A COMPUTATIONAL AND STATISTICAL FRAMEWORK FOR SCREENING NOVEL ANTIMICROBIAL PEPTIDES

by

Daniel Paul Veltri

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

Dr. Jeff Solka — Dr. Jeff Solka, Dissertation Committee Chair

Amarda Shehu — Dr. Amarda Shehu, Dissertation Director

Vaisman — Dr. Iosif Vaisman, Committee Member

Benjamin Matth — Dr. Benjamin Matthews, Committee Member

James D. Willett — Dr. James D. Willett, Director, School of Systems Biology

Donna M. Fox — Dr. Donna M. Fox, Associate Dean, Student Affairs & Special Programs, College of Science

P. Agouris — Dr. Peggy Agouris, Dean, College of Science

Date: 12/1/2015

Fall Semester 2015
George Mason University
Fairfax, VA

A Computational and Statistical Framework for Screening Novel Antimicrobial Peptides

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Daniel Paul Veltri
Master of Science
George Mason University, 2013
Bachelor of Arts
University of Colorado at Boulder, 2006

Chair: Dr. Jeff Solka, Adjunct Professor
School of Systems Biology

Director: Dr. Amarda Shehu, Associate Professor
Department of Computer Science

Fall Semester 2015
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my family, friends, advisers and lab-mates for all of their love, patience and support.

Acknowledgments

This work is supported in part by various seed grants from George Mason University. I would like to thank my committee members, as well as, Drs. Uday Kamath, Elena Rantou, Barney Bishop, Monique van Hoek and Anand Vidyashankar for their useful help and feedback during the progression of this work.

Table of Contents

List of Tables

List of Figures

Abstract

A COMPUTATIONAL AND STATISTICAL FRAMEWORK FOR SCREENING NOVEL ANTIMICROBIAL PEPTIDES

Daniel Paul Veltri, PhD

George Mason University, 2015

Dissertation Director: Dr. Amarda Shehu

Bacterial resistance to antibiotics continues to be a serious concern worldwide. This has motivated a strong research focus on naturally-occurring antimicrobial peptides (AMPs) as templates for new drug development. To date, experiments in the wet laboratory have characterized thousands of AMPs while generally concentrating on measures of antibacterial activity for natural sequences or peptides designed using a limited number of site-directed mutations. Based on these findings, the computational AMP research community seeks to better understand how biological signals and features relate to antimicrobial activity through the use of machine learning and statistical approaches. In this dissertation, we advance our understanding of the determinants for antimicrobial activity by carefully constructing a set of descriptive features for use in AMP classification models. In addition to using physicochemical features, we also construct new sequence-based features which capture information about distal patterns within a peptide. Comparative analysis with state-of-the-art methods in AMP recognition reveal our methods to be among the top performers while still providing a transparent summary of relative feature importance. Moreover, this dissertation applies our features in a new setting to demonstrate for the first time a computational model to recognize if an AMP may perform better against a representative

Gram-positive or Gram-negative bacteria. Work presented is a step forward for *in silico* research seeking to help guide AMP design in the wet laboratory. Our predictive models are made accessible via AMP Scanner, a new publicly-available web server at: `http://www.ampscanner.com`.

Chapter 1: Introduction

1.1 The Problem of Antibiotic Resistance

Drug-resistance in bacteria has been an ongoing problem in hospitals around the world [1–3]. Reports of resistance for the first major clinical antibiotic, penicillin, started to surface in Western hospitals soon after its introduction in the mid 1940's [4, 5]. More recently, other front-line drugs have started to become ineffective against bacteria such as carbapenem resistant enterobacteriaceae and methicillin-resistant *S. aureus* (MRSA). A 2013 report from the US Centers for Disease Control and Prevention cites over two million infections and 23,000 deaths a year from antibiotic-resistant bacterial infections in the US alone [6]. In 2014, US President Barack Obama released Executive Order 13676 [7], a "National Strategy for Combating Antibiotic-Resistant Bacteria," which places a heavy emphasis on promoting new antibiotics and diagnostics. The World Health Organization has also been calling for additional antibiotic development from the medical research community [3]. This dissertation aids in these efforts through the use of computational analysis on naturally-occurring antimicrobial peptides (AMPs) to assist wet laboratory researchers in the discovery of novel treatments effective against drug-resistant bacteria. Given the large financial and regulatory hurdles which have tempered antibiotic development over the past half century, this is also an important academic endeavor [8, 9].

1.2 Antimicrobial Peptides: Nature's Solution

AMPs are a major component of innate immunity and found across all phyla of life [10]. They comprise a number of protein families which can vary in structure, target specificity

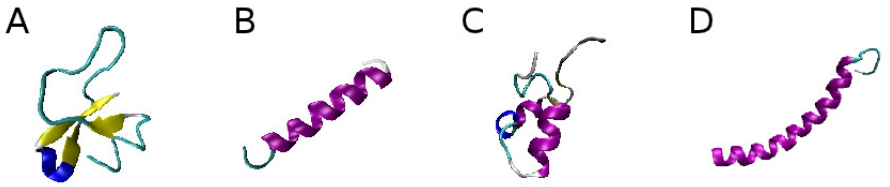

Figure 1.1: Some AMP examples, A) β-defensin 1 from *H. sapiens*, B) Magainin 2 from *X. laevis*, C) Aurelin from *A. aurita*, D) Cathelicidin LL-37 from *H. sapiens*.

and mode of attack. A few major examples include: β-defensins, brevinins, caerins, cathelicidins, magainins and stylins [11–15]. Figure 1.1 shows a few example three-dimensional (3D) structures of AMPs taken from the Protein Data Bank (PDB) [16]. While a number of structural classification systems are found in the literature, the Antimicrobial Database (APD) currently defines eight different structural classes: helical, β-structure, helix and β-unpacked, combined helix and β packed, neither helix nor β-structure, rich in unusual amino acids (AA), disulfide bridge (no 3D structure) and unknown 3D structure [17, 18]. To date, amphipathic α-helical, β-sheet and peptides with highly biased amino acid composition have been the most extensively studied [11].

After millions of years of coevolution alongside their bacterial targets, AMPs have acquired characteristics which make them generally more resilient against resistance compared to conventional drugs [11, 19]. Interest in AMPs as potential novel drugs by the research community is further motivated by their demonstrated ability to kill a broad spectrum of Gram-positive and Gram-negative bacteria in addition to fungi. Some have even been shown effective in breaking down lipopolysaccharide endotoxins [18]. As active AMP fragments (i.e. non-precursor proteins) are generally 50 or fewer amino acids in length, AMPs are also amenable for protein synthesis and manufacturing [11, 15, 20].

Evidence from circular dichroism studies supports the notion that some AMPs, particularly those that are cationic and helical, lack structure prior to contact with a bacterial membrane [13]. This flexibility allows amphipathic AMPs, which are interspersed with cationic and anionic regions, to utilize side-chain packing. The cationic residues rotate until they face the anionic head groups of phospholipids on the surface of a bacterial membrane. This, in turn, directs the peptide toward the membrane via an electrostatic attraction [11, 13, 21]. Such an action is just one example of how a physicochemical property like charge is intrinsically encoded in a peptide AA sequence and can influence an AMP's mechanism of action (MOA). In the case of eukaryotic cells, which excrete AMPs, the presence of cholesterol in their membrane prevents a negative charge from forming on the surface strong enough to attract AMPs to attack itself [11,22]. Some notable exceptions are AMPs with a high hydrophobicity-to-charge ratio such as melittin or those composed of mostly L and K residues. Such peptides are less discerning about membrane type and may demonstrate hemolytic activity against red blood cells [23, 24].

Years of research on a wide variety of AMPs has produced a number of hypotheses for how AMPs kill their targets. The general MOAs thought employed may include: physical membrane disruption (through the creation of pores, charge disruption, lipid clustering, etc.), membrane-bound receptor inactivation, interference with DNA replication or other targets in the cytoplasm, and even roles as signaling molecules to activate adaptive immune responses [11, 18, 21, 25]. Depending on the peptide, some of these may not be mutually exclusive. For example, LL-37 is known both to attack membranes and signal the adaptive immune system in humans [26]. While much progress has been made in understanding the specifics for how these mechanisms work at the molecular level, the subject is still hotly debated [27]. Exciting new progress is being made in the wet laboratory which may help generate more conclusive descriptions of AMP-membrane interactions. For example, Oreopoulos et al. recently used polarized total internal reflection fluorescence microscopy (pTIRFM) to visually and temporally observe AMP-induced reorientation of membrane domains [28]. Unfortunately, AMP databases at present do not characterize any definitive

MOAs. First steps in this direction can be seen with the Database of Antimicrobial Activity and Structure of Peptides (DBAASP) from [29] which allows one to search for peptides that target a keyword like "`lipid bilayer`" or "`DNA`." However, this still lacks sufficient detail to be particularly useful to this dissertation, as mechanisms may not be mutually exclusive; there are also numerous competing models for how an AMP might actually target a lipid bilayer [27]. A few examples, some of which are shown in Figure 1.2, are: the carpet mechanism [21, 30], toroidal pore [21], barrel-stave pore [21, 31], or even β-amyloid-like mechanisms [32]. For this reason, rather than focus on specific MOAs, this dissertation utilizes physicochemical and sequence-based features to predict if a peptide may have generalized antimicrobial activity (AMP vs. non-AMP) or selectivity (Gram+ vs. Gram-). Additional details about current theories on AMP killing mechanisms can be found in a number of comprehensive reviews [27, 33–35].

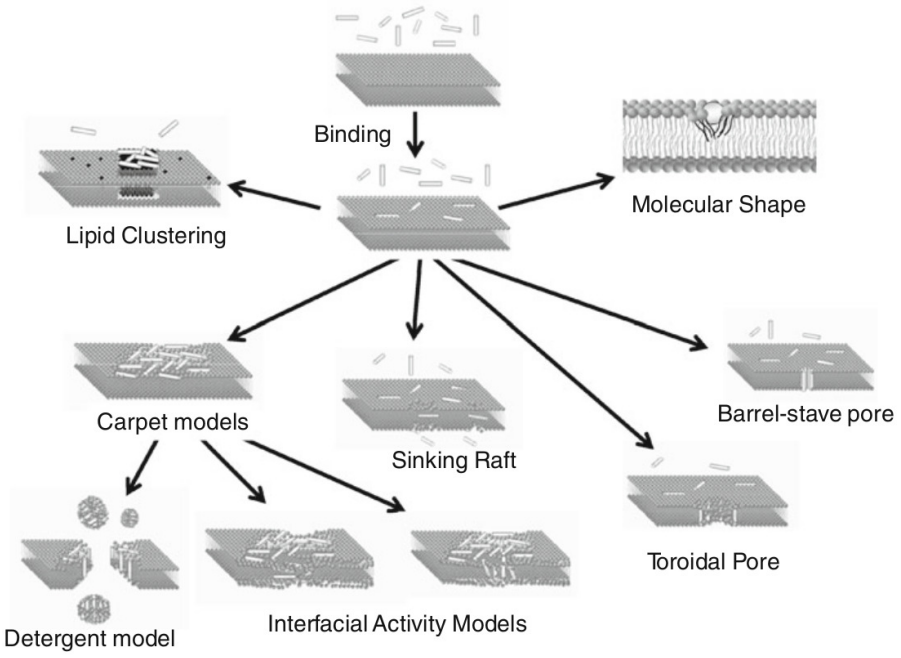

Figure 1.2: Some illustrations of different attack mechanisms proposed for a variety of AMP classes in the literature. These mechanisms are well characterized by experiment but do no apply to all AMPs [27]. Figure originally presented by Wimley and Hristrova in [27] and used with permission from Springer Publishing.

1.3 Role of Computation in AMP Analysis and Design

The development of new drugs in the wet laboratory is an expensive and time-consuming endeavor. Computer-based approaches stand to benefit drug development by screening out low-probability targets and highlighting those with a higher likelihood for success. In turn, this can aid in hypothesis generation to speed up drug development and lower costs. Since AMPs stand as excellent templates for novel antibiotic development, computational investigators ultimately seek to determine the functional and mathematical relationships that link the physicochemical properties innate to an AMP's AA sequence with general antimicrobial activity. Since AMP structures are known to change their conformation based on the surrounding environment as mentioned above, we choose to focus our models on

primary AA sequences as these remain constant.

To date, the field has mostly relied on machine learning methods to build predictive models which can assign an "AMP" or "non-AMP" class label when presented with the sequence of a query peptide as input. These approaches typically rely on sets of quantified physicochemical properties (e.g. hydrophobicity, isoelectric point, charge, etc.) which have been identified as important from previous wet-laboratory experimentation and stored in a feature vector to represent a peptide. However, the process of encoding variable-length protein sequences into a discrete model or fixed-length numerical vector required by existing learning algorithms like support vector machine (SVM) or artificial neural networks (ANN) is not a straightforward task [36]. The main concern, both for AMP recognition and other areas of bioinformatics using biological sequences, is how to represent a complex three-dimensional organic compound like a protein as a vector of numbers and retain adequate sequence-pattern information. For example, taking simple counts of AA frequency in an AMP and storing it in a 20-dimensional vector will lose all information about sequence order. A popular method to address this issue is pseudo-amino acid composition (PseAAC) [37] which captures correlations between relative AA positions in a sequence. The approach has been used in numerous areas for computational proteomics and genetics [38, 38, 38, 39] in addition to work on AMP recognition [40, 41]. Other approaches, such as that of Lata in [42], use AA counts over a pre-determined number of positions at the N- and C-termi; terminal regions are known to have a variety of AMP-related functions [18, 21, 43]. Despite the lack of positional information, the most common approach in the field is to simply average features over the entire length of a peptide [44–49].

Pairing the encoded features mentioned above with various prediction algorithms has resulted in a number of good models reported in the literature with accuracy in the $70-90\%$ range (a summary is provided in Table 2.1 in the next chapter). However, as many of these studies do not make their data or implementation publicly available, it can be difficult to draw a clear comparison of performance. This stems from the lack of a "gold standard" benchmark data set, which remains a problem in the field (and will likely require extensive

collaboration with biochemists and molecular biologists for manual curation of non-AMPs). Examples of some common algorithms utilized for AMP recognition include: hidden Markov models (HMM) [50], ANN [46, 47, 51], random forests (RF) [44], and SVM [42, 44, 52]. A major issue with many of these "black box" methods [53] is a lack of transparency in how features are being used to make predictions. Computational studies on AMPs using many of the aforementioned algorithms typically do not provide a rigorous analysis of features or rank their relative importance. Furthermore, the field has largely ignored the role of interactions between features. While a method like PseAAC may capture simple relationships between AA positions, they have not tried to link more complicated co-occurring (or perhaps mutually exclusive) motifs together within the same sequence.

While there is no doubt much progress has been made in the arena of computational recognition of AMPs, work has focused almost exclusively in a binary setting to predict AMP activity. To date, machine learning (ML) has not been used to addressed the problem of AMP-selectivity at the level where responses against bacterial classes are considered. This is a critical question as responses can vary dramatically. For example, Zhu et al. show in [54] that the peptide RFRRLRCKTRCRLKKI is effective against *E. coli* but ineffective against *S. aureus*. With the eventual goal of computer-assisted *de novo* drug design in mind, methods are needed which not only predict if a peptide has AMP-like activity, but also predict if the AMP can kill a particular strain of clinical importance such as MRSA. Years of research on AMP-membrane interactions in the areas of biochemistry [25, 55–57] and molecular dynamics [34, 58, 59] suggest AMP models need to take into consideration the lipid membrane composition of a target bacteria if this goal is to be achieved.

1.4 Problem Statement and Objectives

It is the position in this dissertation that the current computational efforts for AMP prediction lack a rational design process for generating features that go beyond basic sequence composition. The majority of existing work also lacks adequate transparency and rigor in

considering feature importance and possible interactions in predictive models. This is important, both for experimentalists designing AMPs in the laboratory and for computational researchers seeking an algorithm to reliably generate novel AMP sequences. Furthermore, the field has yet to address the open problem of predicting if a peptide may be more effective against one species of bacteria over another. This dissertation addresses these issues by gradually and iteratively building models to predict both AMP activity and selectivity. Validation of models is considered in both an internal and external context; models are built and evaluated through the use of both cross-validation (CV) and data sets with separate training and testing partitions when selecting features and assessing performance. Where possible, direct comparison against other state-of-the-art approaches for AMP recognition are also carried out and reported. This dissertation offers the following contributions to the field:

- **Predicting AMP Activity:** Predictive models presented in Chapter 6 provide best-in-field average performance on two publicly-available data sets when compared against other publicly-available methods.

- **Predicting AMP Selectivity:** Models provided in Chapter 7 take a first step at designing features for AMPs known effective against Gram-positive and Gram-negative bacteria. An additional model predicts if an AMP may be more active between the clinically important bacteria *E. coli* and *S. aureus*. To the best of our knowledge, this is the first time learned computational models in this setting have been reported.

- **Feature Interactions and Rankings:** In Chapter 4 we show that feature interactions are important for AMP recognition and further incorporate interaction terms into our feature sets.

- **Transparency of Features:** All top models presented in Chapters 4-7 report relative feature rankings and highlight relevance to experimental results from biology when appropriate. Decision trees in Chapter 7 provide visual descriptions for the role features play in the classification process.

- **Transparency of Data and Methods:** All data sets used for training and testing of models in this work are made publicly available. To further aid the research community and for reproducibility, we implement our top predictive models as a free web server available at: `http://www.ampscreen.com`.

Throughout this work, we employ a solid framework to design and evaluate our predictive models based on Chou's 5-step rule [60]. Shown effective in a series of recent publications [38, 61–66], we use the following five guidelines to establish useful sequence-based statistical predictors for a biological system:

1) **Construct or select a valid benchmark data set to train and test the predictor:** In this work, we use a pair of benchmark data sets taken from peer-reviewed literature to allow for a direct and fair performance comparison against other methods. Those, in addition to other data sets which we carefully construct, are composed only of AMPs which have been experimentally-validated. As there is no data set of experimentally validated non-AMPs currently available, best efforts have been made to screen out any potential AMPs using homology-based comparison methods.
2) **Formulate the biological sequence samples in a manner that can truly reflect their intrinsic correlation with the target to be predicted:** We pay careful attention in this work to the task of encoding AMP sequences as feature vectors. In Chapter 5 we introduce novel distal-based features which capture direct correlations between neighboring and non-neighboring motifs within AMP sequences. Non-AMPs are also used in a scoring function to filter out features generated which are not specific to AMPs. Features in Chapter 6 are designed to represent known functional regions of an AMP sequence to better reflect their biology.
3) **Introduce or develop a powerful algorithm (or engine) to operate the prediction:** Feature sets are paired with a variety of established statistical approaches and machine learning algorithms soundly based in statistical theory. We use a variety of methods capable of handling linear and non-linear data.

4) **Properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor:** Where possible, we use separate training and testing data sets to evaluate performance. When CV is performed, we apply feature selection and parameter tuning exclusively to the training fold, and report average performance for the testing fold.

5) **Establish a user-friendly web server for the predictor that is accessible to the public:** As mentioned above, we provide a free and easy-to-use web server to make our method available to the AMP research community. This allows for reproduction of our results and we also provide two helpful settings for the user. "Classify Sequences Mode" classifies a peptide as AMP or non-AMP in the context of the entire sequence. "Proteome Scan Mode" allows the user to upload a proteome to scan for potential AMP segments identified within larger proteins. For sequences selected as potential AMPs, both modes also provide predicted probabilities for the sequence being more effective against *E. coli* or *S. aureus*.

1.5 Dissertation Structure

Having introduced AMPs and their basic biological activity above, we begin in Chapter 2 by providing relevant background information on concepts related to machine learning, performance measurements, and an overview of the current state of the art in computational AMP recognition. As a number of AMP data sets are used repeatedly throughout this work, Chapter 3 acts as a central reference point to describe the composition, origin, and any cleaning or pre-processing steps performed for all data sets. In Chapter 4 we begin the model-building process, starting with a simple and interpretable logistic regression (LR) model using 8 initial physicochemical features to investigate if feature interactions are important for improving AMP classification performance. After showing interactions to be beneficial, Chapter 5 leverages the power of an evolutionary algorithm to generate hundreds of novel features which not only recognize important sequence motifs in AMPs, but go beyond basic composition and capture relevant correlations between neighboring

and non-neighboring motifs. Chapter 6 combines the features from previous chapters and reorganizes them about functional regions within an AMP peptide. Three different feature selection approaches are then compared before we identify and apply a subset of features to generate two final predictive models for AMP classification that improve on the state-of-the-art. Chapter 7 focuses on designing models for target-selectivity and identifies subsets of features to recognize groups of AMPs known effective against Gram-positive and Gram-negative bacteria. It also introduces a model to predict better activity between *E. coli* and *S. aureus*. Chapter 8 details a web server constructed to make our new models freely accessible to the AMP research community and to help individuals mine AMPs from within a proteome. We submit a number of proteomes to the server and introduce a list of top potential candidates from each. Recent AMPs discovered in *Alligator mississippiensis* in [67, 68] provide an excellent opportunity to evaluate the server using external experimental data. This dissertation concludes with Chapter 9 which presents a final overview and discussion of the work presented and possible future research directions for this line of AMP research.

Chapter 2: Additional Background Information and Related Work

In the previous chapter, we introduced the problem of antibiotic resistance and the basics of AMPs and how they work. Throughout the rest of this dissertation, two main lines of investigation will be considered: predicting AMP activity (recognizing if a peptide has antimicrobial activity) and predicting AMP selectivity (recognizing if a peptide has a preference for certain types of bacteria). Before we begin the process of designing predictive models, we provide in this chapter some necessary background information on how we can address these problems in a computational setting. In particular, we discuss the ML and statistical methods used in this work, as well as, other computational approaches employed by the computational AMP community to date.

We start by describing a metric used by experimentalists for quantifying AMP efficacy and then proceed with basic definitions of ML, classification, and related evaluation metrics. Next, we provide a performance summary for different classification methods published in the literature and detail some of the top performers. We conclude with some basic theoretical background on the three different classification methods used to build models in this dissertation which were taken from both the ML and statistics communities. Additional information on the ML algorithms and statistical approaches introduced in this chapter can be found in [69–71].

2.1 Experimental Measurement of Antimicrobial Activity

Minimum inhibitory concentration (MIC) is a basic measurement used in microbiology wet laboratory experiments for evaluating the efficacy of an antimicrobial against a bacterial sample. The value (typically given in units of μM or μg/mL) is determined as the minimum

concentration of antibiotic required to inhibit bacterial growth overnight [72]. Accordingly, a lower MIC value denotes a more powerful antibiotic. Other measurements of activity (e.g. EC50, LD50, etc.) are not considered in this dissertation, as these are sparsely represented in current AMP database entries with information characterizing activity [29, 44, 73].

2.2 Machine Learning Approaches

ML refers to the application of computer algorithms and artificial intelligence for problem solving. When provided a series of observations and features as input, an algorithm attempts to deduce patterns or rules which return an appropriate solution [71]. Various machine learning techniques have successfully been used in a number of bioinformatics applications [74]. For example, artificial neural networks have predicted protein cleavage sites [75], genetic algorithms have been used to determine gene expression levels [76] and DNA promoter binding sites [77], and decision trees have been applied to protein secondary structure prediction [78]. Many different ML approaches have also been used in the field of AMP classification as listed in Table 2.1 and discussed further below. ML is especially helpful when dealing with complex and high-dimensional data. In the context of this dissertation, we use ML and statistical methods to focus on discriminating between positive (AMP) and negative (non-AMP) observations. In other words, given an unknown peptide sequence, we hope to correctly "recognize" or "classify" it as an AMP or non-AMP. In Chapter 7, we also consider discriminating against *E. coli* and *S. aureus* to predict which taxa would have a lower MIC value when challenged with the same AMP.

2.3 Learning Classification Models

Classification, a common statistical and machine learning task, involves assigning an unknown target to a particular group or label. Binary classification problems involve distinguishing between two groups (e.g. 0/1 or true/false). Multi-class problems involve assigning three or more labels (e.g. high/medium/low). A classifier is first trained on positive and

negative examples from a training data set in order to learn patterns from the observations provided. It is important that data used to train a model is a good representation of the natural distribution of observations in the real world. Poor training data comprised of observations and/or features which inadequately represent this natural distribution can generate models which fail to reproduce any new data presented to it. For example, overfitting can occur when a model mistakenly learns from random noise in the training data and confuses it for a real phenomenon. Underfitting can occur if the training data has too few examples or if an inappropriate learning algorithm is selected (e.g. a linear model is paired with nonlinear data) [69, 71]. Learning is considered "supervised" when the training examples have labels of a known class [71]. In this setting, once a model has been trained, predictions can then be made on a separate testing set where each observation is assigned to a class. If labels are available for the testing data, one can perform validation as described below. In "unsupervised" learning, there are no labels used to train the data, so an algorithm will try its best to cluster and assign observations into groups based either on rules or a user-designated number [69]. For this dissertation, we will use supervised learning to train and build our models using AMPs and non-AMPs taken from a number of data sets which are detailed in the next chapter.

2.4 Validating Learned Classification Models

If a testing or validation set (not used to train a model) is provided with labels for all observations, performance can be evaluated by comparing the actual classes against those predicted by an algorithm. Predictions in agreement with reality are referred to as "true positives" (TP) or "true negatives" (TN). Erroneously classified observations are known as "false positives" (FP) or "false negatives" (FN). For a fair validation, it is important that observations do not overlap between the training and testing sets (this would essentially be peeking at the answer) and that they are pulled from the same distribution. In cases where few observations are available and no separate testing or validation set can be formed, one

can perform K-fold CV using all of the available data in a manner similar to bootstrapping in statistics [69–71]. Here, K refers to the number of groups the data set has been split into (typical numbers are 5 or 10). After splitting the data into K groups, observations from $K-1$ groups are pooled together for training, and testing is performed on the remaining data. This process is then repeated until all K splits have had a chance to act as the testing set and performance is then averaged over the K separate runs of the algorithm. The idea behind CV is to try and estimate the error associated with a model by simulating training on only part of the data in much the same way any training data set is just a sample of observations taken from the real world distribution [69, 71].

A number of metrics are available for quantifying classification performance. Determining the appropriate performance metric and acceptable level of type I error (when a null hypothesis is mistakenly rejected) and type II error (when a null hypothesis is mistakenly accepted) depends on the given problem at hand. Due to the high cost of producing AMPs in the lab, this dissertation will prioritize generating the fewest number of FP predictions as possible. Some common performance metrics used for classification and employed throughout this dissertation are briefly outlined in the next section.

2.5 Evaluation Metrics Considered for Model Performance

2.5.1 Sensitivity and Specificity

Sensitivity assesses type II error and is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

Specificity assesses type I error and is defined as:

$$\text{Specificity} = \frac{TN}{TNs + FP}.$$

In both cases, values range from 0 to 1 and higher values equate to lower respective error rates.

2.5.2 Accuracy

Accuracy (ACC) represents how repeatable measurements are when conditions remain constant [79]. In the context of classification, this represents how well a method can predict sample classes and is defined as:

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \times 100(\%).$$

Values can range from 0 to 100 percent. Larger values equate to better classification performance given that the number of positive and negative examples are comparable in size.

2.5.3 Mathew's Correlation Coefficient

Matthew's correlation coefficient (MCC) [80] is another summary statistic to evaluate binary classification performance and is more stringent compared to ACC, as it weighs the impact of false predictions more heavily. MCC is defined as:

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

Values can range from 0 to 1. Larger values equate to better classification performance.

2.5.4 Brier Score

Brier score (BS) is a metric used to quantify the overall accuracy of model predictions and is defined to be the average of the squared deviations between the observed and the predicted

values (probabilities) of a binary response variable [81]. It is defined as:

$$BS = \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i)^2 .$$

Brier scores range from 0 (an ideal model) to 0.25 (a poor model) and are very close in concept to residual deviance [81]. The squared deviations in the BS can also be thought of as the squared residuals.

2.5.5 Akaike Information Criterion

Akaike Information Criterion (AIC) is a score used to compare and rank the performance of different models. It is a regularization method which measures the loss of information from a model and penalizes the addition of any superfluous predictors to help prevent overfitting [82]. When comparing models for a given data set, the best performer corresponds to the one with the smallest AIC value. Work has shown that lower-scoring models in terms of AIC can generate a better estimate of the original data set of interest [83]. It should be noted that AIC can only be used for comparison when considering models generated from the same data set.

2.5.6 Receiver Operating Characteristic

A receiver operating characteristic (ROC) [84] curve provides a convenient graphical representation of classification accuracy and can be seen in Chapter 4 Fig. 4.3. The curve is generated using classifier predictions ranked in descending order of confidence. A cutoff is moved along the ranking list to assess how the number of true and false positives below the cutoff line change. An ROC curve captures the true positive rate as a function of the false positive rate as this cutoff moves along the ranking list [84]. The ROC score is a summary statistic which refers to the area under the curve (auROC). Random ranking would be expected to yield a score ~ 0.5 and would be represented by a straight diagonal line from the

lower left to top right of the graph. Scores greater than 0.5 bend the line towards the upper left corner as classification performance improves. A large bend to the lower right might indicate that observations are mislabeled. An auROC score reaches 1 (or 100 when given as a percent) if a classifier correctly predicts the class for all observations.

2.5.7 Precision and Precision Recall Curves

Precision, also known as positive predictive value, represents the proportion of positive and negative results and ability for information retrieval. It is defined by:

$$\text{Prec.} = \frac{TP}{TP + FP}.$$

Values range from 0 to 1, with higher values equating to a better ability to make positive predictions. A precision recall curve (PRC) is a graphical representation of how precision changes as a cutoff moves along a ranking list similar to that for ROC discussed above. It may be better suited for imbalanced data compared to ROC and, like auROC, the area under the PRC (auPRC) can also be used as a summary statistic for how a model performs.

2.5.8 Information Gain

Information gain (IG), also commonly referred to as the KullbackLeibler divergence, is a way to assess a feature's relevance based on a change in information entropy [85, 86]. For a given data set D, with classes ranging from $i = 1$ to k, entropy I is given by:

$$I(D) = -\sum_{i=1}^{k} P(C_i, D) \cdot \log(P(C_i, D)).$$

For a feature f taking on values(f) different values in D, the weighted sum of its expected information (over splits of the data set D according to the different values of f into D_v

subsets, with v ranging from 1 to values(f)) is given by:

$$Info_f(D) = -\sum_{v=1}^{\text{values(f)}} \frac{|D_v|}{|D|} \cdot I(D_v).$$

The IG for a feature f over a data set D is then given by:

$$IG(D, f) = I(D) - Info_f(D).$$

2.6 AMP Classification Methods

Many machine learning methods have been employed to recognize (classify) AMPs amongst a set of non-AMP peptides. This has generally been performed with all AMP families grouped together regardless of structural class and in a binary setting [41, 42, 44, 46, 47]. In other words, methods assign a query peptide an "AMP" or "non-AMP" label but ignore the question of how more or less potent one peptide might be compared to another. Table 2.1 below summarizes the state of the art for binary classification using standard ML metrics defined in the previous section. Algorithm abbreviations used throughout this dissertation are as follows: ANFIS (Artificial Neural Fuzzy-Interface-System), ANN (Artificial Neural Network), DA (Discriminant Analysis), FKNN (Fuzzy K-Nearest Neighbor), HMM (Hidden Markov Models), LR, NNA (Nearest-Neighbor Algorithm), RF (Random Forest), SVM (Support Vector Machine).

2.6.1 Features Used for AMP Classification

A variety of features have been applied to the problem of AMP recognition. These are most-often generated for primary protein sequences and utilize either basic composition [40–42, 44, 87], PseAAC [40, 41], or employ experimentally-determined values for each of the 20 standard AAs. For the latter, typical examples of physicochemical properties used include: acid dissociation constants [41], "Bowman Index" (binding potential based on side chain

free energies when a peptide transfers from cyclohexane to water) [11,73], charge [40,44,46, 47,52,88], hydrophobicity [40,41,44,46,47,52,88], isoelectric point [40,41,52,88], molecular weight and volume [40,41,52,88], polarity [40,44,52,88], secondary structure predictors [40, 44, 46, 47, 52, 88], solvent accessibility surface area [44], van der Waals volume [44, 52, 88], and calculations for likelihood for peptide aggregation [46,47,52,88]. Jenssen and colleagues in [50] build a mathematical model using principal component analysis and partial least squares to relate features based on 3D quantitative structure-activity relationships (QSAR) to antimicrobial activity using adjusted MIC values. This work was further improved using neural networks in [51]. One problem facing both of these QSAR-based approaches is their reliance on predicted 3D structures of AMPs from computational modeling. Their features can be biased if a poor energy function is used during the initial modeling stage.

2.6.2 Survey of Machine Learning Approaches Used to Date

Table 2.1: A summary of AMP prediction algorithms and data sets is shown. The first column includes a list of algorithms and the citation for where it is applied to AMP recognition. Columns 2-4 include training, validation and testing performance when reported in terms of MCC. The final column includes the database or source of AMP observations used for testing.

Method	**MCC**			**Database Source**
	Training Data Set	Validation Data Set	Testing Data Set	
HMM [50]		0.98		AMPer
ANN [51]		0.88		RANDOM
DA [44]	0.75		0.74	CAMP
RF [44]	0.86		0.86	CAMP
SVM [44]	0.88		0.82	CAMP
SVM [87]			0.84	AntiBP2
NNA [40]			0.73	CAMP
SVM [52]			0.80	APD
ANFIS [47]		0.94		APD
ANN [47]		0.85		APD
FKNN [41]	0.73		0.84	APD

Direct comparisons between these approaches are hard to draw due to the great diversity amongst the features and data sets considered. However, results do show that good recognition accuracy can currently be obtained in a controlled setting. For instance, work in [47] achieves an MCC value of 0.94 for recognizing AMPs using only features of length (number of amino acids) and propensity for aggregation. Interestingly, physicochemical properties shown useful in the wet laboratory for modifying antimicrobial activity for various AMPs (such as, hydrophobicity, propensity for certain secondary structures, and more) were shown not to be important for automatic recognition [47]. An issue facing ML approaches is the reliance on hand-picked negative data sets for training and testing. Many methods rely on database keywords and assume that, if the record is not listed as antimicrobial, then it can be used as a negative example without explicit additional testing [41, 46, 47]. The use of a poor quality negative data set can cause model bias and may select for a set of spurious features.

AMP sequences can be quite diverse and other approaches have attempted to overcome this by focusing on a fixed number of terminal amino acids and employing simple features based on amino acid composition prior to using a machine learning classifier [42, 44, 45, 87]. Such approaches discriminates between AMPs and non-AMPs with accuracies in the $80 - 90\%$ range. However, what these features capture and how to properly interpret them is difficult considering AMPs are highly-constrained peptides in terms of physicochemical and structural properties [10, 11]. Further problems with some machine learning techniques, such as SVM, is that they can act as a "black box" and obfuscate why exactly a peptide might be classified as an AMP. Accordingly, their value for drawing rules to design novel AMP-based drugs in a wet laboratory setting has had limited success. For example, a recent survey of the American alligator (*Alligator mississippiensis*) by Bishop et al. in [67] detected 8 potential AMPs and determined 5 to have antibacterial activity. They tested these peptides using three publicly available AMP prediction servers: *CAMP* (`http://www.camp.bicnirrh.res.in/predict`) from [44] which supports DA, RF and SVM as seen in rows 5-7 in Table 2.1 above, *AntiBP2* (`http://www.imtech.res.in/raghava/`

antibp2) from [42] which uses only SVM as seen in row 8 in Table 2.1 above, and a predictor service from the APD (http://aps.unmc.edu/AP/prediction/prediction_main.php) which uses their database and a number of physicochemical properties but does not detail the specifics of their algorithm. The authors found *CAMP* could not accurately predict any active peptides, and only the DA algorithm correctly rejected 2 of the inactive ones. They submitted 6 peptides to *AntiBP2* and found all predictions to be incorrect. They did obtain 5/8 correct predictions with the APD predictor but this method is not transparent and only allows the user to submit one peptide at a time. These results illustrate that there still remains sizable room for improvement in the field of AMP recognition.

2.6.3 Classification Methods Used in this Work

Binary Logistic Regression

LR is a general linear model first introduced by Cox in [89]. In brief, using a set of observations X with k features or predictors (X $= x_1, \ldots, x_k$), LR allows us to build a model to predict an ordinal response variable y. In the binary case, where we assume y follows a Bernoulli distribution, y is typically $\{0, 1\}$ but can represent any two class labels. Throughout this dissertation we use 0 to represent a non-AMP and 1 to denote an AMP observation.

LR allows us to consider a single or multiple predictors. For the single predictor case, where no interactions are modeled among features, if x_j is the j^{th} feature for $j = 1, \ldots, k$, then one can define:

$$P(y = 1|x_j) = \frac{e^{b_0+b_1x_j}}{1 + e^{b_0+b_1x_j}}.$$

Here, $P(x_j)$ in our case is the probability that a peptide is an AMP as a function of the j^{th} feature. If we denote this probability as p, then the logit or inverse logistic function can be defined as $ln\frac{p}{1-p} = b_0 + b_1x$. This defines the b coefficients as the linear regression coefficients of the logit function. This is useful since we can take any input in $[-\infty, +\infty]$

and output a response value in then range $[0, 1]$. Such models can be extended to the case of having multiple predictors or predictor interactions. In the case of K-predictors, the above probabilities are calculated as:

$$P(y = 1|\mathbf{x}) = \frac{e^{b_0+\mathbf{x}^T\mathbf{b}}}{1 + e^{b_0+\mathbf{x}^T\mathbf{b}}}.$$

Here, $\mathbf{x}$ and $\mathbf{b}$ represent the k-dimensional vectors of the predictors and the regression coefficients, respectively. The coefficients can then be used to measure the importance of individual predictors to the model or to calculate an odds ratio where:

$$OddsRatio = \frac{e^{b_0+b_1(x+1)}}{e^{b_0+b_1x}}.$$

Additional background information on the topic of LR is available from Kuhn and Johnson in [70] and Hastie et al. in [69].

Multivariate Adaptive Regression Splines (MARS)

MARS is a powerful non-linear extension of multivariate regression which can handle continuous or ordinal data and developed by Friedman and first introduced in [90]. The acronym "MARS" is trademarked and licensed to *Salford Systems* (San Diego, CA). The method employs "hinge" functions to help partition the data in a non-linear feature space. The potential benefits of MARS over standard linear regression can seen in the toy example in Figure 2.1.

Models are built to approximate a function as follows:

$$\hat{f}(x) = \sum_{i=1}^{k} = a_k \mathrm{B}_k(x),$$

where B_k is a basis function of the form:

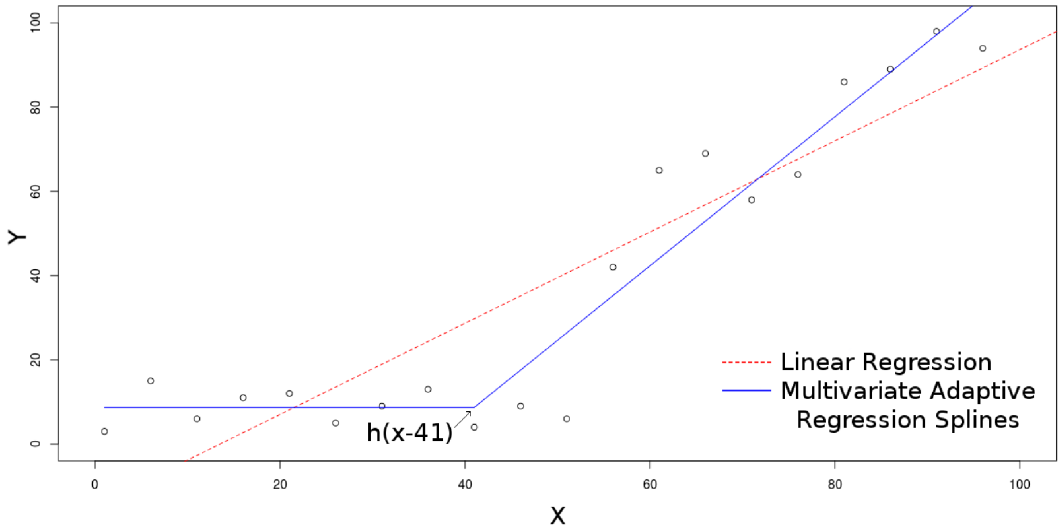

Figure 2.1: Seen here is a toy example to demonstrate how multivariate adaptive regression splines can use a hinge function to partition non-linear data. A dashed line fit with classic linear regression is shown in red, while a solid blue line shows a fit generated with multivariate adaptive regression splines. The respective residuals sum square (RSS) and adjusted R-squared values for linear regression are 4693.1 and 0.795. For multivariate adaptive regression splines, the respective RSS and R-squared value is 1626.4 and 0.933 (GR^2:0.916, GCV:112.6). The later method selected two terms, a y-intercept at 8.7 and the hinge function $h(x-41)$ (labeled on graph with a black arrow) which it used to partition the data into disjoint regions. Values for x to the left of this point take $max(0, 41-x)$ while those to the right take $max(0, x-41)$.

$$\mathrm{B}_k = I[x \in R_k].$$

Above, a are coefficients, I is an indicator function (set to 1 if true, 0 if false), and R_k is a sub-region of the feature space. The basis function can consist of a single intercept term, a hinge function, or interactions between two or more hinge functions. The hinge function is mirrored about a point c (the "knot") which continuously separates the sub-region R into R_{left} and R_{right} according to $max(0, x - c)$ and $max(0, c - x)$ respectively. In the example of Figure 2.1 above, $c = 41$ which sets x equal to the intercept when ≤ 41 yet allows x to increase when > 41.

The model building process uses recursive partitioning (similar to RF discussed below) and takes place in two stages. The first stage uses a greedy approach to exhaustively adds pairs (or higher-order terms) of hinges to the list of basis functions. This is done until either a user-set limit on the number of terms is reached, or the residual error no longer changes when terms are added. The second stage then proceeds in the reverse direction to prune hinges which may contribute to over-fitting. Subsets of the data are compared using generalized cross validation (GCV) to try and minimize the number of features (i.e. complexity) while maximizing performance. Calculation of GCV attempts to approximate leave-one-out CV by using RSS and further details can be found in [90, 91]. A lower GCV is considered better and results are often reported with a related R-squared value (GR^2) which estimates R^2 performance based on subsets of the data rather than the entire data set.

Some major benefits of multivariate adaptive regression splines is that selected variables are interpretable like linear regression with the added benefit of the model-building process performing internal feature selection to remove unimportant terms. The method is also fast once models are built and relatively robust against the influence of outliers. Additional details about MARS can be found in Friedman's original paper [90] and in Kuhn and Johnson in [70].

Random Forest

RF is a method based on using a collection of decision trees to solve classification or regression problems and was introduced by Breiman in [92]. A decision tree is a simple graph which breaks a complex problem into simpler parts to characterize a given query. Features can be numeric or nominal and play the part of nodes which ask simple questions such as: "does the query match feature k (`TRUE`/`FALSE`)?" or "is feature k for this query $\geq$ to a certain value?" An example of a simple decision tree is given below in Figure 2.2 to recognize if a query fruit as an apple, orange or banana. Three Boolean (`TRUE`/`FALSE`), features are shown as nodes in the tree: `SphericalShape?`, `Poisonous?`, `RedColor?`. One can rank features by assessing which does the best job at correctly separating the observations in half. A final decision tree can then be constructed with the best feature as the parent, followed by the next-best features in the following level and so-on until all features are used [69–71, 92].

Using our example problem, if we only have observations which are apples, bananas and oranges, the feature `Poisonous?` can be seen as unnecessary since none of the three fruits are poisonous. However, in more complex problems with many features present, it may not be obvious which are unnecessary or potentially correlated. As poor selections when pruning features can lead to models which over- or under-fit the data, Brieman introduced a concept similar to bootstrapping for decision trees known as "bagging" in [93]. This basic concept builds n_{trees} trees, each from a random sample of observations without replacement, and assigns the response found by the majority of trees. It has been shown that the use of random samples help in warding against the problem of overfitting [71, 93]. However, some researchers have noted that the use of sampling without replacement could be a potential point of feature selection bias, particularly, for binary features [94].

RF adds an additional sampling element on top of bagging by only using a smaller random subset of K features to construct a tree (in other words, if RF is set to use all features to build trees, it is equivalent to bagging). By selecting only a subset of features (for example square root of (k)), and assessing how well their trees perform on the data not selected (during the bootstrap step) for their construction, one can also make an error

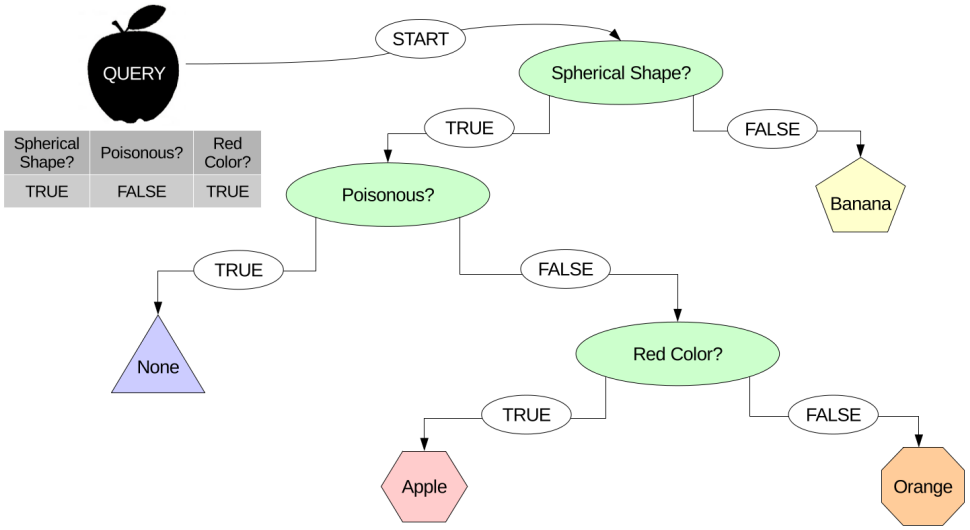

Figure 2.2: Seen here is a toy example of a decision tree to recognize a query fruit. A vector containing Boolean values for the three features: `SphericalShape?`, `Poisonous?`, `RedColor?` is listed with the example query. The decision tree turns each of these features into a node with a simple `YES` / `NO` question. The query begins at the top (parent) node and progresses until it reaches a leaf or terminal which assigns it with a class label. Given the query fruit shown, it would be recognized as an apple. Note, if no poisonous fruits are considered, the `Poisonous?` feature does not aid in the classification process. Some decision tree methods would recognize this and "prune" or remove the node and proceed to reconstruct the tree without it. In this manner, tree construction can work as a basic feature selection method and remove unhelpful or redundant features.

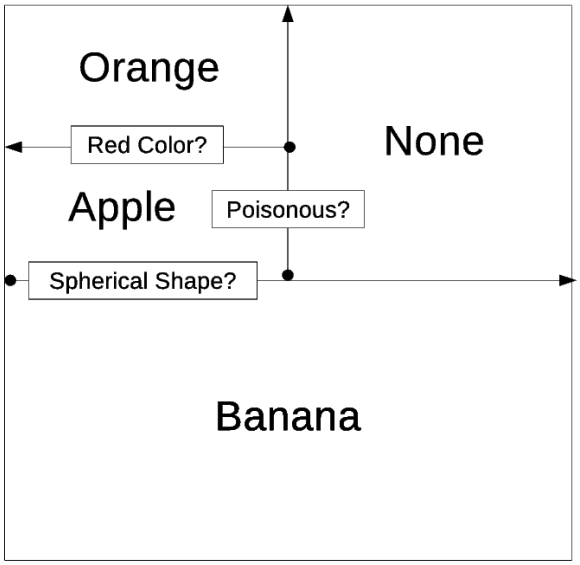

Figure 2.3: Seen here is another representation of the toy example shown in Figure 2.2. We begin with the feature space as a square and draw the feature `SphericalShape?` as a vector to separate the space in half. An answer of `TRUE` partitions the top half, an answer of `FALSE` assigns a label of banana in the bottom half. The next feature `Poisonous?` further separates the top half of the space in half, and the process is repeated for `RedColor?` until a query can obtain any of the labels. In the case of numeric features, vectors could be shown as number lines to further delineate how the space is partitioned [70].

estimate (known by the method as the "out-of-bag" error) for how well the classifier will perform [70, 92]. This is similar to the concept of CV or GCV discussed above. Additional feature ranking can also be performed by keeping track of which features are consistently used to construct the best performing trees.

While we limit our discussion here to the context of classification, decision trees can also make predictions for numeric responses. Figure 2.3 shows an alternate representation of Figure 2.2 and illustrates how points can be located in a feature space for use in a regression problem [70].

Additional information about decision trees, RF, bagging, and related concepts (e.g. "boosting") can be found in [69–71, 92].

Chapter 3: AMP Data Sets

In order to generate the models considered throughout this work, a variety of data sets consisting of AMP and non-AMP observations are employed. As many are used repeatedly, we detail them here rather than repeat their description multiple times. When possible, data sets provided from other reviewed literature in the field are utilized to allow for direct performance comparisons. This chapter describes the data sources, sizes, training and testing partitions, and any cleaning or pre-processing procedures applied to each set. Any cases where training and testing partitions from different data sets are combined to train a model will be detailed in its respective methods section. Table 3.1 identifies the source of AMP and non-AMPs observations for all data sets and highlights any overlaps. The protein sequences for all data sets are provided in FASTA format in Appendix A.

3.1 Fernandes et al. (2012)

The Fernandes data set consists of 115 AMP and 116 non-AMP peptides ranging between 10 to 100 AA in length and is provided in the supplemental material of Fernandes et al. in [47]. Due to its small size, this set lacks separate training and testing partitions and is used either as a training or testing partition on its own, or in the context of cross-validation. All AMP observations share $\leq 50\%$ sequence identity and cover a variety of known AMP classes. Sequences were selected from the APD [18] based on if they have a corresponding structure in the Protein Data Bank (PDB) [16]. The set of non-AMPs observations are also taken from the PDB and share the same sequence identity and length cutoffs. Fernandes and colleagues performed screening with the Phobius server [95] to restrict all sequences to intracellular proteins. Further details on this set can be found in [47].

3.2 CAMP Database

To compliment the Fernandes data set when it is used as a training partition, the 'CAMP Database set' provides an accompanying testing set of 216 AMPs obtained from the CAMP database [44]. Similar to the Fernandes set, AMPs are filtered to include only those between 10 and 100 AA in length and sharing $\leq 50\%$ sequence identity. To provide non-AMP observations we pair these with a subset of 145 non-AMPs taken from the Xiao data set [41] (described below) which share the same length and sequence identity cutoffs.

3.3 Xiao et al. (2013)

The Xiao data set consists of separate training and testing partitions taken from the supplemental material provided by Xiao et al. in [41]. The original training partition contains 770 AMPs and 2405 non-AMPs while the original testing partition has 920 AMPs and 920 non-AMPs. The negative examples are taken from the UniProt database [96] and pairwise sequence identity is limited to $< 40\%$. To prevent the selection of extracellular peptides, the UniProt keyword 'cellular location' was set to 'cytoplasm' when extracting the non-AMP examples. Further details on this data set can be found in [41].

An additional cleaning step is applied to this data set by using BLAT [97] (with default settings and the *-prot* flag for a protein vs. protein search) to map all AMPs found in the APD to the the non-AMP training and testing partitions. Sequences with any positive matches are removed and the number of training non-AMP examples is reduced to 2368 while the testing non-AMP set is reduced to 897 observations. The size of the AMP examples remained unchanged.

3.4 APD Gram-based Sets

The APD Gram-based data set consists of AMPs taken from the APD in January 2014. These are split into three separate subsets based on bacterial specificity found using the

'Database Search' page. Sequences in each individual set were removed if they shared $> 70\%$ shared sequence identity using the CD-HIT program [98]. The Gram-negative (GN) set consists of 128 AMPs only active against Gram-negative bacteria and found by selecting the 'GRAM- ONLY' options under the 'Antimicrobial Activity' category. Similarly, the Gram-positive (GP) set consists of 271 AMPs only active against Gram-positive bacteria using the 'GRAM+ ONLY' option. A mix of 1103 AMPs active against both GP and GN is also selected using the 'GRAM+/GRAM-' option. Similar to the CAMP Database set described above, non-AMPs from the Xiao et al. data set are used to pair negative observations.

3.5 DBAASP Database *E. coli* vs. *S. aureus* Set

The DBAASP database [29] set consists of AMP sequences with a recorded MIC value against both *E. coli* (DBAASP Target Species ID 3232) and *S. aureus* (DBAASP Target Species ID 3164) as of July 2015. Peptides are selected to have lengths between 10 and 100, no unusual AA's, and MIC responses against both taxa units of μg/mL or μM with the later converted to μg/mL using (Molecular Weight x μM)/1000. Average molecular weights for conversions are obtained using the `Pepstats` program from EMBOSS Vr. 6.6.0 [99]. We select peptides based on MIC responses as other measurements (e.g. EC50, IC50, etc.) have far fewer observations available. Sequences in this set are not reduced by identity as we are interested in capturing changes in MIC that can occur from differences of just a few AA. AMPs with multiple MICs reported for the same bacteria, are assigned the median MIC value. Responses are then converted to a class label of 0 if activity reported against *E. coli* was lower (better) than for *S. aureus* and less than 50μg/mL. A class of 1 is assigned if the same is true in the reverse case. All peptides with ties are removed. This process results in a total of 158 samples with 62 sequences considered more effective against *E. coli*, and 96 considered more effective against *S. aureus*. To balance the classes, 62 of the 96 *S. aureus* observations were randomly sampled using R for a final total of 124 AMP observations with

Table 3.1: Overview of Data Set Sources

Source Database:	APD	CAMP	DBAASP	PDB	Uniprot
Data Set Name					
APD [18]	◊				⧫
CAMP [44]		◊			⧫
Fernandes [47]	◊			◊†	⧫
Xiao [41]	◊				⧫
DBAASP			◊		
Annotation	◊	◊			

◊ AMPs — ⧫ Non-AMPs

† All APD entries are also present in the PDB

62 members of each class.

3.6 Server Annotation Set

The server annotation set consists of AMPs taken from the APD and CAMP databases and is used to search proteomes for potential AMP-like segments with the predictive web server introduced in Chapter 8. A total of 2564 AMPs from the APD and 2006 experimentally-verified AMPs from the CAMP database were downloaded in July of 2015. After removing redundant entries, the list of sequences consists of 3061 unique AMPs with lengths ranging from 2 to 517. To better match the above training and testing data sets, peptides with lengths < 10 and > 100 AA were removed. To prevent over-represented AMP classes in databases from dominating the search process, we use CD-HIT to remove sequences sharing $> 70\%$ sequence identity. Additional sequences with > 10 consecutive G residues and those listed as "putative" in the APD were also removed, resulting in a final data set of 1367 AMPs.

Chapter 4: Building a Classifier for AMP Activity: Considering Feature Interactions

4.1 Introduction

With this chapter we begin the process of building a predictive model for AMP recognition. We seek to build a comprehensive set F of p physicochemical features $F = f_1, \ldots, f_p$ which we can use with an appropriate model to predict if a query peptide possesses antibacterial activity. In later chapters, this will become a non-trivial task as we add new features and encounter $p >> n$ due to the limited availability of AMP examples. However, in this chapter we start with fewer features as we are focused on the question of feature interactions. In particular, can interactions between physicochemical features really help us improve recognition performance? To answer this question, it is important to have a model that is easily interpretable and capable of handling higher-order interactions. Accordingly, we begin with 8 features first introduced in the context of AMP recognition in [46] and further described below. Work by Fernandes et al. in [47] also used these features and determined that 1 of them was not significant with their data set. As we use the same data set here (described in detail in Chapter 3 Section 3.1), we try to reconfirm their findings by subjecting all 8 features to a randomization test. To build our predictive model we use binary LR, as it is interpretable and implementations in the R programming language [100] easily allow one to include interactions and perform model evaluation.

Work in this chapter was originally presented in [48] and [49] and was co-authored with Dr. Elena Rantou and Dr. Amarda Shehu. The idea of exploring feature interactions come from all authors. Coding responsibilities were shared. ER was responsible for the idea of exploring interactions using LR, the permutation test and the deviance reduction test. DV

was responsible for selecting the data sets and features, conducting biological analysis and interpretation and coding an accompanying web server to make models accessible.

4.2 Methods

We first describe the data set and our initial 8 features from [46]. This is followed by a detailed description of the randomization tests, model construction, selection, and evaluation procedures employed. All performance measurements are used in the context of binary classification as described in Chapter 2 Section 2.5.

4.2.1 AMP Data Sets Used

In this chapter, we use the Fernandes, CAMP Database, and Xiao data sets which are detailed in Chapter 3. The feature randomization tests and model training are performed using the 115 AMPs and 116 non-AMP peptides of the Fernandes set. Testing is performed using the 216 AMPs in the CAMP set paired with 145 non-AMPs from the Xiao set. We choose not to use the testing AMPs provided in Xiao as they are taken from the APD and overlap with our training observations (see Chapter 3 Table 3.1).

4.2.2 Peptide Features Employed

Models in this chapter utilize whole-peptide or "global" features which are features calculated for each amino acid position and then averaged over the length of a peptide. We start with 8 well-studied physicochemical features: α-helix, β-sheet, β-turn and *in vitro* aggregation propensities calculated from the Tango server [101]; *in vivo* aggregation propensity calculated from the AGGRESCAN server [102]; isoelectric point provided by the ExPASy server [103]; hydrophobic mean (based on the GRAVY scale [104]); and peptide length (number of amino acids). These features are a result of wet-laboratory insight obtained from years of experimental research. Many of these features capture known structural propensities among AMPs. For example, aggregation propensity relies on observations that AMPs

may aggregate at the surface of bacterial membranes to aid in membrane insertion [32,105]. Others, such as hydrophobic mean, relate to how many of these peptides are attracted to the surface of bacterial membranes [13]. Further details about these features are available in [46].

4.2.3 Randomization Tests

The significance of the 8 features described above has been previously discussed and measured in [47] through independent sample t-tests. The motivation being that mean values for AMPs and non-AMPs should be significantly different from each other if a feature allows separation between AMPs and non-AMPs. However, there are a number of assumptions made when employing t-tests. For example, in order to correctly employ a t-test, the data should be randomly sampled from a normal population. Also, the positive and negative data sets should have equal population variances. These assumptions are, in this case, translated to assuming equal variances for the populations of AMPs and non-AMPs. It also assumes the data was selected from independent random samples for both groups. However, the fact that in nature AMPs represent only a small fraction of the full spectrum of peptides makes the above assumptions questionable.

Randomization tests provide a rigorous, non-parametric technique to determine significance. They are based on the idea that, if the populations of AMPs and non-AMPs do not differ, then all possible permutations of the observations are equally likely. A standard randomization test is typically conducted as follows. For each feature, one generates N shufflings or permutations. Let the original sample be of size n, where the first q values are AMPs and the next $n - q$ values are for non-AMPs. In this work, N is set to be 10000, q is the size of the positive training observations (115 AMPs) and $n - q$ is the size of the negative training observations (116 non-AMPs). The mean difference between AMPs and non-AMPs is then calculated over each of the N permutations separately. The result of this process effectively reconstructs the sampling distribution for the mean difference.

This raises the following question: if the two populations do not differ, what is the

probability that the observed (real) mean difference is a typical value from this distribution? This probability represents the p-value of the test. Let μ^+ and μ^-, denote the means of the AMP and non-AMP distributions, respectively. The difference of the two means from the observed training data set can then be represented by $D = \mu^+ - \mu^-$. For the i^{th} permutation, where $1 \leq i \leq N$, one derives a mean difference D_i. We also record m, the number of times that $|D_i| > |D|$. Accordingly, we can use m/N to obtain the p-value of the randomization distribution of the mean differences.

We employ this procedure to obtain a p-value with each of the 8 features separately. If the p-value for a feature is < 0.05, then the observed difference in the training data is an atypical result from this distribution. This is interpreted as strong evidence against the hypothesis of identical populations between AMPs and non-AMPs. In other words, the feature is significant and has discriminatory power to separate AMPs from non-AMPs. Such a feature is retained in our methodology for the purpose of model construction, which is now detailed below.

4.2.4 Model Construction and Selection

Using the features detailed above, we can build a predictive model via binary LR (described in Chapter 2 Section 2.6.3) to predict the probability that a query peptide is an AMP. LR models were chosen due their ease of implementation, interoperability, and ability to encode either a single feature or multiple features which may, or may not, include interactions.

Our main concern when selecting features for a model is which, if any, of them increase the probability of an individual peptide being an AMP. The features can be thought of as acting individually, or interacting together as pairs, triplets, etc. In order to use the best features, we start by assessing the performance of individual features (as their own model) in addition to all features together as an additive model without interactions. From there, we can make strategic decisions about including or removing features which are clearly helping or hurting our prediction accuracy. This becomes particularly helpful when we begin the more computationally expensive task of including interaction terms.

As detailed in the results section below, two features appear to be of foremost importance to predicting AMP activity so we fix them to our models. We further restrict our attention to a total of 4 features with a maximum of 4-way interactions in order to control the dimensionality of the parameter space. While it may be appealing to try all possible interactions, it is important that one controls the number of resulting predictors. Encoding all interactions among K selected features results in a model with $2^K - 1$ variables. Great care needs to be exercised so that this number does not exceed or compare to the size of the training data set. As we limit ourselves to 4 features, the maximum number of variables considered in a single model is 15 predictors ($2^4 - 1$). The web server mentioned in the chapter introduction only considers 2-way interactions so that the number of variables does not exceed the size of the training data set.

A standard step-iterative process, implemented through the step function in R, is used to rank and simplify each of these maximal models based on AIC score. In determining a top model we also take into consideration Brier score (described along with AIC in Chapter 2 Section 2.5) and residual deviance which we detail next.

4.2.5 Classification with a Binary Response Model

Once a best model is determined, it is not straightforward as to how to employ it for the purpose of classification. Recall that models here associate a probability from $0 - 1$ with an unseen sequence. A high probability means the peptide is closer to being an AMP than a non-AMP, but where does one draw the cutoff? A typical default assumption would be 0.5, however, we prefer to address this question in a non-parametric way that does not depend on the distribution of probability values observed over the positive and negative observations.

The non-parametric technique we employ proceeds as follows. Let $p_1, p_2, \ldots, p_n$ be the ordered values of the probabilities estimated by the best model for each of the n sequences in the training data set. The goal is to find a change point p_k, where $1 \leq k \leq n$, that separates the probabilities into two groups: $p_1, p_2, \ldots, p_k$ and $p_{k+1}, \ldots, p_n$. In our case,

the group with lower probability values corresponds to non-AMPs, and higher values to AMPs. If we consider μ to be the mean of the predicted probabilities, we can then define the deviance as: $DV = \sum_{i=1}^{n} (p_i - \mu)^2$, where n is the sample size and $i \in 1, \ldots, n$. When the set of probabilities is divided into two groups, the sum of the deviance for these two groups is always less than, or equal to, the deviance of the entire set [106]. Each possible threshold produces a deviance reduction: $RD_k = DV - (DV_{\leq k} + DV_{>k})$, where DV is the deviance for the entire set, $DV_{\leq k}$ is the deviance for the sequence $p_1, \ldots, p_k$ and $DV_{>k}$ is the deviance for the sequence $p_{k+1}, \ldots, p_n$ for $i = 1, \ldots, n$. The threshold probability is then the value p_k that maximizes the deviance reduction RD. This non-parametric procedure was first introduced by Qian and colleagues in [106] in a different context, but we generalize it here to evaluate a binary response model based on LR in the context of AMP classification. We do so in order to compare the best model with other machine learning techniques for AMP recognition, as detailed in our comparative analysis in Section 4.3.6 below.

4.2.6 Implementation Details

As detailed above, 15 pilot models are constructed, each encoding all possible single predictors among the selected four features and their 2-way, 3-way, and 4-way interactions. R's step function is used to simplify each of these models. Taken together, the construction, simplification, and evaluation of models took less than 5 minutes using a single core of an Intel i5 2.5Ghz CPU with 4GB of RAM. Our implementation of the methodology and analysis detailed was done with R. The web server implementation (no longer available) used additional CGI code written in Perl.

4.3 Results

Before proceeding to relate results on the statistical significance of each of the 8 features, we show first the separation power of a simpler linear LR model that uses all features with the Fernandes data set. We then consider the significance results of individual features based

on the randomization tests and then move on to consider only models based on statistically significant features. We begin with performance results using an additive model using no interactions and then take a more exhaustive view and systematically search the feature space as described above using up to 4-way interactions between features. Analysis of the top models, detailed below, shows that interactions allow for better performing models and that peptide length and *in-vitro* aggregation are consistently found to play a central role in predicting AMPs. Discarding these features results in a reduction of the model's predictive power.

4.3.1 Separation of AMPs is Possible with Global Features

We start by showing the separation power of a simple linear LR model that uses all 8 features with no interactions. As expected from the already demonstrated classification power of an ANN-based method that uses all 8 features in [47], the features allow our linear model to separate AMPs from non-AMPs using the same data set. Figure 4.1 illustrates this by showing the predicted value of the response variable y_i for each of the peptides in our data set, where $i = 1, \ldots, 231$ as described above in Methods. Values for the known AMPs in the data set (peptides corresponding to $i = 1, \ldots, 115$) are drawn on the left in blue, whereas those for known non-AMPs in the data set (peptides corresponding to $i = 116, \ldots, 231$) are drawn on the right in black. Figure 4.1 shows that the majority of AMPs have predicted response values above 0.75, whereas the majority of non-AMPs have predicted response values around 0.0.

Next, we proceed to measure the usefulness of each feature through the randomization tests and see that not all 8 features are equally useful. We also demonstrate that more powerful models can be built when considering not all 8 features but only those that are statistically significant.

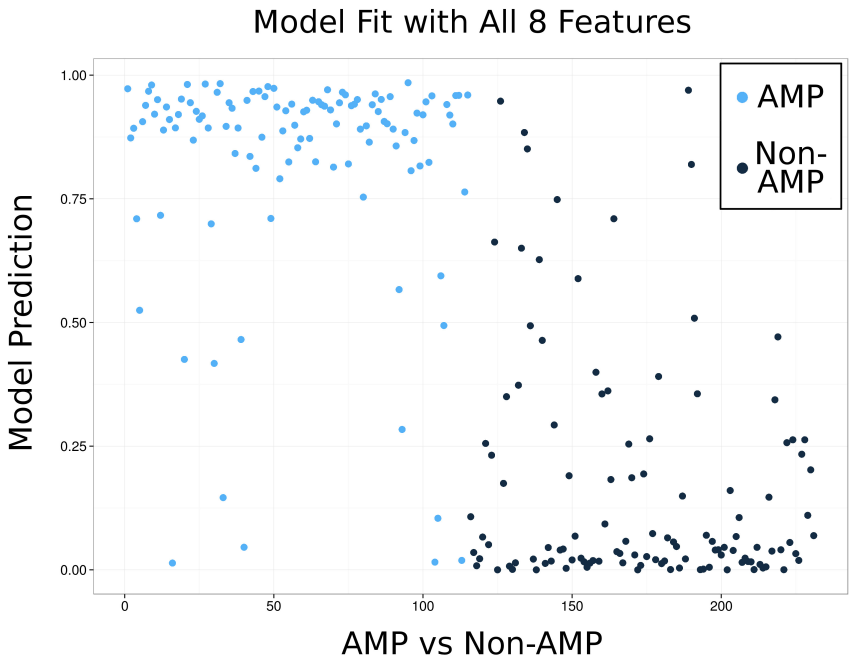

Figure 4.1: Model fitness using all 8 features is shown for peptides in the data set. The first 115 sequences are AMPs, and values for the predicted response variables are drawn in blue. The last 116 sequences are non-AMPs, and predicted response variables are drawn in black. Separation of AMPs from non-AMPs demonstrates that a collective model with all 8 features is viable. However, improved performance can be achieved through the procedure we describe above in Methods and evaluate below.

4.3.2 Individual Feature Significance

The procedure described above in section 4.2.3 generates p-values for all 8 features. We recall that these features, first presented in [46], are peptide length, isoelectric point, hydrophobic mean, β-sheet propensity, α-helix propensity, β-turn propensity, *in vitro* aggregation, and *in vivo* aggregation. For each of these features, the randomization results verify the results of the t-tests previously performed on this data set in [47] as one feature is shown to not be statistically significant.

Table 4.1: Randomization p-values for the 8 features from [46]. All features except β-turn propensity are highly significant. Applying Bonferroni [107] or Benjamini & Hochberg [108] p-value adjustments cannot be observed within 4 decimal places.

Physicochemical Feature	**P-Value**
1. Isoelectric Point	0.0001
2. Peptide Length	0.0000
3. β-Turn Propensity	0.2396
4. β-Sheet Propensity	0.0000
5. Helix Propensity	0.0000
6. *In vitro* Aggregation	0.0000
7. *In vivo* Aggregation	0.0000
8.Hydrophobic Mean	0.0000

The p-values obtained from the randomization tests for each of the 8 features are shown in Table 4.1. These results suggest that, with the exception of β-turn propensity, the other p-values make it highly unlikely for the observed accuracy to be obtained by chance given that there is no significant difference between the means of AMPs and non-AMPs. The remainder of the results use the 7 features with low p-values as model features to predict the probability of a peptide having AMP activity given the significance level of each feature.

4.3.3 Model Evaluations and Building a Model with 3 Features

We first examine the multiple LR model that includes all 7 features but without cross-interactions. Parameters of this model are listed in Table 4.2. In this model, peptide length and *in vitro* aggregation have highly significant coefficients (at the 1% significance level). The β-sheet propensity is shown to be significant at the 10% level. In the case of a LR model, the value of the estimated coefficient gives the change in the value of the predicted probability, when the independent variable increases by one unit of measurement. For instance, the coefficient of -0.0775 obtained for the peptide length feature in this model means that, when the length increases by one amino acid, and all the other features are kept constant, the probability of AMP activity decreases by 0.0775.

Table 4.2: Estimated coefficients and p-values for each of the 7 independent variables/features used in the all-features model.

Model 1. Features	**Estimated Coefficient**	**P-Value**
1. Isoelectric Point	0.0410	0.7023
2. Peptide Length	-0.0775	1.65e-05
3. β-Sheet Propensity	0.0009	0.6382
4. Helix Propensity	0.0082	0.0688
5. *In vitro* Aggregation	-0.0029	1.20e-06
6. *In vivo* Aggregation	0.0030	0.9171
7. Hydrophobic Mean	-0.8816	0.3365

Three other simple logistic models, each consisting of a single feature, are found to have good performance: peptide length, *in vitro* aggregation, and β-sheet propensity, respectively. Parameters of each of these models are listed in Table 4.3. In addition, predicted probabilities from a model can be plotted against the independent variable along with $(1-\alpha)\cdot 100\%$ confidence intervals. Two of the models show the probability of AMP activity to decrease as the value of the predicting feature increases (data not shown). This is in

agreement with the negative values obtained for the estimated coefficients in these models (namely, length and *in vitro* aggregation). Figure 4.2 illustrates this point on the model using peptide length as its only independent variable, AMP-activity can be seen as less probable as length increases. This is an important observation and corollary with known biological observations. AMPs tend to be short peptides compared to non-AMPs [18].

Table 4.3: Estimated coefficients and p-values for each of the three single-feature models.

Model Feature	**Estimated Coeffs.**	**P-Value**
Peptide Length	-8.7e-02	2.1e-7
In vitro Aggregation	-6.1e-03	1.8e-7
β-Sheet Propensity	8.6e-03	0.0351

Using R, the three features above are tested for possible cross-interactions through automated model selection. "Model 1" is produced when the three features are combined with a two-way interaction term between peptide length and *in vitro* aggregation. Parameters of this model are listed in Table 4.4.

Table 4.4: Estimated coefficients and p-values for Model 1

(Best) Model 5. Features	**Estimated Coeffs.**	**P-Value**
Peptide Length	-8.65e-02	2.13e-07
In vitro Aggregation	-6.145e-03	1.80e-07
β-Sheet Propensity	8.63e-03	0.0351
Length $*$ *in vitro* Aggregation	5.437e-03	0.0036

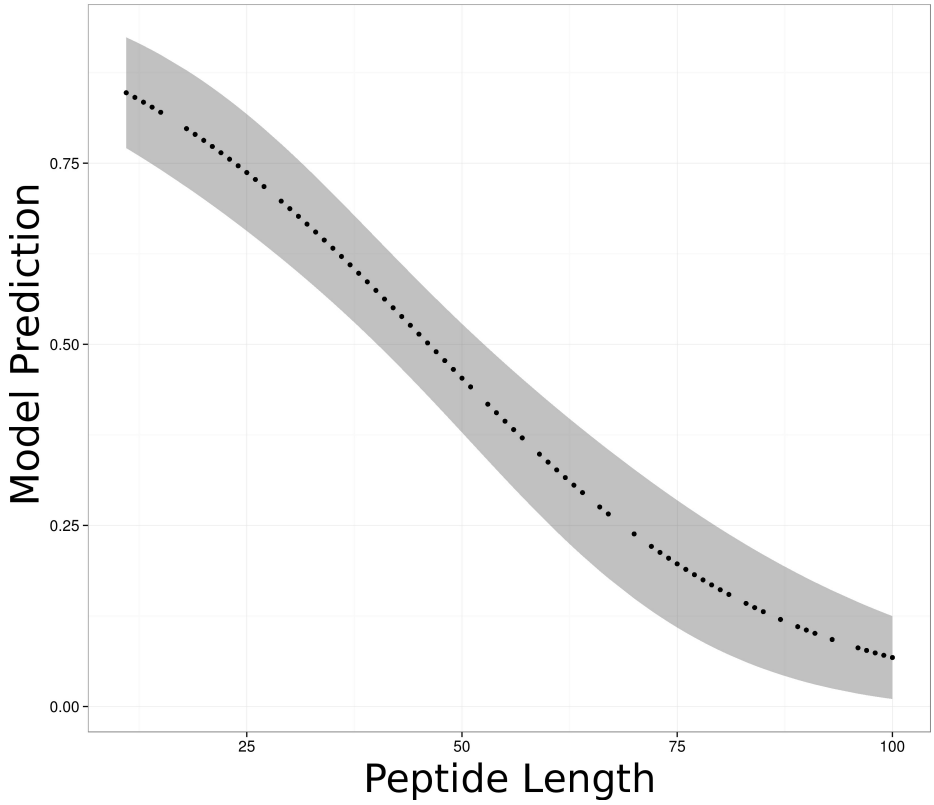

Figure 4.2: Probability of activity decreases with length.

4.3.4 Model Construction with 4 Features

We now investigate all possible interactions among 4 selected features, resulting in pilot models of $2^4 - 1$ predictors or variables. Based on our results so far, we know length and *in vitro* aggregation to be very important in AMP recognition so we fix these as 2 of the 4 considered. The step-iterative process in R is again used and removes predictors from the model if they cause a notable reduction in AIC [109]. From now on, we refer to the top 3 resulting models (in terms of lowest AIC), which we analyze in great detail and denote as Model 2, Model 3, and Model 4 to compliment our initial Model 1.

Table 4.5: Features and interactions for the four models considered on the training data set.

Model	**Features**
Model 1	length + *in vitro* aggregation + β-sheet propensity + (length * *in vitro* aggregation)
Model 2	length + *in vitro* aggregation + hydrophobic mean + (length * *in vitro* aggregation) + (*in vitro* aggregation * hydrophobic mean)
Model 3	length + *in vitro* aggregation + hydrophobic mean + *in vivo* aggregation + (length * *in vitro* aggregation) + (hydrophobic mean * *in vivo* aggregation)
Model 4	length + *in vitro* aggregation + β-sheet propensity + hydrophobic mean + (length * *in vitro* aggregation) + (hydrophobic mean * *in vitro* aggregation)

4.3.5 Selecting a Top Model

Table 4.6 summarizes model diagnostic measures using AIC, residual deviance, and Brier score for all shown models. Recall for all three measures, a smaller value equates to a more reliable model. The model that is selected as best has the lowest AIC, residual deviance and Brier score.

The models evaluated on the training data are additionally compared in terms of ROC

Table 4.6: Comparison of all four models constructed with 4 features. Performance is evaluated in terms of AIC, Residual Deviance, and Brier Score.

Model	**AIC**	**Residual Deviance**	**Brier Score**
Model 1	149.99	137.99	0.0845
Model 2	147.98	135.98	0.0834
Model 3	149.33	135.33	0.0832
Model 4	**147.34**	**133.34**	**0.0813**

curves, shown in Fig. 4.3. Here, the true positive rate and false negative rate are computed as the probability threshold for determining whether a peptide is AMP or not changes from 0.0 to 1.0 in increments of 0.01. Comparison of ROC curves shows that all models have high discriminatory power over AMPs and non-AMPs in the training data set.

Further details are provided on the models in terms of how they separate AMPs from non-AMPs in the training data set. A visual examination is performed via plotting the predicted probabilities for peptides in the data set. Fig. 4.4 plots AMPs in blue and non-AMPs in black.

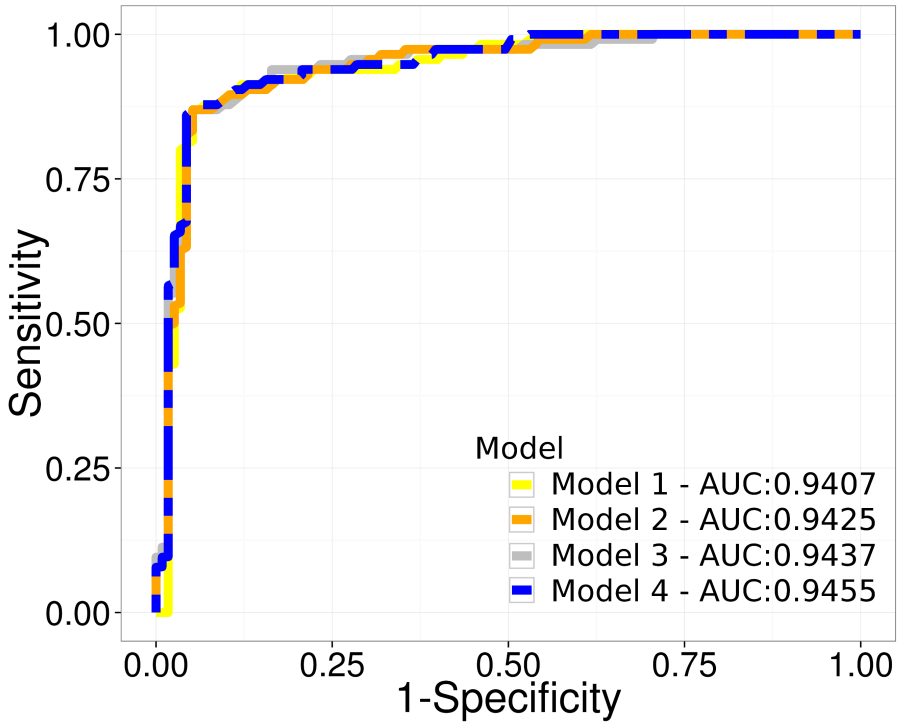

Figure 4.3: ROC curves denoting the performance of Models 1-4 on the training data set from [47]. Area under the curve (AUC) is listed for each reported model.

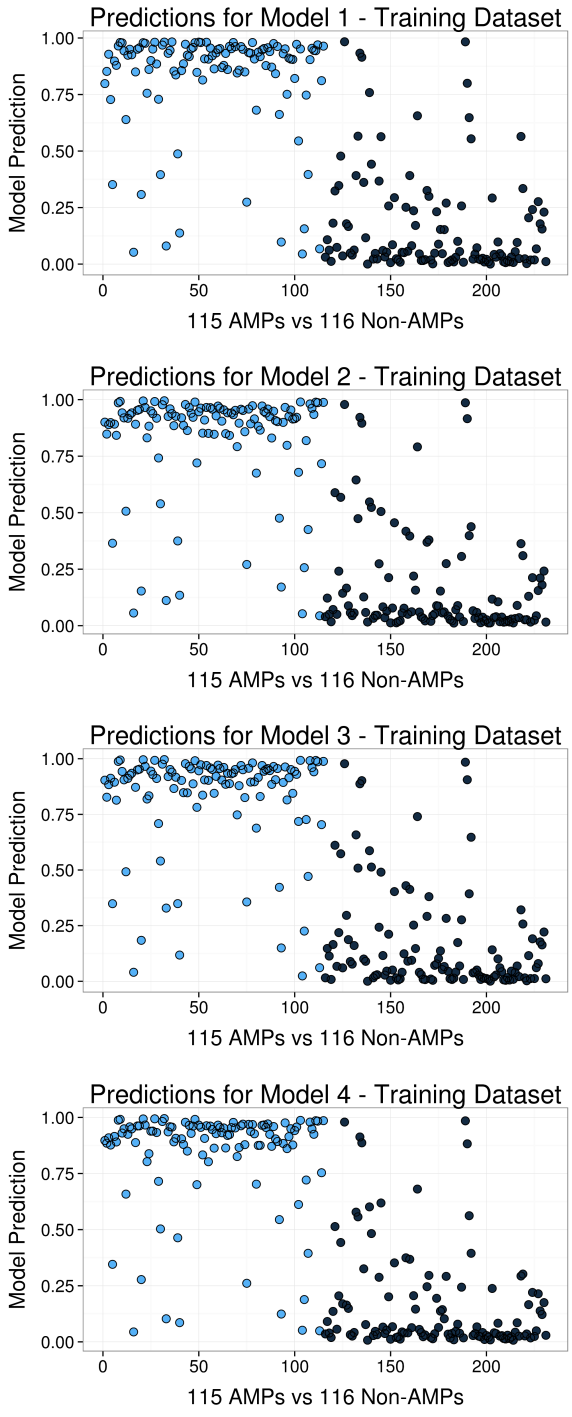

Figure 4.4: Predicted probabilities from each of the models on the training data set. AMPs are shown in blue and non-AMPs in black.

Table 4.7: Mean and median of predicted probabilities for AMPs and non-AMPs for all 4 models.

Model	**AMP Mean**	**AMP Median**	**Non-AMP Mean**	**Non-AMP Median**
Model 1	0.8259	0.9210	0.1726	0.0563
Model 2	0.8294	0.9198	0.1692	0.0586
Model 3	0.8290	0.9157	0.1695	0.0681
Model 4	0.8331	0.9270	0.1654	0.0507

Additionally, table 4.7 shows the mean and median values over predicted probabilities, separately for the positive and negative training observations. All models associate high mean and median probabilities with AMPs and low mean and median probabilities with non-AMPs.

Models are now compared in the context of classification performance, comparing the models in terms of ACC and MCC values. In order to do so, a probability threshold is first determined for each model through the non-parametric technique detailed in section 4.2.5. The non-parametric technique reports thresholds slightly above 0.5 for each model. Fig. 4.5 shows that the probability density functions have long tails. While there is significant difference between the means or medians, the long tails cause the non-parametric technique to report a probability threshold of 0.5. This threshold is not to be interpreted as a fair coin model, as the interpretation is limited to actual distributions. For instance, for all models, more than 89.66% of the non-AMPs lie below 0.5, and more than 89.57% of the AMPs lie above. The effect on classification performance on an independent test set can be seen in Table 4.8, which shows that all models have high sensitivity, specificity, accuracy and MCC.

4.3.6 Comparative Analysis on Classification

We now further showcase the classification performance of the best model (Model 4). Details on how the model separates AMPs from non-AMPs and the ROC curve obtained on the

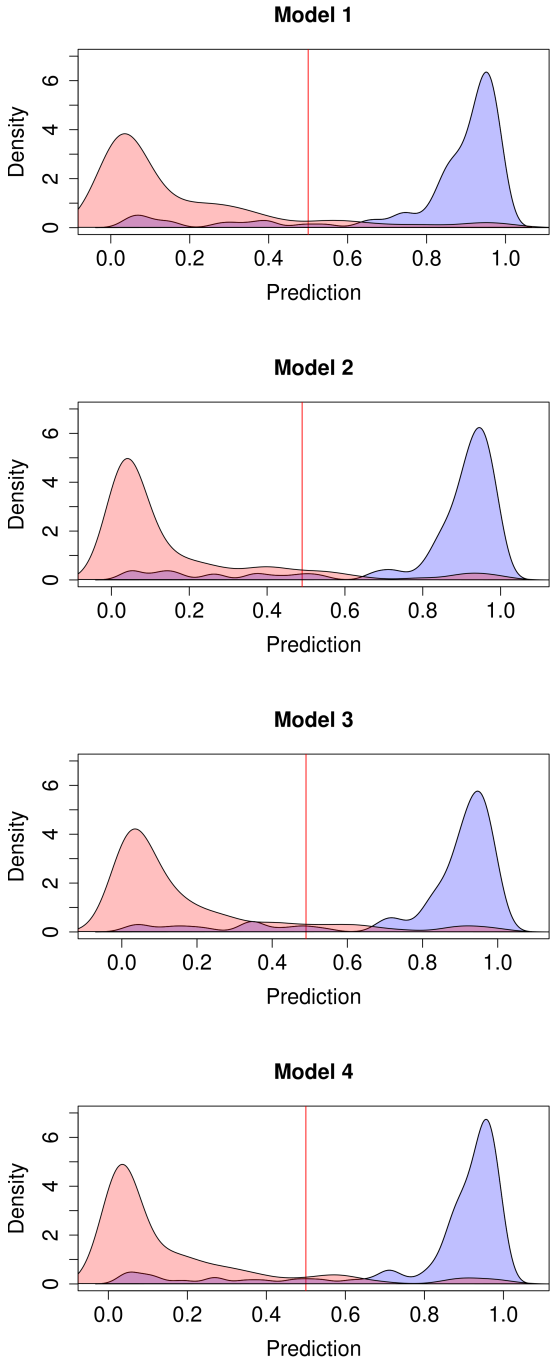

Figure 4.5: Density functions for 4 models on the training data set, with AMPs in blue on the right and non-AMPs in red on the left. The vertical red bar represents the optimal threshold for a prediction being an AMP vs a non-AMP.

Table 4.8: Classification performance of all four models on training data set is evaluated in terms of sensitivity, specificity, ACC and MCC.

Model	**Sensitivity**	**Specificity**	**ACC(%)**	**MCC**
Model 1	0.89	0.90	89	0.7836
Model 2	0.89	0.90	89	0.7836
Model 3	0.88	0.89	88	0.7662
Model 4	0.90	0.90	90	0.7922

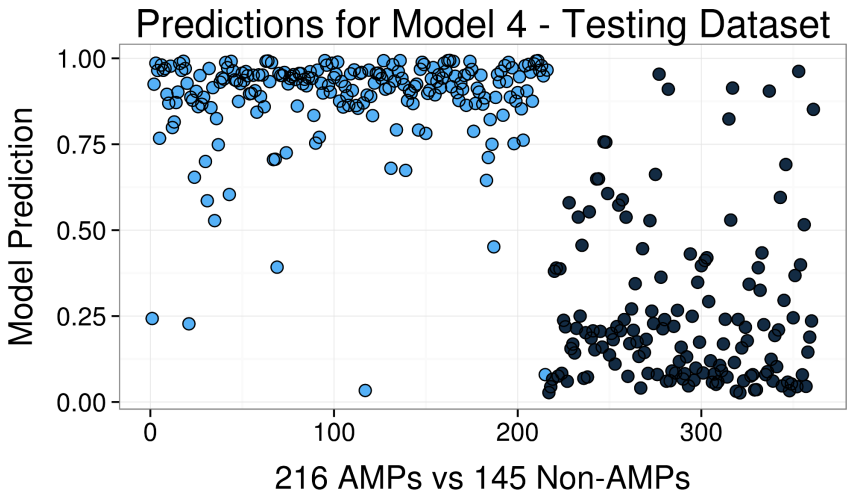

Figure 4.6: Probabilities predicted by model 4 on testing data set. AMPs are drawn in blue and non-AMPs in black.

testing data set are shown in Fig. 4.6 and 4.7. MCC values are also reported and compared to other recent machine learning applications for AMPs. Comparison is limited to MCC values reported in the literature. The model proposed here obtains comparable or higher MCC values on the training or testing data set. We note that a higher MCC is obtained on the testing data set than on the training data set, similarly to values reported by the FKNN [41] method. This is due to the fact the testing data set, a combination of peptides from CAMP [44] and [41], appears to be easier for classification than the Fernandes data set employed for training. Inspection of the data sets reveals that there is more separation

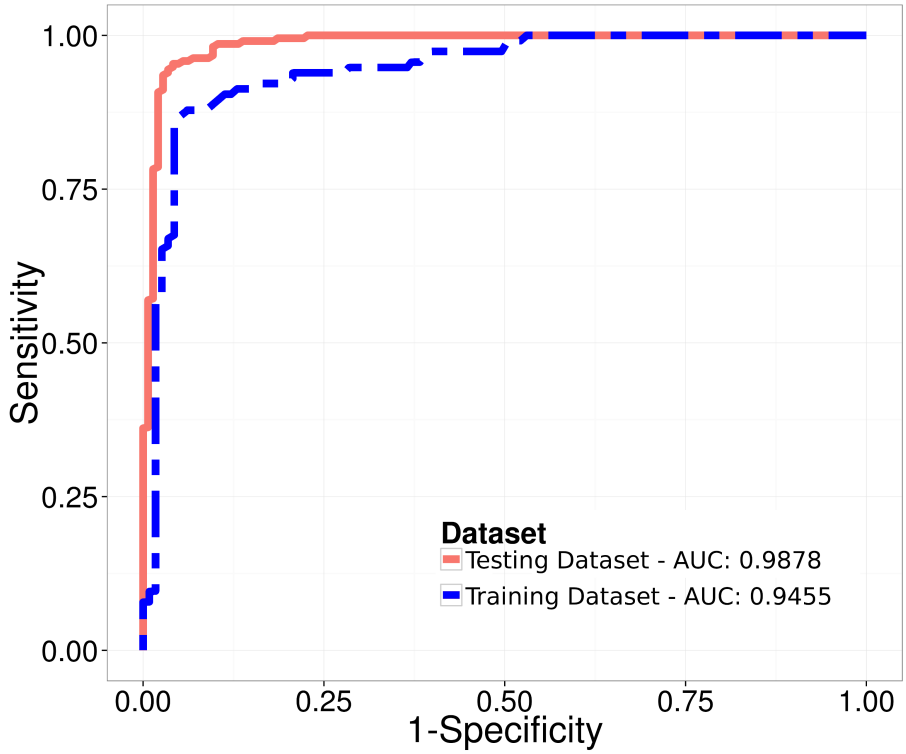

Figure 4.7: ROC curves are shown for the Model 4 applied to the training data set (dashed blue) and the testing data set (solid red). Area under the curve (AUC) is also provided.

between the length distributions of AMPs and non-AMPs in the testing data set, in turn, making AMP classification easier.

4.4 Conclusion and Chapter Summary

This chapter has demonstrated that feature interactions are important and can help improve AMP recognition. It has also presented a general methodology to assess features in terms of relevance for antimicrobial activity and built a LR-based predictive model using statistically-significant features. The usefulness of the methodology has been showcased and validated on global features suggested by experimental studies on AMPs. A predictive model that considers interactions between features is shown to separate AMPs from non-AMPs with

Algorithm	MCC			AMP Database
	Training data set	Validation Data Set	Testing Data Set	
HMM [50]		0.98		AMPer
ANN [51]		0.88		RANDOM
DA [44]	0.75		0.74	CAMP
RF [44]	0.86		0.86	CAMP
SVM [44]	0.88		0.82	CAMP
SVM [87]			0.84	AntiBP2
NNA [40]			0.73	CAMP
SVM [52]			0.80	APD
ANFIS [47]		0.94		APD
ANN [47]		0.85		APD
FKNN [41]	0.73		0.84	APD
LR (Model 1) [48]			0.78	APD
LR (Model 4) [49]	**0.79**		**0.82**	APD,CAMP

Table 4.9: Summary of algorithms and their performance on data sets drawn from different databases. Performance for Models 4 presented in this chapter is given in the last row.

similar accuracy to other machine learning methods for AMP recognition. While all features allow separating AMPs from non-AMPs, length and *in vitro* aggregation appear to be the most important. This finding is in agreement with those reported in [47]. For the features considered in this chapter, we show that the best predictive model is obtained when considering an interaction between these two features. Elucidating which features confer to AMPs their direct antimicrobial activity has general relevance for the computational community. The web server mentioned in the chapter introduction provided researchers with the ability to test their own novel features using our methods. We hope to bring this service back online in the near future when resources become available.

Up until now, we have only considered global "whole-peptide" features. While these have been shown in this chapter as relevant, studies have shown that the position of residues in an AMP can play an important role. For example, Epand and colleagues show in [25] that changing just few residues in the human LL-37 AMP can have a major impact on its ability to induce lipid segregation in a membrane. As different classes of bacteria have different lipid compositions, they demonstrate how these small mutations can essentially switch on or

off the ability for certain peptides to kill Gram-positive and/or Gram-negative bacteria. As such localized differences could easily be missed by features averaged over an entire peptide sequence, in the next chapter we consider features which can capture position-based motifs. Through the use of a powerful evolutionary algorithm, we design novel features which can capture correlations between both neighboring or distal AA positions.

Chapter 5: Building a Classifier for AMP Activity: Considering Distal Features

5.1 Introduction

In this chapter, we introduce a method for novel feature construction and selection to generate novel sequence-based features that are able to capture and encode relationships between distal portions of a peptide sequence. This is motivated in part by detailed biological studies on the behavior and mechanism of action of characterized AMPs which point to the fact that different parts of an AMP sequence may be used for different purposes. Flexible termini may be important to disrupt membranes, and specific hydrophobic regions may serve as anchors to initiate interactions [11, 13]. Based on this biophysical insight, what makes an AMP a potent antimicrobial is probably not just an average hydrophobicity score or the presence of a few short sequence motifs. Therefore, we propose here features that capture the contribution from different parts of an AMP and serve as complex yet transparent descriptors of antibacterial activity. Additional motivation for this method comes from recent work on DNA by Kamath et al. (2012), where they showed features to be more effective at various recognition problems when they contained distal information [110, 111].

The essential goal of these new features is to uncover the underlying "grammar" of AMPs. The basic idea being to construct features of important motifs which captures both position and correlations (shown important in the previous chapter) with other motifs. The hope is to go beyond features based on simple AA composition as in some previous works [44, 46,87,112]. In these latter cases, the only description of a sequence is in the form "it contains these many counts of this k-mer or motif" (where k is the number of consecutive amino acids recorded in a motif). By using motifs as a foundational building block, we design more

informative features using boolean combinations via operators {AND, OR, NOT}. This allows for a grammar-based process (founded upon predicate logic) of feature construction. Motifs and sequence positions play the role of terminals, while boolean operators and other powerful constructs play the role of non-terminals. The representation of such features allows for using an evolutionary algorithm based on Genetic Programming to explore the potentially vast space of such complex features in search of those that discriminate between AMPs and non-AMPs in a supervised classification setting. We refer to this algorithm as EFC for Evolutionary Feature Construction.

Evolutionary algorithms, such as the EFC algorithm proposed here, are particularly effective at searching large feature spaces even when dealing with complex features. If one were to approach this process through other generative models, such as HMMs, the explosion in the number of states and transitions between states would make the HMM unwieldy and its training very difficult, given the scarcity of characterized AMPs.

The method described below pairs the EFC algorithm with the fast correlation-based filter (FCBF) selection algorithm. We use FCBF here, first presented in [113], to reduce the large size of our initially constructed feature set to a smaller but informative one with low redundancy. This is particularly needed in the case of AMPs which only have a few thousand instances to train on [44,114]. When these two algorithms are combined, we refer to it as the EFC-FCBF method. We use a series of experiments to thoroughly show that the EFC-FCBF features offer significant improvements in AMP recognition over the state of the art. Our testing of these features is performed in the context of supervised classification via binary LR as in the previous chapter. More importantly, these new features are more transparent compared to those of the previous chapter and provide an intuitive summary of what they capturing. We hope this can additionally allow for more informative design choices for those designing novel AMPs in the wet laboratory.

This chapter is structured as two experiments. The first is a proof of concept to establish that distal-based features contribute more to AMP recognition than simple composition-based ones using the Fernandes data set. The second experiment employs a more challenging

data set and assess if further performance gains can be found by combining our new distal features with global features of the kind used in the previous chapter. The work presented in this chapter was first introduced in [115, 116] and co-authored with Dr. Uday Kamath and Dr. Amarda Shehu. Code for the EFC framework was written by UK original for DNA analysis. Modifications for AMP recognition were prepared by DV and AS. Selection of data sets and features, feature representation, biological analysis and interpretation was performed by DV. Supplemental material such as all code and data are provided online at: `http://dveltri.com/shehulab/TCBB15_Website`. The full list of EFC features for the models presented in this chapter are available in the Appendix C. We now proceed with the methods used for feature construction and selection using EFC-FCBF.

5.2 Methods

We begin by describing a reduced alphabet used to represent a peptide sequence with our method. We then summarize the EFC algorithm to construct features and the FCBF algorithm to obtain a reduced feature set. Finally, we describe our validation of such features in the context of supervised binary classification via LR and the performance measurements employed. All references to Weka [117], a publicly-available package for machine learning, are for Vr. 3.7.

5.2.1 Representing Peptides with a Reduced Alphabet

EFC builds complex features over motifs or k-mers drawn from a peptide sequence. If a 20-letter alphabet to designate the 20 standard amino acids is used, 20^k k-mers can be constructed. Building more complex features by stacking boolean operators on k-mers results in a combinatorial explosion of the size of the feature space. To reduce the size of this space, we employ a reduced alphabet. In particular, we make use of the GBMR4 alphabet which is composed of only 4 letters and originally proposed in [118] for protein fold assignments. While any 4 unique letters can be selected for the GBMR4 alphabet, we choose to employ A, C, G, and T. Table 5.1 shows the mapping between the letters in this

alphabet to the standard amino acids.

Table 5.1: The four letters in the GBMR4 alphabet employed here are mapped to the standard amino acids, together with a description of what amino-acid properties they capture.

Amino Acid	**Mapping**	**Notes**
ADKE RNTSQ	A	Trends small and for special turns
CFLI VMYWH	C	Non-polar and/or aromatic
G	G	Flexible
P	T	Rigid

5.2.2 Evolutionary Feature Construction

EFC is an evolutionary algorithm originally presented in [110] for DNA sequence analysis. The algorithm makes use of a generalized representation of sequence-based features as Genetic Programming trees. The leaf nodes are k-mers over the GBMR4 alphabet. Here we limit k between 1 and 8. Operators are used to combine these building blocks into more complex features. Four operators are employed: *matches*, *matchesAtPosition*, *matchesAtPositionWithShift*, and *matchesCorrelatingPosition*. This allows for building compositional features (which capture only the presence of a motif anywhere in a sequence), positional features (which capture the presence of a motif at a specific sequence position), position-shifted features (that provide a tolerance upstream and downstream for positional features) and correlated features (which match a position-shifted feature upstream or downstream from another motif), respectively. Boolean operators (AND, OR, NOT) additionally enable the construction of more complex features as illustrated in Figure 5.1.

EFC makes use of the concept of a population, which is a set of feature trees that evolve over a fixed number of generations. The initial population of N features is carefully

constructed to contain a variety of tree shapes with maximum depth D. Rather than keep a fixed population size over each generation, EFC uses an implosion mechanism, reducing the population size by $r\%$ over the previous generation to avoid convergence pitfalls. The top (fittest) ℓ features of each generation are copied into a "hall of fame" set. The hall of fame contributes m features, drawn at random, to serve as parents in the next generation.

The parents are subjected to reproductive operators to obtain child features in a generation. As in [110], both mutation and crossover are employed. The mutation operator is performed with probability p, whereas crossover with probability $1-p$. Bloat, or the growth of overly-complex aggregate features through reproductive operators which do not provide additional gains in discriminatory power, is controlled in parent selection as in [110].

Features in a generation are evaluated (and compared) via a fitness function Fitness(f). The function makes use of a labeled (training) data set of AMPs and non-AMPs as in:

$$\text{Fitness}(f) = \frac{C_{+,f}}{C_+} \cdot |C_{+,f} - C_{-,f}|$$

Here f refers to a feature, $C_{+,f}$ and $C_{-,f}$ are the number of positive (AMP) and negative (non-AMP) training sequences that contain feature f, respectively, and C_+ is the total number of positive training sequences. This fitness function tracks the occurrence of a feature only in AMPs, as non-AMPs may not share relevant features. The fitness function penalizes non-discriminating features (those equally found in positive and negative training sequences).

5.2.3 Filter-Based Feature Selection

After termination of the EFC algorithm, the features in the hall of fame are submitted to a feature selection algorithm to obtain a smaller set of relevant features. The FCBF algorithm presented in [113] is employed for this purpose. The algorithm uses the concept of *entropy* from information theory to maximize the relevance between features and classes

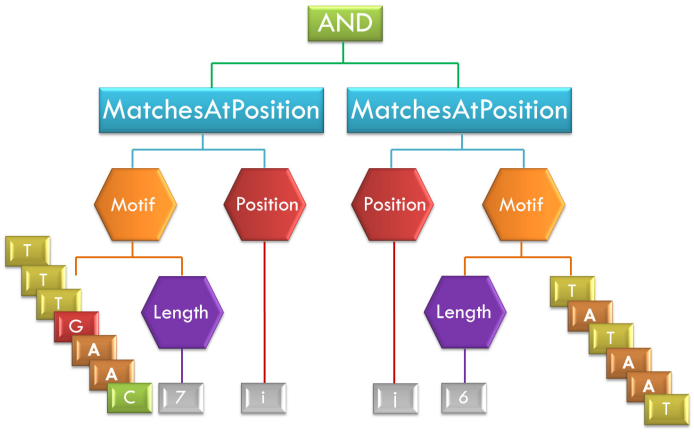

Figure 5.1: This is an example of a conjunctive (correlative) feature which encodes the co-occurrence of two motifs and is an example of features constructed by EFC.

while minimizing correlation amongst features. This provides a set of highly-relevant features with low redundancy. The particular implementation used here is the FCBF option from the publicly-available Weka package for machine learning [117].

5.2.4 Evaluation of Features and Performance Measurements

Selected features are evaluated in the context of supervised classification through Weka's implementation of LR with the regularization parameter set to $1e-8$. We choose to demonstrate results obtained using LR, as LR provides a smooth probabilistic transition between two classes in addition to controlling for over-fitting [69].

The performance of the LR model is evaluated through standard measures in machine learning as described in Chapter 2 Section 2.5.

5.2.5 Implementation Details

All experiments in this chapter were performed on an Intel 2X quad-core machine with 3.2Ghz CPU and 8GB of RAM. EFC is written in Java. Since EFC is stochastic, it is run 30 times per experiment, and average results with standard deviations are reported.

One run of EFC takes about 1 hour of CPU time. The maximum motif length in EFC is set to $k = 8$ in all the EFC runs, as smaller maximal values yielded slightly lower performance. The other parameters in EFC are set as follows: $N = 10,000$, $D = 5$, $r = 10$, $G = 30$, $\ell = 500$, and $m = 100$. The mutation and crossover operators are performed with probability 0.3 and 0.7, respectively. Weka is used to apply FCBF to EFC-obtained features in the hall of fame and select a subset of 40 features after an EFC run. The method is run with *numToSelect*=-1 and using the *SymmetricalUncertAttributeSetEval* option. FCBF typically takes $5 - 10$ minutes of CPU time. The final predictive model is then built using LR, which is also available in Weka.

5.3 Results

We conduct a comparative performance analysis in two distinct experimental settings. The first demonstrates the advantage of employing complex features (capable of capturing both local and distal relationships in a peptide sequence) as opposed to simple composition-based features. Superior performance is demonstrated in the context of 10-fold cross-validation on the Fernandes data set. In the second experimental setting, we use a more recent training and testing benchmark data set from Xiao. Both data sets are described in Chapter 3. We next compare our EFC-FCBF method to several other state-of-the-art methods for AMP recognition available as web servers. After demonstrating comparable performance to some of the top performers, we demonstrate how our results can be even further improved by combining our sequence-based features with physicochemical ones. This specific setting demonstrates how a wet-laboratory researcher could combine our sequence-based features with their additional domain-specific knowledge of AMPs to generate even better predictive models. We conclude this section by examining in detail the biological relevance of the top 10 features obtained by our EFC-FCBF method and providing the first steps into possible employment of the complex features proposed here for discriminating among different mechanisms of action within AMPs.

5.3.1 Proof of Concept Comparison of EFC-FCBF vs. K-mer SVM

Data Set and Experimental Setup

We use here the 115 AMPs and 116 non-AMPs from the Fernandes data set detailed in Chapter 3. All peptides in the training data set are first converted to the GBMR4 alphabet mentioned above. Our EFC-FCBF method is compared on this data set to k-mer SVM. The latter is freely available at the Rätsch Lab Galaxy Server (`https://galaxy.cbio.mskcc.org`) under the "SVM Toolbox." We use the spectrum kernel, together with other default settings, except for the number of cross-validations, which we set to 10. We run the k-mer SVM method with different values of k, ranging between 5-8.

The EFC-FCBF method is applied using a maximal motif length of $k = 8$ (other parameters are set to the values listed above). Peptide sequences are represented as binary feature vectors of 40 dimensions (with a 0 denoting the absence and 1 the presence of a particular feature in a sequence; 40 corresponds to the 40 features selected by FCBF). The LR implementation from Weka is used to train and apply the final predictive model. The entire process of running EFC to obtain a hall of fame, running FCBF to select 40 features from it, and then building an LR model is repeated 30 times (given that EFC is stochastic) to obtain average performance results. We note the features selected in each run remain relatively consistent in rank, with the top 10 not changing across runs. As validation is performed using 10-fold cross-validation, the 30 runs of EFC-FCBF are applied to each fold separately.

Performance Comparison

Performance is reported in Table 5.2 in terms of sensitivity, specificity, MCC, auROC, and auPRC (described in Chapter 2). Average values are reported for EFC-FCBF, with standard deviations shown in parentheses. The results in Table 5.2 show that EFC-FCBF clearly outperforms k-mer SVM on all the performance measurements. In particular, an improvement of more than 14% is obtained on auROC and auPRC.

Table 5.2: Performance comparison using 10-fold cross-validation between EFC-FCBF and k-mer SVM on the Fernandes data set. Bold font is used to highlight higher performance by EFC-FCBF.

Method	Sens. (%)	Spec. (%)	MCC	auROC	auPRC
5-kmer-SVM	75.7	75.0	0.54	0.81	0.79
6-kmer-SVM	74.8	74.1	0.46	0.79	0.79
7-kmer-SVM	73.0	72.4	0.40	0.78	0.78
8-kmer-SVM	73.0	72.4	0.36	0.72	0.70
EFC-FCBF	**87.1** (0.11)	**87.2** (0.12)	**0.76** (0.01)	**0.95** (0.00)	**0.94** (0.00)

These results suggest that the quality of the features obtained by EFC-FCBF is much higher than that of (compositional) spectrum k-mer features. Combining distal information affords higher classification performance. However, it also appears that global features do play an important role for AMP recognition. For example, 10-fold cross-validation on the Fernandes data set using Model 4 from the previous chapter gives a comparable (yet slightly higher) MCC of 0.78 compared to 0.76 for EFC-FCBF. In the next experiment, we show how a dramatic improvement in performance can be achieved by combining both distal and global features together.

5.3.2 Comparing EFC-FCBF to Other Servers and Adding Global Features

Data Set and Experimental Setup

For this experiment we use both the training and testing partitions from the Xiao data set described in Chapter 3. Performance of EFC-FCBF is measured on the Xiao testing data set to four methods (SVM, RF, ANN, and DA) provided as part of the *CAMP* prediction server [44] (`http://www.camp.bicnirrh.res.in/predict`) and to one other method, *iAMP-2L*, provided through Xiao's own server (`http://www.jci-bioinfo.cn/iAMP-2L`). Since neither *CAMP* nor *iAMP-2L* are trained for peptides encoded in the GBMR4 alphabet, the testing set submitted to these methods is left in the standard 20-letter amino acid

alphabet. EFC-FCBF uses the GBMR4 alphabet encoding.

Performance Comparison

Performance is reported in Table 5.3 in terms of sensitivity, specificity, MCC, auROC, and auPRC. Average values are reported for EFC-FCBF over 30 runs, with standard deviations shown in parentheses. For methods which provide continuous prediction values, we report auPRC. Otherwise, "NA" is shown, when methods only report a binary (AMP or non-AMP) prediction. The results in rows $2-7$ in Table 5.3 show that EFC-FCBF outperforms all the learned models provided by the CAMP AMP-Prediction Server on the Xiao testing data set for most of the performance measurements, including MCC, auROC, and auPRC. This is not surprising, as the features employed by these models are a mixture of compositional and physicochemical ones and do not encode distal information. The comparison with the *iAMP-2L* server shows that EFC-FCBF on its own remains competitive but only performs better on the auROC measurement. It is important to note that the features employed by the *iAMP-2L* server combine correlational pseudo-amino acid counts with a fuzzy logic-based algorithm, which explains the closer performance to EFC-FCBF.

Better performance is obtained by our method when physicochemical features are added to the pool of sequence-based ones prior to feature selection by FCBF. The physicochemical features include the 8 global peptide features introduced in Chapter 4. We note the feature "turn propensity" is also retained as it was found to be highly significant for the Xiao training set using both the randomization test detailed in the previous chapter (mean randomized difference: 0.717, mean true difference: 16.173, p ¡ 0.00000) and the Wilcoxon rank sum test with continuity correction (W=450150, p ¡ $2.2E-16$). We also include 299 peptide-averaged physicochemical features extracted from the AAIndex database [119]. This database documents 544 attributes, but only 299 remain after randomly selecting attributes and removing related entries sharing $\pm 80\%$ correlation. These latter features have also been used to classify AMPs through SVM [52].

We designate this setup, when the 307 physicochemical features are included with

sequence-based ones prior to feature selection, as "EFC+307-FCBF" and show its performance in row 8 of Table 5.3. Better performance is obtained by EFC+307-FCBF over *iAMP-2L* overall, including in performance measurements, such as auROC. ROC curves drawn in Figure 5.2 additionally shows EFC+307-FCBF and *iAMP-2L* to be the top two performers. These results demonstrate that there is some orthogonal information in physicochemical features not captured directly in sequence-based ones (possibly lost due to the reduced alphabet), and the best performance can be obtained when combining both.

Table 5.3: Performance comparison on the Xiao testing data set between EFC-FCBF and various methods available online as prediction servers for AMPs. Bold font is used to highlight highest performance on a specific metric.

Method	**Sens. (%)**	**Spec. (%)**	**MCC**	**auROC**	**auPRC**
CAMP SVM	95.8	39.8	0.43	0.64	0.53
CAMP RF	97.1	33.5	0.40	0.73	0.76
CAMP ANN	89.1	70.9	0.61	0.80	NA
CAMP DA	94.1	49.5	0.49	0.81	0.76
iAMP-2L	**97.7**	92.0	**0.90**	0.95	NA
EFC-FCBF	92.1 (0.70)	90.0 (2.30)	0.73 (0.07)	0.96 (0.00)	0.95 (0.00)
EFC+307-FCBF	92.4 (1.10)	**96.1** (0.20)	0.86 (0.02)	**0.98** (0.00)	**0.98** (0.00)

5.3.3 Information Gain Analysis

We now provide a more detailed analysis of the top 10 features consistently selected by FCBF over 30 different halls of fame (independent runs of EFC, where constructed features are evaluated over the Xiao training data set, adding the physicochemical features prior to feature selection). Table 5.4 shows the IG of these features over the Xiao testing data set.

Features with rank 2 and 6 in Table 5.4 reproduce discovery made by computational and wet laboratory studies [13, 46, 47]. Charge (the feature with rank 2) is considered to be important for attracting AMPs toward their target bacterial membranes [12, 13]. It is also

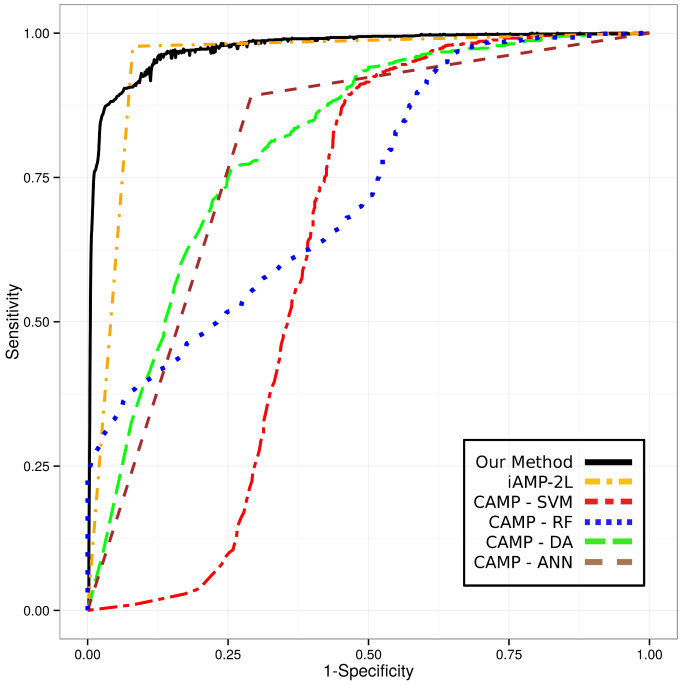

Figure 5.2: ROC curves on the Xiao testing set. Here, "Our Method" refers to the combined EFC+307-FCBF features described in this chapter. Curves are plotted using actual predictions for methods that provide them. Since the CAMP ANN and iAMP-2L methods only provide binary predictions, their curves are generated using the ROCR package [120]. Area under the curves are reported for each method in the 5th column of Table 5.3.

Table 5.4: The top 10 EFC+307-FCBF features are ranked here by their information gain, shown on column 2. Additional feature rankings can be found in the Appendix. The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Position-Shift features are motifs found at a residue position within a given tolerance. For example, the top feature shows that 'GGGA' can range from position 34 to 40 (37 ± 3). Global Motifs are patterns which can be found at any position within a peptide. Additional information for AAIndex features is available from the AAIndex database [121]

Rank	Info. Gain	Feature Source	Feature Description
1	0.1965	EFC: Position-Shift	GGGA at position 37 ± 3
2	0.1956	AAIndex: FAUJ880112	Negative charge
3	0.1438	AAIndex: FINA910104	Contribution to helix termination
4	0.1361	AAIndex: YUTK870103	Activation Gibbs energy at pH 7.0
5	0.1201	EFC: Position-Shift	CA at position 53 ± 3
6	0.1161	One of 8 features from [46]	*In vitro* peptide aggregation from Tango Server [101]
7	0.0884	EFC: Position-Shift	CA at position 27 ± 3
8	0.0882	EFC: Global Motif	CCCG at any position
9	0.0812	AAIndex: GEOR030101	Helix linker propensity
10	0.0663	AAIndex: AURR980118	Normalized residue freq. at C" helix termini

thought that aggregation of peptides at the membrane surface (captured in the feature with rank 6) may contribute to many of the pore-forming abilities of helical AMPs [32]. As a major portion of AMPs in both the training and testing sets are helical, it is not surprising that many helix-related features such as those with rank 3, 9 and 10 are also selected in the top 10.

Sequence-based features constructed by EFC, indicated by an "EFC" prefix in Table 5.4, provide novel information. Three of the top 10 features are of the Position-Shift type, which captures the presence of a specific sequence motif about a specific position with some tolerance. Features with rank 1 and 7 capture the C-termini of AMPs. It is interesting to note that the position of the motifs captured in these features indicate the characteristic length of AMPs in the training data set (where average peptide length was 32 amino acids).

More importantly, the feature with rank 1 captures a consecutive segment of flexible amino acids followed by a small amino acid found in special turns. Such a feature, found on the C-terminus, may capture an important biological signal that AMPs use to form pores as they attack the membrane surface [13,18]. The feature with rank 7 captures a non-polar or aromatic amino acid followed by a small amino acid towards the C-terminus. The rank 5 feature captures the same but for longer AMPs, possibly pointing to a biological signal important for the mechanism of action in certain AMPs.

5.4 Conclusion and Chapter Summary

In this chapter we have proposed a new method, EFC-FCBF, for deducing complex, yet easily interpretable, sequence-based features for AMP recognition. An evolutionary feature construction algorithm is employed to generate novel sequence-based features capable of encoding the presence of distal motifs within an AMP sequence. The algorithm is highly appealing and effective for this purpose compared to other generative models that require the explicit specification of structure in the sought data.

The method addresses the possible scarcity of labeled data by selecting highly informative yet non-redundant features. The fast correlation-based filter selection algorithm is

used for this purpose. LR is used to evaluate these features in the context of supervised classification.

The results related in this chapter show that the computed features are highly informative and discriminating. Detailed comparisons with a comprehensive list of state-of-the-art methods show EFC-FCBF to be among the top performers as seen in Table 5.3 above and the general performance summary across various data sets in Table 5.5 below. We further demonstrate that there is orthogonal information in physicochemical features like the kind used in the previous chapter. Including them for selection by FCBF improves the performance of our EFC-FCBF method. This illustrates how a wet-laboratory researcher can combine our sequence-based features with domain-specific knowledge, for example with a specific class of AMPs, to generate better predictive models.

Table 5.5: Summary of EFC-FCBF+307 Performance with other AMP Prediction Algorithms and Data Sets

Algorithm	**MCC**			**Database Source**
	Training Data Set	Validation Data Set	Testing Data Set	
HMM [50]		0.98		AMPer
ANN [51]		0.88		RANDOM
DA [44]	0.75		0.74	CAMP
RF [44]	0.86		0.86	CAMP
SVM [44]	0.88		0.82	CAMP
SVM [87]			0.84	AntiBP2
NNA [40]			0.73	CAMP
SVM [52]			0.80	APD
ANFIS [47]		0.94		APD
ANN [47]		0.85		APD
FKNN [41]	0.73		0.84	APD
LR (Ch.4 Model 1) [48]			0.78	APD
LR (Ch.4 Model 4) [49]	0.79		0.82	APD,CAMP
LR (EFC+307-FCBF) [115, 116]			**0.86**	APD,CAMP

In the next chapter we demonstrate how we can improve classification performance by

recalculating some of the global features used here to represent specific regions of a peptide. In effect, we capture additional information about the N- and C-termini, as well as, the most hydrophobic window contained in an AMP sequence.

Chapter 6: Building a Classifier for AMP Activity: Considering Regional Features to Capture AMP Function

6.1 Introduction

This chapter finalizes the design of our model to recognize AMP antibacterial activity. Our goal is to select a feature set and a classifier with best performance to incorporate into a predictive web server introduced later in Chapter 8. Accordingly, we try to balance high classification accuracy with reasonable performance speeds so that users can upload potentially thousands of query peptides. From previous chapters, we have seen the importance of considering feature interactions and both distal *and* global features to improve AMP recognition performance. We now expand our current feature set by applying some of our global features specifically to functional regions within an AMP sequence. More specifically, we average these features about the N- and C-termini, as well as, an amphipathic window we locate within the sequence. This allows us to create more descriptive features which can aid in the design of novel AMP sequences. For example, rather than simply state: "feature X appears important for AMPs," using these new features we might be able to say: "feature X appears important at the N-termini."

An obvious question arises: why choose the N-termini, C-termini and an amphipathic window for our three functional regions? Support for these choices comes from earlier work on classification, molecular dynamics and biological studies of AMPs. The terminal regions have been shown in previous computational work to be helpful in AMP recognition problems [42, 52, 87, 88]. Torrent and colleagues have also used the concept of a sliding window approach to score and identify segments with potential antimicrobial activity for their

AMPA prediction server [46, 112]. Molecular dynamics studies on AMPs such as alamethicin and melittin suggest the N-terminus plays a key role in initiating membrane pore formation and that when both termini are highly charged, they form important interactions with polar membrane head groups [59, 122]. Many biological studies have truncated or "swapped" the termini between various AMPs resulting in marked differences in antibacterial activity [43, 123–126]. For some AMPs thought to aggregate at the membrane surface (e.g. LL-37), the N-terminus has also been shown important to induce peptide oligomerization [21]. Our additional focus on the amphipathic window stems from studies in the AMP literature which suggest it plays a critical role in initiating AMP-membrane interactions and can also help anchor an AMP inside the bilayer once inserted [11, 13, 18].

We begin this chapter describing the changes to our feature set and how our feature vectors are reorganized to represent the new functional regions just described. We then show how our feature set performs with and without additional feature selection using LR and a variety of other linear and non-linear binary classifiers. Our objective is to pair an informative feature set with a best-performing classifier to implement with our server. We conclude with an analysis of features rankings by two of our top models. One which results in the best overall classification performance and another with a more transparent view of how features and interactions are utilized.

6.2 Methods

We start this section with a general description of features added or removed to our set compared to previous chapters. Additional details for features retained up to this point can be found in Chapters 4 and 5. Next, we describe the three specific segments of a peptide we use to generate our new regional features. Throughout the rest of this dissertation we denote these segments as: NT (N-termini), CT (C-termini) and AS (Amphipathic Segment). We conclude with a brief description of implementation details and parameter settings for the different classifiers employed for AMP recognition. Additional background and theory for

the top classifiers shown in Results (RF and MARS) can be found in Chapter 2 Section 2.6.3.

All references to Weka [117] are for Vr. 3.7 of the program. Calculations for the MARS method are performed using R and the *Earth* (Enhanced Adaptive Regression Through Hinges) package Vr. 4.4.2 by Stephen Milborrow [127]. The package uses a different acronym as the term "MARS" is trademarked and licensed to *Salford Systems* (San Diego, CA) and tables containing results for MARS-based models are labeled as "Earth" from here on.

6.2.1 Changes to the Feature Set

For this chapter we add a number of new features. The first additions are numeric features which describe solvent accessibility surface area (SASA) taken from Creamer et al. in [128]. SASA has been used in other AMP predictors [44] and separate values are used here to describe regions for polar only, apolar only, or both types of resides. Work in [128] provides separate upper- and lower-bounds values for both backbone or side-chain atoms in residues. As the upper- and lower-bound values are highly correlated (Pearson Correlation Coefficient of 0.95 and 0.98 for backbone and side-chain respectively), after a random selection, we use the upper-bound values for the backbone and lower-bound values for the side-chain categories to avoid redundancy. In total, we add 6 new SASA features to our set. The motivation for using these SASA features comes from the importance of AMP side-chain interactions (propagated through surrounding water molecules) with membrane lipid head groups. These interactions can induce detrimental effects to membrane structure and cause leakage through various means such as: disrupting normal lipid packing order, changing membrane curvature, or inducing membrane thinning [24]. We also add 11 new numerical features from the `emboss` Vr. 6.6.0 package [99]. These include basic physicochemical descriptors such as: overall charge, molecular weight, and percentages for the number of tiny, small, aliphatic, aromatic, apolar, polar, charged, basic and acidic residues. The `emboss` package also calculates isoelectric point which we use to replace the separate implementation of this feature from Chapter 4 for improved efficiency.

From the previous chapter, we combine 208 unique binary EFC features (generated

from all different runs prior to applying FCBF) with 299 AAIndex features. From Chapter 4, aside from replacing the source of isoelectric point, we remove the *in vivo* aggregation feature generated from [102]. The rational for this stems from the lack of a stand-alone executable available to calculate the feature without the aid of the AGGRESCAN web-based server. This is a problem as we later need to encode user-submitted data sets into our own features for our own web server implementation. Fortunately, this feature was not a top contender in either Model 4 or our EFC+307-FCBF model in previous chapters. We keep the remaining Chapter 4 features and add the two feature interactions "length `x` *in vitro* aggregation" and "length `x` GRAVY" (recall GRAVY is a measure of hydrophobic moment) found important for Model 4 (see Chapter 4 for details).

Finally, we add 3 new numeric features used in the Heliquest server [129] specifically for the AS segment which we describe in more detail in the next section. While we now have 530 features to describe a peptide in full, we will reapply many of these features to also describe our three new functional regions.

Defining 3 Regions of an AMP

As mentioned above, we designate NT, CT and AS as three functional regions of an AMP. We define NT as the first 10 AA and CT the last 10 AA of a peptide. Accordingly, we now place a minimum length limit on our model and only consider a query peptide ≥ 10 AA in length. If a peptide is exactly 10 AA long, the NT and CT regions are identical. AS is defined as the best-scoring 18 AA window which we detail in the next section. The number 18 is chosen as it is the number of positions in a typical Edmundson projection [130] (also known as a helical wheel diagram) as seen in Fig. 6.1. If a peptide is between $10 - 17$ AA in length, for compatibility reasons we append placeholder "X" characters until the query is 18 AA long, but we do not include these when features are calculated for the segment. We now describe how the AS window is located.

Locating the Amphipathic Window

AS is determined by looking at a "window" of 18 consecutive AA, starting at the N-terminus and sliding one residue at a time until the end of the C-terminus is reached. For each window, we thread the sequence into an Edmundson projection like the one seen in Fig. 6.1 and generate the parameters: mean amphipathic moment ($\widehat{\mu H}$), hydrophobic face length (f), and the net charge of the hydrophobic face at pH=7.4 (z). These are all calculated in the same manner as the Heliquest server [129] as follows:

$$\widehat{\mu H} = \frac{1}{N} \times \left(\left(\sum_{i=1}^{N} H_i \times sin\left(\delta\right) \right)^2 + \left(\sum_{i=1}^{N} H_i \times cos\left(\delta\right) \right)^2 \right)^{1/2}$$

Here, N is the total number of AA in a peptide, H_i the hydrophobicity of the i^{th} AA in the sequence and δ is the angle separating side chains along the backbone (either 100° for α-helices or 160° for β-sheets [130, 131]). There are a number of scales available for hydrophobicity and Heliquest uses that of Eisenberg [132]. Using our code with the Eisenberg scale produced identical results to the Heliquest site for all of the peptides we tested. However, we change our hydrophobicity scale to a newer one provided by Wimley and White in [133] as it is empirically determined for residues at membrane interfaces which better reflects the biological action of AMPs. As in [129], we calculate f by adding up the number of consecutive hydrophobic residues (A,L,M,F,W,Y,I,V) along the face of the wheel diagram and make an exception for G if it is the final or second-to-last position on either end of this face segment. Calculating charge z assumes H as neutral, and is found by adding $+1$ for all K and R residues, and -1 for all E and D residues contained in the face segment. From all of the windows generated along the sequence, we select the one with the highest $\widehat{\mu H}$ value. Since $\widehat{\mu H}$, f and z pertain only to wheel diagrams, we only use these 3 features explicitly with the AS region.

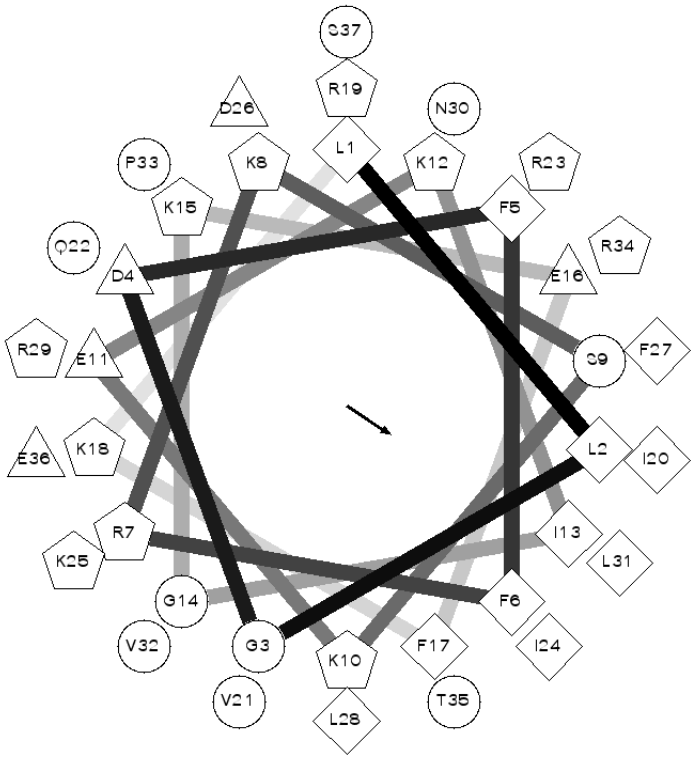

Figure 6.1: An example of an Edmundson projection [130] or helical wheel diagram using the human AMP peptide LL-37 (Uniprot ID: P49913). Hydrophilic residues are drawn as circles, hydrophobic residues as diamonds, potentially negatively charged residues as triangles, and potentially positively charged residues as pentagons. Inside each shape is an AA letter followed by its position in the sequence (e.g. *L1* is a leucine and the first AA of the peptide). A standard α-helix will have an angle of rotation of 100° and repeat every 18 residues. The arrow in the center of the circle represents the direction of the hydrophobic moment and can be thought of as pointing towards the center of a hydrophobic face about the circle. If we were to draw an imaginary line through the center of the circle perpendicular to this arrow, it would be a best attempt to segregate hydrophobic and hydrophilic residues into opposite faces. An amphipathic β-sheet, which alternates between hydrophilic and hydrophobic residues, can also be represented in a similar fashion using a 160° angle of rotation [131]. This figure was created using the *Wheel* Vr. 1.4 program by Don Armstrong and Raphael Zidovetzki (`http://rzlab.ucr.edu/scripts/wheel/wheel.cgi`).

As we do not know the secondary structure of a query peptide, we must make a basic prediction of secondary structure to select a value of δ to use when scoring a window. A number of programs are available to predict protein secondary structure given a primary AA sequence as input. Some examples are: `PSIPRED` [134], `JPred4` [135], and `PredictProtein` [136]. Generally, these programs perform best when they can compare a query sequence against a large database of known sequences [134, 135, 137]. This is both a time and memory-intensive step which severely limits the number of queries a web-server (particularly ours with limited resources) can handle. Testing with `PSIPRED` Vr. 3.5 and a recent version of the UniRef90 database [138], as well as, the new `Jpred4` REST API web service consistently took over 5 minutes to complete per query. We next tried using the `PSIPRED_SINGLE` stand-alone executable included with the `PSIPRED` Vr. 3.5 package. This version of the program does not compare against a reference database and, while less accurate compared to the main program, finishes a prediction in a few seconds [134]. An example discrepancy between both programs can be seen in Table 6.1 below for the AMP "Gomesin" (APD ID: AP00191) which is listed in the APD as a β-sheet.

Table 6.1: A comparison of outputs for `PSIPRED` Vr. 3.5 with UniRef90 vs. `PSIPRED_SINGLE` executables from [134] on the β-sheet AMP Gomesin (APD ID: AP00191; Sequence: QCRRLCYKQRCVTYCRGR). Program output shows the position confidence scores (Conf.) which range [0,9], the residue predicted structure (Pred.) which can be H (helical), C (coil) or E (sheet), and the original sequence (AA). The `PSIPRED_SINGLE` version of the program (right) incorrectly predicts a majority of residues as helical (H) while regular `PSIPRED` (left) predicts two patches as sheets (E) and is consistent with the known β-sheet structure of the AMP.

PSIPRED with UniRef90	**PSIPRED SINGLE**
`Conf: 941203203123113169`	`Conf: 906445553245552279`
`Pred: CCCCEEECCCEEEEECCC`	`Pred: CCHHHHHHHHHHHHHCCC`
`AA: QCRRLCYKQRCVTYCRGR`	`AA: QCRRLCYKQRCVTYCRGR`

As a basic assessment of PSIPRED_SINGLE's ability to discriminate between α-helical and β-sheet AMPs, all sequences from the APD with a 'helix' or 'beta' designation under secondary structure were downloaded. This resulted in a set of 353 α-helix and 102 β-sheet AMPs. We then submitted each sequence to PSIPRED_SINGLE mode and, based on the per-residue output predictions (like those seen in Table 6.1), classified a peptide as 'helix' or 'beta' if it contained more H or E characters respectively (residues predicted as C were ignored). This resulted in reasonably good accuracy, with 90.29% of sequences being correctly labeled and an MCC score of 0.722. However, one concern is that there are many AMPs with mixed helical and β-sheet structures [18] and the hydrophobic moment associated with the secondary structure is not considered in PSIPRED's approach. Recall our main interest is in the secondary structure for segments with a large $\widehat{\mu H}$ score.

To address address this concern, we design a simple decision tree based on two key observations for the α-helix and β-sheet peptides downloaded form the APD. The first is that, β-sheet AMPs, such as those in the defensin family, tend to have more C residues compared to α-helical AMPs. The other observation based on looking at some average $\widehat{\mu H}$ scores for α-helix and β-sheet AMPs (using their respective δ values) is that β-sheets tend to produce larger values for $\widehat{\mu H}$. This led to the following basic decision tree shown here in pseudo-code:

```
# Find window with the highest muH_hat score assuming alpha-helix then beta-sheet
muH_alpha = get_muH_hat(query_peptide, delta=100);
muH_beta = get_muH_hat(query_peptide, delta=160);

# Select beta-sheet if our query has >= theta cysteines
if (query_peptide).count('C') >= theta
        return muH_beta;

# Or select beta-sheet if it is gamma times larger than muH_alpha
```

```
elsif muH_beta > (gamma * muH_alpha)
        return muH_beta;

# Otherwise select alpha-helix
else
        return muH_alpha;
end
```

To determine values for the parameters "theta" and "gamma" above, we perform an exhaustive grid search in the context of 10-fold cross validation. We first separately split our helical and β-sheet AMPs into folds $1, \ldots, 10$ and then merge the helical and β-sheet observations with the same fold number together (ensuring all folds contain observations with both secondary structures types). Next, we combine 9/10 folds as training (reserving the last fold for testing) and exhaustively try all combinations of theta in [0,10] (stepping by 1) and gamma in [0,10] (stepping by 0.01) and recording the values that provide the highest ACC and MCC on the training fold. This is then repeated so that all folds are held out as a testing set and we average theta and gamma values from the 10 trials. These averaged values are then used on the testing folds separately so that we can observe a standard deviation (SD) when calculating ACC and MCC. The averaged value for theta (rounded to the nearest whole number of cysteines) was 5 (SD: 0.99), and mean gamma was determined as 4.62 (SD:0.90). This produces a mean ACC of 90.51% (SD:0.04) and mean MCC of 0.716 (SD:0.118) across the testing folds.

Using this simple decision tree approach results in a slighter higher ACC (0.22% higher) and similar MCC (0.006 lower) value compared to using `PSIPRED_SINGLE` for the AMPs tested. We note the decision tree approach is slightly faster and only considers secondary structure for portions of the AMP sequence with a high $\widehat{\mu H}$ value. Given a query peptide of mixed helical and β secondary structure, it has a better chance of selecting the portion of the sequence relevant for AMP-activity to designate as our AS region. Since the β-sheet

calculation of $\widehat{\mu H}$ considers every other residue, this method can still locate a high-scoring hydrophobic stretch of amino acids even in a peptide with a random coil structure as well.

Final Feature Vector Representation

After incorporating the changes discussed above to our feature set, our final vector to represent a peptide is comprised of 1276 numerical and 208 binary values. We group these 1484 features into sections as follows:

$$\left\langle FULL_1, \ldots, FULL_{319}, SASA_1, \ldots, SASA_{18}, NT_1, \ldots, NT_{312}, AS_f, AS_z, AS_{\widehat{\mu H}}, AS_1, \ldots, AS_{312}, \ldots, CT_1, \ldots, CT_{312}, EFC_1, \ldots, EFC_{208}, C \right\rangle$$

Here, $FULL$ stands for global features averaged over the full peptide and includes the 299 AAIndex features from Chapter 5, 7 features plus 2 interactions from Chapter 4, and 11 new `emboss` features. SASA features include the 6 new SASA categories mentioned above averaged for the full peptide, NT and CT segments. NT, CT and AS are features averaged over those specific functional regions and include 299 AAIndex features plus 11 new ones from `emboss`. Additionally, GRAVY and isoelectric point from Chapter 4 are also included since they do not require the full peptide sequence for calculation. AS also has the three additional features for: f, z and $\widehat{\mu H}$. Finally, C represents a class label using 1 for AMP and 0 for non-AMP. The full list of global and features, including descriptions and sources can be found in Appendix B. EFC features are defined in Appendix C.

6.2.2 Features Selection and Evaluation

In addition to using the complete feature set described above, we also apply feature reduction using three different selection methods implemented in Weka on the Xiao et al. training data set. The first is FCBF which selects a subset of 33 features and is detailed in the previous chapter. We also try a "BestFirst" search, which starts with the entire feature set and incrementally deletes a feature one at a time while assessing the impact the

removed feature has on predictive performance. The method does not rank features but retains those which hurt performance when removed and abandons those which improve performance when removed. Using this method, a subset of 80 features is selected. The final feature selection method we employ is the "GreedyStepwise" selection algorithm which operates in a similar manner to BestFirst. The algorithm starts with all features and removes each until it exhaustively encounters all subsets of the data. It first removes features which improve performance the most before removing features which hurt performance the least. As it processes each of these deletions, it generates a ranking based on the order in which features are removed and a merit score to describe the remaining subset of features. We select a subset of 75 features which provided the highest merit score. The list of features selected by each of these methods is provided in the next section.

The predictive performance of all features and feature subsets are evaluated in the context of supervised classification using Weka's implementation of the ZeroR, NaiveBayes, LR, Kernel LR (using the PolyKernel), J48 [139] (also known as C4.5) and RF algorithms. Additionally, the LibSVM package is used to implement SVM with the linear and RBF kernels (parameters selected using the included `grid.py` script according to the LibSVM documentation [140]). Lastly, the *Earth* package in R is used to implement MARS trying both 2- and 3-way interactions. As output from the *Earth* package only produces prediction values between $0 - 1$, we use the deviance reduction thresholding technique introduced in Chapter 4 Section 4.2.5 to select a cutoff value for assigning AMP and non-AMP labels.

As in the previous chapter, the Xiao training data set is used to train each model, and performance is evaluated using both the Fernandes and Xiao testing data sets for comparison. Due to the $\geq$ 10 AA length limit imposed by our new model, consideration has to be taken into how to treat sequences below this limit. While the number of observations $<$ 10 AA in the data sets used are relatively few in number, those that do fall into this category are almost all non-AMPs. Accordingly, rather than simply reject a sequence $<$ 10 AA as a non-AMP, we choose to exclude these sequences when reporting performance statistics as this would artificially inflate our numbers without truly

evaluating these sequences for antimicrobial activity. Where possible, other publicly available methods are also tested on the same data sets while also excluding < 10 AA sequences. Servers used include: `CAMP` [44](`http://www.camp.bicnirrh.re.in/predict`), `AntiBP2` [87] (`http://www.imtech.res.in/raghava/antibp2`) and `iAMP-2L` [41] (`http://www.jci-bioinfo.cn/bioinfo/iAMP-2L`). Attempts to use the server from [40] was unsuccessful as the services was no longer available at the time of this writing.

We conclude with an analysis on the features selected for two of the top performing methods.

6.2.3 Implementation Details

All classification experiments in this chapter were performed on an Intel i5 quad-core machine with 2.5Ghz CPU and 6GB of RAM. Weka is used to apply the three feature selection methods using FCBF (as in Chapter 5), the BestFirst approach (using the bi-direction option $D = 2$ and the `searchTermination` parameter set to 5) and a GreedyStepwise search (using backwards search with the `searchBackwards`:True and `generateRanking`:True options). Weka is also used to implement the following classifiers with settings given in parenthesis: ZeroR (`globalBlend`:20), NaiveBayes, LR (`regularization`:$1e-8$, Kernel LR (`kernel`:polyKernel [`exponent`:1], `lambda`:0.01), J48/C4.5 (`confidenceFactor`:0.25, `numFolds`:3), RF (`numTrees`:50, 100, 250, 500, 750, 775 and 800). The number of trees selected for RF started at 50 and were increased until performance no longer improved on the Xiao training data set. The LibSVM Vr. 3.18 package [141] is used with the linear kernel (Using `c`:2.0 for all feature subsets) and RBF kernel (Using all features: `c`:2.0, `g`:0.031; Using FCBF features: `c`:2.0, `g`:0.50; Using BestFirst Features: `c`:32.0, `g`:0.125; Using GreedyStepwise Features: `c`:8, `g`:0.5). Parameters were selected using the included `grid.py` script according to LibSVM documentation [141]. Finally, R and the `earth` package are used to implement MARS (`pmethod`:backward, `degree`:2 and 3).

Run times for training and testing take between 5 minutes to 12 hours depending on the classifier when all features are used. This is reduced to between $1 - 45$ minutes after any

of the feature selection methods are applied as fewer than 100 features are used to build models.

6.3 Results

This section describes evaluation results for the features described above when paired with a number of binary classifier algorithms. From this point on, we will refer to this complete feature set as "CFS" (where all features are considered). When the FCBF, BestFirst or GreedyStepwise methods are applied to CFS for feature selection, we refer to these reduced subsets of features as "CFS-FCBF," "CFS-BF," and "CFS-GS" respectively.

We begin using CFS features applied to the Xiao training data set and evaluate performance in the context of 10-fold CV as in Chapter 5. As all sequences in the training set are $\geq$ 10 AA, none are removed. We note, the EFC features generated in Chapter 5 are reused here so methods are not stochastic as in the previous chapter. Performance is then shown when our trained models are applied to the Xiao testing set, followed by the Fernandes data set. Both of these data sets are the same used in previous chapters and detailed in Chapter 3. However, we note the $\geq$ 10 AA length cutoff requirement removes 3 AMPs and 82 non-AMPs for consideration from the Xiao testing set. All sequences in the Fernandes data set are $\geq$ 10 AA, so none are removed.

Next, we list the features selected when the CFS-FCBF, CFS-BF, and CFS-GS feature reduction methods are applied and repeat the same training and testing procedures used for CFS. In order to compare our performance with other AMP recognition methods, we also submit these testing sets (using only the $\geq$ 10 AA Xiao testing sequences) to other AMP classifiers with publicly available servers and report results.

We conclude with an updated state summary of the field and present a detailed analysis of features and interactions (when present) used by two of our top models.

6.3.1 Performance Evaluation with the Complete Feature Set

Classifier Performance Comparison using CFS on the Xiao Training Set

Results using CFS features on the Xiao training data set are presented in Table 6.2 in the context of 10-fold CV below. Classifiers were implemented using the parameters outlined above. For the RF method we present results using various number of trees ranging between $50 - 800$ to observe where performance improvements level off. Adding more trees than necessary will slow performance without any gain in recognition performance. Across all values for the number of trees tried, the RF out of bag error ranges between $9 - 17\%$ and Weka randomly selected 11 features for each tree (based on program defaults). Two separate entries for MARS are given based on the inclusion of second or third-order interactions. Best overall predictive performance for the training set is found with MARS using third-order interactions according to all metrics except specificity. The baseline classifier ZeroR obtains the best specificity by simply labeling all observations as non-AMPs as the training set has more non-AMP than AMP observations. As all models are generated from this training set data, ZeroR obtains 100% specificity and 0% sensitivity for all experiments in this chapter. Excluding the ZeroR classifier, general performance appears good for all classifiers considered with MCC values ranging from $0.609 - 0.842$ and accuracies between $85 - 94\%$. This demonstrates that CFS contains features relevant for recognizing AMP-activity.

Classifier Performance Comparison using CFS on the Xiao Testing Set

Results using CFS features trained on the Xiao training data and applied to the Xiao testing data set can be seen below in Table 6.3. MCC and ACC values range between $0.809 - 0.950$ and $90 - 98\%$ respectively for all methods excluding ZeroR. RF appears to perform well regardless of the number of trees selected, with the best MCC seen using 250 trees. However, SVM using the RBF kernel achieves top performance according to sensitivity, ACC and MCC. Non-linear methods appear to outperform linear methods. For example, the RBF kernel performs better than linear kernel and other non-linear methods consistently outperform LR.

Table 6.2: 10-fold CV performance for CFS and various classification algorithms is shown using the Xiao training data set. The best performance for each metric is highlighted in bold. Descriptions of the parameters used for each method are described above in Methods Section 6.2.3. For RF, separate entries represent the number of trees selected (beginning with 50 trees and incremented until performance improvements leveled off). Trees were constructed using 11 random features based on Weka's default settings. The RF out of bag error ranged between $9 - 17\%$. Earth is an R implementation of MARS and separate entries with "Deg2" and "Deg3" refer to the inclusion of second and third-order interactions, respectively. Best performance according to sensitivity, ACC, MCC and auROC is seen with Earth-Deg3. The ZeroR classifier is provided as a baseline for comparison. It ignores features and simply predicts based on the mode of the class labels.

Method	**Sensitivity (%)**	**Specificity (%)**	**ACC (%)**	**MCC**	**auROC (%)**
ZeroR	0.00	**100**	73.7	NA	49.8
LR	72.5	89.1	84.7	0.609	88.5
Kernel-LR	79.4	93.3	89.3	0.731	93.7
NaiveBayes	81.2	85.8	84.6	0.634	86.8
J48	73.4	90.7	86.2	0.642	80.1
RF-50Trees	73.9	96.2	90.3	0.742	95.5
RF-100Trees	74.7	96.0	90.4	0.744	95.8
RF-200Trees	75.5	96.0	90.6	0.751	95.9
RF-250Trees	76.2	96.0	90.8	0.755	96.0
RF-500Trees	75.4	95.9	90.5	0.749	95.9
RF-750Trees	75.8	95.9	90.6	0.750	96.0
RF-775Trees	76.0	95.8	90.6	0.751	96.0
RF-800Trees	75.5	95.8	90.4	0.746	96.0
SVM-Linear	79.0	93.7	89.8	0.735	94.8
SVM-RBF	79.3	95.6	91.3	0.771	96.1
Earth-Deg2	86.9	94.3	92.4	0.805	96.9
Earth-Deg3	**87.8**	96.1	**93.9**	**0.842**	**98.1**

Since LR using CFS obtains a MCC of 0.81, which is slightly lower than the MCC of 0.86 obtained using LR with the EFC+307-FCBF model in Chapter 5, it seems that CFS contains a number of extraneous features that hurt recognition performance. We see in later sections of this chapter that feature reduction methods allow LR to perform even better than EFC+307-FCBF after these unhelpful features are removed.

Table 6.3: Performance for CFS and various classification algorithms is shown for the Xiao testing data after training on the Xiao training data set. All models match the ones used with the Xiao training data 10-fold CV experiment seen in Table 6.2. Best performance according to sensitivity, ACC and MCC can be seen with SVM using the RBF kernel. RF using any number of trees achieves higher performance according to specificity and auROC.

Method	**Sensitivity (%)**	**Specificity (%)**	**ACC (%)**	**MCC**	**auROC (%)**
ZeroR	0.00	**100**	47.1	NA	50.0
LR	85.6	95.3	90.1	0.809	97.2
Kernel-LR	91.7	99.4	95.3	0.910	98.7
NaiveBayes	88.5	97.7	92.8	0.861	97.5
J48	86.4	96.8	91.3	0.832	90.4
RF-50Trees	94.0	**100**	96.8	0.938	**100**
RF-100Trees	94.1	**100**	96.9	0.939	**100**
RF-200Trees	94.2	**100**	96.9	0.941	**100**
RF-250Trees	94.8	**100**	97.2	0.946	**100**
RF-500Trees	94.5	**100**	97.1	0.944	**100**
RF-750Trees	94.5	**100**	97.1	0.944	**100**
RF-775Trees	94.7	**100**	97.2	0.945	**100**
RF-800Trees	94.5	**100**	97.1	0.944	**100**
SVM-Linear	91.7	98.9	95.1	0.905	99.3
SVM-RBF	**95.3**	99.9	**97.5**	**0.950**	99.8
Earth-Deg2	94.7	99.3	96.8	0.938	99.4
Earth-Deg3	91.7	98.5	94.9	0.901	99.5

Classifier Performance Comparison using CFS on the Fernandes Testing Set

Results using CFS features trained on the Xiao training data and applied to the Fernandes testing data set can be seen below in Table 6.4. MCC and ACC values are generally lower compared to the Xiao testing set and range between $0.322 - 0.897$ and $62 - 95\%$ respectively for methods aside from ZeroR. As with the Xiao testing set, RF again performs well. It obtains the best performance using 200 trees in terms of sensitivity, ACC, MCC and auROC.

If one considers average performance between the Xiao and Fernandes test sets, best performance is also seen with RF using 200 trees according to MCC (0.919) and ACC (96%). This shows tha RF appears to have good general performance and does just as well, or better, using between $100 - 250$ trees compared to larger tree values. Kernel-LR

Table 6.4: Performance for CFS and various classification algorithms on the Fernandes data set after training on the Xiao training data set is shown. All models match the ones used with the Xiao training data 10-fold CV experiment seen in Table 6.2. Best performance according to sensitivity, ACC, MCC and auROC can be seen with RF using 200 trees. The ZeroR classifier is provided as a baseline for comparison, it ignores features and simply predicts the mode of class labels.

Method	**Sensitivity (%)**	**Specificity (%)**	**ACC (%)**	**MCC**	**auROC (%)**
ZeroR	0.00	**100**	50.2	NA	50.0
LR	88.7	82.8	85.7	0.716	90.6
Kernel-LR	83.5	50.0	62.1	0.332	91.1
NaiveBayes	81.7	80.2	81.0	0.619	86.1
J48	86.1	87.1	86.6	0.732	87.8
RF-50Trees	91.3	94.0	92.6	0.853	98.3
RF-100Trees	**93.0**	95.7	94.4	0.888	98.4
RF-200Trees	**93.0**	96.6	**94.8**	**0.897**	**98.5**
RF-250Trees	**93.0**	95.7	94.4	0.888	98.4
RF-500Trees	**93.0**	95.7	94.4	0.888	98.4
RF-750Trees	92.2	95.7	94.0	0.879	**98.5**
RF-775Trees	**93.0**	95.7	94.4	0.888	**98.5**
RF-800Trees	**93.0**	94.8	93.9	0.879	**98.5**
SVM-Linear	60.6	98.3	75.0	0.589	91.1
SVM-RBF	77.4	93.1	85.3	0.714	95.0
Earth-Deg2	80.9	91.4	86.1	0.727	92.7
Earth-Deg3	82.6	90.5	86.6	0.734	93.3

and SVM both see substantial drops in performance on the Fernandes testing set compared to the Xiao data. While MARS using third-order interactions also performs worse on the Fernandes data, it still demonstrates good overall performance and manages an average ACC of 91% between both testing sets.

6.3.2 Performance Evaluation with Reduced Feature Sets from CFS

We now perform a side-by-side comparison of the FCBF, BestFirst and GreedyStepwise feature selection methods from Weka. Results show that comparable recognition performance for AMPs can be maintained using smaller subsets of more relevant features compared to CFS. After detailing the features selected by each method, we show performance as we

did above for CFS. First, in the context of 10-fold CV on the Xiao training data before assessing performance on both the Xiao testing and Fernandes data sets. It is notable that the whole-peptide averaged feature for molecular weight was selected by all methods and was chosen as the top feature for methods with rankings. Additionally, the top 5 features for FCBF are also selected by the other two methods.

Features Selected by FCBF

When FCBF is applied to CFS using Weka as in Chapter 5, it selects a reduced set of 32 features. These are listed below in Table 6.5. Of these, 7 features are global (whole-peptide averaged) while 17 are binary EFC features retained from the previous chapter. The remaining 8 features comprise new regional features including: 4 for NT, 2 for CT and 2 for AS. None of the new SASA features were selected by the method.

Table 6.5: A list of 32 FCBF-selected features from CFS on the Xiao training data are shown. Columns from left-to-right show the rank (ordered from 1 as most important to 32 as least-important), feature name, source and a brief description of the feature. Names may start with the prefix FULL (whole-peptide averaged), EFC (evolutionary feature constructed from Chapter 5), NT, AS or CT to distinguish the functional region it represents. Additional information for AAIndex features is available from the AAIndex database [121]. Motifs for EFC features are given in the 4-letter alphabet detailed in Chapter 5 Table 5.1.

Rank	Name	Source	Description
1	FULL_MolWeight	emboss	Global averaged molecular weight
2	FULL_AURR980107	AAIndex	Global averaged normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
3	FULL_YUTK870103	AAIndex	Global averaged activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
4	FULL_OOBM850104	AAIndex	Global averaged optimized average non-bonded energy per atom (Oobatake et al 1985)
5	FULL_FUKS010112	AAIndex	Global averaged entire chain composition of amino acids in nuclear proteins (%) (Fukuchi-Nishikawa 2001)
6	FULL_FINA910104	AAIndex	Global averaged helix termination parameter at position j-plus1 (Finkelstein et al 1991)
7	NT_LEVM760103	AAIndex	N-termini averaged side chain angle theta(AAR) (Levitt 1976)
8	NT_DAYM780201	AAIndex	N-termini average relative mutability (Dayhoff et al 1978b)
9	FULL_BUNA790103	AAIndex	Global averaged spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich 1979)
10	CT_FINA910102	AAIndex	C-termini average helix initiation parameter at position ii-plus1i-plus2 (Finkelstein et al 1991)
11	AS_MAXF760103	AAIndex	Amphipathic segment average normalized frequency of zeta R (Maxfield-Scheraga 1976)
12	NT_WERD780102	AAIndex	N-termini average free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
13	NT_QIAN880137	AAIndex	N-termini average weights for coil at the window position of 4 (Qian-Sejnowski 1988)

14	EFC_107	EFC from Ch 5	Match CGA at position 11 ± 3
15	EFC_37	EFC from Ch 5	Match GGC at position 11 ± 3
16	CT_MEEJ810101	AAIndex	C-termini average retention coefficient in NaCL04 (Meek-Rossetti 1981)
17	EFC_113	EFC from Ch 5	Match AGCC at position 14 ± 3
18	EFC_4	EFC from Ch 5	Match GA and GTCGC at any position
19	AS_BEGF750102	AAIndex	Amphipathic segment average conformational parameter of beta-structure (Beghin-Dirkx 1975)
20	EFC_19	EFC from Ch 5	Match TATTT at any position
21	EFC_24	EFC from Ch 5	Match ACCCCTG at any position
22	EFC_205	EFC from Ch 5	Match ATAATC at position 11 ± 3
23	EFC_92	EFC from Ch 5	Match AGGAACG at any position
24	EFC_25	EFC from Ch 5	Match AGCCAC at any position
25	EFC_130	EFC from Ch 5	Match ACCTACG at position 4 ± 3
26	EFC_191	EFC from Ch 5	Match TGT at position 36 ± 3
27	EFC_50	EFC from Ch 5	Match TTG and GCGT at any position
28	EFC_73	EFC from Ch 5	Match TTAT at position 12
29	EFC_119	EFC from Ch 5	Match TATG at position 33 ± 3
30	EFC_104	EFC from Ch 5	Match TCACAGCT at any position
31	EFC_172	EFC from Ch 5	Match TCTCAT at any position
32	EFC_101	EFC from Ch 5	Match GACCATGA at any position

Features Selected by BestFirst Search

The BestFirst feature selection method was applied in Weka using the bi-directional search option to generate a reduced set of 80 features. These are listed below in Table 6.6. Of these, 30 features are global (whole-peptide averaged) features, 10 are SASA features (of these 5 are whole-peptide, 3 for NT and 2 for CT) and 15 are binary EFC features retained from the previous chapter. The remaining 25 features comprise other averaged regional features including: 16 for NT, 5 for CT and 4 for AS. BestFirst does not generate a set ranking of features so they are listed in Table 6.6 in the same order they were provided to the method as in Section 6.2.1 above.

Table 6.6: 80 BestFirst-selected features from CFS on the Xiao training data. The BestFirst method does not provide a specific feature ranking so columns from left-to-right show the feature name, source and a brief description of the feature. Names may start with the prefix FULL (whole-peptide averaged), EFC (evolutionary feature constructed from Chapter 5), NT, AS or CT to distinguish the functional region it represents. Additional information for AAIndex features is available from the AAIndex database [121]. Motifs for EFC features are given in the 4-letter alphabet detailed in Chapter 5 Table 5.1.

Name	**Source**	**Description**
FULL_MolWeight	`emboss`	Global averaged molecular weight
FULL_AcidicMolPerc	`emboss`	Global average mole% acidic residues
FULL_HelixPropensity	Ch 4 and [46]	Global averaged propensity to form helix secondary structure
FULL_TurnPropensity	Ch 4 and [46]	Global averaged propensity to form turn secondary structure
FULL_BetaPropensity	Ch 4 and [46]	Global averaged propensity to form beta secondary structure
FULL_Length_x_*in vitro*_Aggregation	Ch 4 and [46]	Global averaged feature interaction between length and *in vitro* aggregation
FULL_GRAVY_x_*in vitro*_Aggregation	Ch 4 and [46]	Global averaged feature interaction between hydrophobic moment and *in vitro* aggregation
FULL_AURR980105	AAIndex	Global averaged normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
FULL_AURR980106	AAIndex	Global averaged normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
FULL_AURR980107	AAIndex	Global averaged normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
FULL_BUNA790103	AAIndex	Global averaged spin-spin coupling constants 3JH alpha-NH (Bundi-Wuthrich 1979)
FULL_CHAM830102	AAIndex	Global averaged parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)
FULL_CHOP780204	AAIndex	Global averaged normalized frequency of N-terminal helix (Chou-Fasman 1978b)
FULL_CHOP780207	AAIndex	Global averaged normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
FULL_DAYM780201	AAIndex	Global averaged relative mutability (Dayhoff et al 1978b)
FULL_FINA910101	AAIndex	Global averaged helix initiation parameter at posision i-minus1 (Finkelstein et al 1991)
FULL_FINA910104	AAIndex	Global averaged helix termination parameter at posision j-plus1 (Finkelstein et al 1991)
FULL_FUKS010112	AAIndex	Global averaged entire chain composition of amino acids in nuclear proteins (%) (Fukuchi-Nishikawa 2001)
FULL_GEOR030101	AAIndex	Global averaged linker propensity from all dataset (George-Heringa 2003)
FULL_KARP850103	AAIndex	Global averaged flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
FULL_KLEP840101	AAIndex	Global averaged net charge (Klein et al 1984)
FULL_NAKH900102	AAIndex	Global averaged SD of AA composition of total proteins (Nakashima et al 1990)
FULL_OOBM850104	AAIndex	Global averaged optimized average non-bonded energy per atom (Oobatake et al 1985)
FULL_QIAN880136	AAIndex	Global averaged weights for coil at the window position of 3 (Qian-Sejnowski 1988)
FULL_RACS820101	AAIndex	Global averaged mean relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
FULL_RACS820102	AAIndex	Global averaged mean relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
FULL_RACS820104	AAIndex	Global averaged mean relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)
FULL_ROBB760107	AAIndex	Global averaged information measure for extended without H-bond (Robson-Suzuki 1976)

FULL_ROSM880103	AAIndex	Global averaged loss of Side chain hydropathy by helix formation (Roseman 1988)
FULL_YUTK870103	AAIndex	Global averaged activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
FULL_BBU_All	SASA Feature from [128]	Global average SASA for all backbone atoms (upper-bound)
FULL_BBU_A	SASA Feature from [128]	Global average SASA for apolar backbone atoms (upper-bound)
FULL_SCL_All	SASA Feature from [128]	Global average SASA for all side-chain atoms (lower-bound)
FULL_SCL_P	SASA Feature from [128]	Global average SASA for polar side-chain atoms (lower-bound)
FULL_SCL_A	SASA Feature from [128]	Global average SASA for apolar side-chain atoms (lower-bound)
NT_BBU_A	SASA Feature from [128]	N-termini average SASA for apolar backbone atoms (upper-bound)
NT_SCL_All	SASA Feature from [128]	N-termini average SASA for all side-chain atoms (lower-bound)
NT_SCL_A	SASA Feature from [128]	N-termini average SASA for apolar side-chain atoms (lower-bound)
CT_SCL_All	SASA Feature from [128]	C-termini average SASA for all side-chain atoms (lower-bound)
CT_SCL_A	SASA Feature from [128]	C-termini average SASA for apolar side-chain atoms (lower-bound)
NT_BUNA790103	AAIndex	N-termini average spin-spin coupling constants 3JH alpha-NH (Bundi-Wuthrich 1979)
NT_CHOP780207	AAIndex	N-termini average normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
NT_CHOP780215	AAIndex	N-termini average frequency of the 4th residue in turn (Chou-Fasman 1978b)
NT_DAYM780201	AAIndex	N-termini average relative mutability (Dayhoff et al 1978b)
NT_GEIM800104	AAIndex	N-termini average alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
NT_GEOR030101	AAIndex	N-termini average linker propensity from all dataset (George-Heringa 2003)
NT_KARP850103	AAIndex	N-termini average flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
NT_KHAG800101	AAIndex	N-termini average the Kerr-constant increments (Khanarian-Moore 1980)
NT_MAXF760105	AAIndex	N-termini average normalized frequency of zeta L (Maxfield-Scheraga 1976)
NT_NAKH900102	AAIndex	N-termini average SD of AA composition of total proteins (Nakashima et al 1990)
NT_QIAN880131	AAIndex	N-termini average weights for coil at the window position of -2 (Qian-Sejnowski 1988)
NT_QIAN880137	AAIndex	N-termini average weights for coil at the window position of 4 (Qian-Sejnowski 1988)
NT_RICJ880101	AAIndex	N-termini average relative preference value at Ntt (Richardson-Richardson 1988)
NT_WERD780102	AAIndex	N-termini average free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
NT_WILM950103	AAIndex	N-termini average hydrophobicity coefficient in RP-HPLC C4 with 01% TFA-MeCN-H2O (Wilce et al 1995)
NT_WOLS870102	AAIndex	N-termini average principal property value z2 (Wold et al 1987)
AS_DAYM780201	AAIndex	Amphipathic segment average relative mutability (Dayhoff et al 1978b)
AS_FAUJ880110	AAIndex	Amphipathic segment average number of full non-bonding orbitals (Fauchere et al 1988)
AS_FINA910101	AAIndex	Amphipathic segment average helix initiation parameter at position i-minus1 (Finkelstein et al 1991)
AS_ROSM880103	AAIndex	Amphipathic segment average loss of Side chain hydropathy by helix formation (Roseman 1988)
CT_Charge	`emboss`	C-termini average mean charge at C-termini
CT_AURR980102	AAIndex	C-termini average normalized positional residue frequency at helix termini Nttt (Aurora-Rose 1998)

CT_FINA910102	AAIndex	C-termini average helix initiation parameter at position ii-plus1 i-plus2 (Finkelstein et al 1991)
CT_WILM950104	AAIndex	C-termini average hydrophobicity coefficient in RP-HPLC C18 with 01% TFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
CT_WIMW960101	AAIndex	C-termini average free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White 1996)
EFC_4	EFC from Ch 5	Match GA and GTCGC at any position
EFC_33	EFC from Ch 5	Match GC at position 1 ± 3
EFC_59	EFC from Ch 5	Match CA at position 59
EFC_84	EFC from Ch 5	Match CTATT at any position
EFC_87	EFC from Ch 5	Match GGGGGG at any position
EFC_92	EFC from Ch 5	Match AGGAACG at any position
EFC_121	EFC from Ch 5	Match ACCAC at position 43 ± 3
EFC_130	EFC from Ch 5	Match ACCTACG at position 4 ± 3
EFC_134	EFC from Ch 5	Match CACC and TA at any position
EFC_175	EFC from Ch 5	Match GTACACA at any position
EFC_185	EFC from Ch 5	Match CGA at position 9 ± 3
EFC_188	EFC from Ch 5	Match GCC at position 10 ± 3
EFC_192	EFC from Ch 5	Match AAAA at position 34 ± 3
EFC_195	EFC from Ch 5	Match GCCA at position 3 ± 3
EFC_205	EFC from Ch 5	Match ATAATC at position 11 ± 3

Features Selected by GreedyStepwise Search

The GreedyStepwise feature selection method was applied in Weka using the backwards search option to generate a reduced set of 74 features. These are listed below in Table 6.7. Of these, 30 features are global (whole-peptide averaged) features, 10 are SASA features (of these 5 are for the whole-peptide, 3 for NT and 2 for CT) and 10 are binary EFC features retained from the previous chapter. The remaining 24 features comprise other averaged regional features including: 15 for NT, 5 for CT and 4 for AS.

Table 6.7: 74 GreedyStepwise-selected features from CFS on the Xiao training data are shown. Columns from left-to-right list the feature rank (ordered from 1 as most important to 74 as least-important), score, name, source, and a brief description of the feature. The score is provided by the method and represents the merit of the remaining subset after that feature is removed. Accordingly, a lower score means the feature was removed later from the subset because it has higher importance. For example, the global feature for molecular weight at Rank 1 was the last feature removed which leaves an empty feature set with a 0.000 merit score. Conversely, the EFC feature at rank 74 is the least important feature listed and once removed leaves a subset with a merit score of 0.386. Feature names may start with the prefix FULL (whole-peptide averaged), EFC (evolutionary feature constructed from Chapter 5), NT, AS or CT to distinguish the functional region it represents. Additional information for AAIndex features is available from the AAIndex database [121]. Motifs for EFC features are given in the 4-letter alphabet detailed in Chapter 5 Table 5.1.

Rank	Score	Name	Source	Description
1	0.000	FULL_MolWeight	`emboss`	Global averaged molecular weight
2	0.246	CT_SCL_A	SASA Feature from [128]	C-termini average SASA for apolar side-chain atoms (lower-bound)
3	0.278	FULL_AURR980107	AAIndex	Normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
4	0.292	FULL_YUTK870103	AAIndex	Activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
5	0.306	CT_SCL_All	SASA Feature from [128]	C-termini average SASA for all side-chain atoms (lower-bound)
6	0.317	FULL_OOBM850104	AAIndex	Optimized average non-bonded energy per atom (Oobatake et al 1985)
7	0.326	FULL_Length_x_*in vitro*_Aggregation	Ch 4 and [46]	Global averaged feature interaction between length and *in vitro* aggregation
8	0.334	FULL_FINA910104	AAIndex	Helix termination parameter at position j-plus1 (Finkelstein et al 1991)
9	0.340	FULL_DAYM780201	AAIndex	Relative mutability (Dayhoff et al 1978b)
10	0.345	FULL_AcidicMolPerc	`emboss`	Global average mole% acidic residues
11	0.348	NT_MAXF760105	AAIndex	Normalized frequency of zeta L (Maxfield-Scheraga 1976)
12	0.352	NT_BBU_A	SASA Feature from [128]	N-termini average SASA for apolar backbone atoms (upper-bound)
13	0.355	FULL_ROSM880103	AAIndex	Loss of Side chain hydropathy by helix formation (Roseman 1988)
14	0.358	FULL_BUNA790103	AAIndex	Spin-spin coupling constants 3JH alpha-NH (Bundi-Wuthrich 1979)
15	0.360	FULL_FUKS010112	AAIndex	Entire chain composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa 2001)
16	0.362	FULL_SCL_P	SASA Feature from [128]	Global average SASA for polar side-chain atoms (lower-bound)
17	0.364	NT_DAYM780201	AAIndex	Relative mutability (Dayhoff et al 1978b)
18	0.366	FULL_HelixPropensity	Ch 4 and [46]	Global averaged propensity to form helix secondary structure
19	0.368	EFC_33	EFC from Ch 5	Match AGC at position 9 ± 3
20	0.369	FULL_KLEP840101	AAIndex	Net charge (Klein et al 1984)
21	0.370	FULL_GEOR030101	AAIndex	Linker propensity from all dataset (George-Heringa 2003)
22	0.372	NT_BUNA790103	AAIndex	Spin-spin coupling constants 3JH alpha-NH (Bundi-Wuthrich 1979)
23	0.373	NT_SCL_All	SASA Feature from [128]	N-termini average SASA for all side-chain atoms (lower-bound)
24	0.373	FULL_CHOP780204	AAIndex	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
25	0.374	FULL_RACS820104	AAIndex	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)

26	0.374	NT_SCL_A	SASA Feature from [128]	N-termini average SASA for apolar side-chain atoms (lower-bound)
27	0.375	AS_ROSM880103	AAIndex	Loss of Side chain hydropathy by helix formation (Roseman 1988)
28	0.376	EFC_187	EFC from Ch 5	Match GCC at position 10 ± 3
29	0.376	EFC_185	EFC from Ch 5	Match CGT at position 10 ± 3
30	0.377	FULL_BetaPropensity	Ch 4 and [46]	Global averaged propensity to form beta secondary structure
31	0.377	NT_NAKH900102	AAIndex	SD of AA composition of total proteins (Nakashima et al 1990)
32	0.378	CT_WILM950104	AAIndex	Hydrophobicity coefficient in RP-HPLC C18 with 01% TFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
33	0.378	NT_CHOP780207	AAIndex	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
34	0.379	FULL_AURR980106	AAIndex	Normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
35	0.379	FULL_FINA910101	AAIndex	Helix initiation parameter at position i-minus1 (Finkelstein et al 1991)
36	0.379	NT_GEIM800104	AAIndex	Alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
37	0.380	EFC_195	EFC from Ch 5	Match GGAT at position 62 ± 3
38	0.380	CT_FINA910102	AAIndex	Helix initiation parameter at position ii-plus1 i-plus2 (Finkelstein et al 1991)
39	0.380	EFC_134	EFC from Ch 5	Match AG at any position any ATG at position 71 ± 3
40	0.381	FULL_KARP850103	AAIndex	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
41	0.381	FULL_RACS820101	AAIndex	Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
42	0.381	EFC_59	EFC from Ch 5	Match GC at position 2
43	0.382	NT_KHAG800101	AAIndex	The Kerr-constant increments (Khanarian-Moore 1980)
44	0.382	FULL_RACS820102	AAIndex	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
45	0.382	FULL_SCL_All	SASA Feature from [128]	Global average SASA for all side-chain atoms (lower-bound)
46	0.383	AS_FAUJ880110	AAIndex	Number of full non-bonding orbitals (Fauchere et al 1988)
47	0.383	NT_GEOR030101	AAIndex	Linker propensity from all dataset (George-Heringa 2003)
48	0.383	FULL_SCL_A	SASA Feature from [128]	Global average SASA for apolar side-chain atoms (lower-bound)
49	0.383	CT_AURR980102	AAIndex	Normalized positional residue frequency at helix termini Nttt (Aurora-Rose 1998)
50	0.384	FULL_NAKH900102	AAIndex	SD of AA composition of total proteins (Nakashima et al 1990)
51	0.384	NT_QIAN880137	AAIndex	Weights for coil at the window position of 4 (Qian-Sejnowski 1988)
52	0.384	EFC_192	EFC from Ch 5	Match AGCC at position 9 ± 3
53	0.384	NT_WERD780102	AAIndex	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
54	0.384	FULL_CHOP780207	AAIndex	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
55	0.384	AS_FINA910101	AAIndex	Helix initiation parameter at position i-minus1 (Finkelstein et al 1991)
56	0.385	NT_CHOP780215	AAIndex	Frequency of the 4th residue in turn (Chou-Fasman 1978b)
57	0.385	FULL_BBU_All	SASA Feature from [128]	Global average SASA for all backbone atoms (upper-bound)
58	0.385	CT_CHOP780204	AAIndex	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
59	0.385	FULL_ROBB760107	AAIndex	Information measure for extended without H-bond (Robson-Suzuki 1976)
60	0.385	EFC_121	EFC from Ch 5	Match ACGTA at position 1 ± 3

61	0.385	FULL_AURR980105	AAIndex	Normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
62	0.385	NT_KARP850103	AAIndex	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
63	0.385	NT_NAKH900109	AAIndex	AA composition of membrane proteins (Nakashima et al 1990)
64	0.385	FULL_*in vitro*_Aggregation	Ch 4 and [46]	Global averaged propensity for *in vitro* aggregation
65	0.385	AS_DAYM780201	AAIndex	Relative mutability (Dayhoff et al 1978b)
66	0.385	FULL_TurnPropensity	Ch 4 and [46]	Global averaged propensity to form turn secondary structure
67	0.385	EFC_31	EFC from Ch 5	Match GA at position 12 ± 3
68	0.385	FULL_BBU_A	SASA Feature from [128]	Global average SASA for apolar backbone atoms (upper-bound)
69	0.385	CT_HOPA770101	AAIndex	Hydration number (Hopfinger 1971) Cited by Charton-Charton (1982)
70	0.386	NT_HUTJ700101	AAIndex	Heat capacity (Hutchens 1970)
71	0.386	NT_RICJ880101	AAIndex	Relative preference value at Ntt (Richardson-Richardson 1988)
72	0.386	FULL_CHAM830102	AAIndex	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)
73	0.386	FULL_QIAN880136	AAIndex	Weights for coil at the window position of 3 (Qian-Sejnowski 1988)
74	0.386	EFC_37	EFC from Ch 5	Match GGC at position 12 ± 3

Performance Comparison Using Reduced Feature Sets on the Xiao Training Set

Table 6.8 below shows classification performance using all three feature reduction methods on the Xiao training set in the context of 10-fold CV. Under each metric, CFS-FCBF values are denoted in parenthesis "( )", CFS-BF in square brackets "[ ]", and CFS-GS in curly brackets "{}." Comparable performance is seen across all metrics between the reduced feature sets and CFS (listed in Table 6.2). This is notable as less than 100 of the 1484 features in CFS appear necessary for good AMP recognition. CFS-FCBF, with only 32 features, performs about on par with CFS-GS which uses 74 features. However, CFS-BF with 80 features consistently shows the best performance across all metrics other than sensitivity on the training data. The difference amongst classifiers also appears similar to that seen for CFS in Table 6.2. The Earth-Deg3 method using CFS-BF obtains the best performance with a MCC of 0.868 and ACC of 94.9%. This is a slight improvement over the MCC of 0.842 and ACC of 93.9% obtained using CFS.

Table 6.8: 10-fold CV performance using the CFS-FCBF, CFS-BF and CFS-GS feature subsets paired with various classification algorithms (shown in rows) is shown for the Xiao training data. The best performance for each metric (shown in columns) is highlighted in bold and numbers separated for CFS-FCBF in parenthesis "( )", CFS-BF in square brackets "[ ]", and CFS-GS in curly brackets "{}" under each metric. The number of features considered to construct trees by the RF classifiers based on Weka defaults was 6, 7 and 7 for CFS-FCBF, CFS-BF, and CFS-GS respectively.

Method	Sensitivity (%)	Specificity (%)	ACC (%)	MCC	auROC (%)
ZeroR	(0.00) [0.00] {0.00}	(**100**) [**100**] {**100**}	(73.7) [73.7] {73.7}	(NA) [NA] {NA}	(49.8) [49.8] {49.8}
LR	(73.4) [78.4] {78.5}	(92.9) [93.3] {92.6}	(87.8) [89.4] {88.9}	(0.679) [0.725] {0.712}	(92.9) [94.8] {94.4}
Kernel-LR	(73.4) [79.1] {79.1}	(92.9) [93.2] {92.8}	(87.8) [89.5] {89.2}	(0.679) [0.727] {0.721}	(92.9) [94.9] {94.6}
NaiveBayes	(79.4) [89.5] {90.1}	(88.6) [84.1] {84.1}	(86.2) [85.5] {85.7}	(0.658) [0.678] {0.683}	(91.2) [91.9] {92.1}
J48	(69.7) [74.3] {74.5}	(90.7) [92.2] {92.3}	(85.2) [87.5] {87.6}	(0.612) [0.674] {0.676}	(83.8) [81.8] {82.3}
RF-50Trees	(75.5) [81.0] {80.2}	(94.8) [95.2] {94.6}	(89.8) [91.5] {90.9}	(0.729) [0.810] {0.762}	(94.9) [96.0] {96.0}
RF-100Trees	(76.7) [81.9] {80.7}	(94.8) [95.1] {94.9}	(90.0) [91.6] {91.2}	(0.736) [0.781] {0.770}	(95.1) [96.2] {96.2}
RF-200Trees	(77.1) [81.3] {80.1}	(94.9) [95.1] {94.8}	(90.2) [91.5] {91.2}	(0.741) [0.777] {0.769}	(95.2) [96.2] {96.2}
RF-250Trees	(77.2) [81.1] {81.3}	(94.9) [95.0] {94.9}	(90.2) [91.3] {91.4}	(0.742) [0.773] {0.774}	(95.2) [96.2] {96.3}
RF-500Trees	(77.7) [81.1] {81.4}	(95.1) [95.1] {95.1}	(90.5) [91.4] {91.5}	(0.750) [0.776] {0.777}	(95.3) [96.2] {96.3}
RF-750Trees	(77.3) [81.3] {81.6}	(95.1) [95.2] {95.0}	(90.4) [91.5] {91.5}	(0.747) [0.779] {0.777}	(95.3) [96.3] {96.3}
RF-775Trees	(77.2) [81.3] {81.7}	(95.1) [95.2] {95.0}	(90.4) [91.5] {91.5}	(0.746) [0.779] {0.778}	(95.3) [96.3] {96.3}
RF-800Trees	(77.3) [81.3] {81.3}	(95.0) [95.2] {95.1}	(90.4) [91.5] {91.4}	(0.746) [0.779] {0.776}	(95.3) [96.3] {96.3}
SVM-Linear	(75.1) [78.4] {78.4}	(92.1) [92.9] {92.7}	(87.6) [89.1] {88.9}	(0.678) [0.717] {0.713}	(92.5) [94.4] {94.2}
SVM-RBF	(79.3) [81.3] {81.2}	(94.5) [94.0] {94.7}	(90.5) [90.7] {91.2}	(0.751) [0.758] {0.769}	(95.1) [95.0] {95.4}
Earth-Deg2	(83.0) [86.3] {86.3}	(93.3) [94.0] {93.8}	(90.6) [92.0] {91.8}	(0.758) [0.796] {0.792}	(95.8) [96.6] {93.3}
Earth-Deg3	(84.6) [**92.7**] {86.3}	(94.5) [95.7] {93.8}	(92.0) [**94.9**] {91.8}	(0.793) [**0.868**] {0.792}	(96.7) [**97.7**] {93.3}

Performance Comparison Using Reduced Feature Sets on the Xiao Testing Set

Table 6.9 below shows classification performance using all three feature reduction methods on the Xiao testing set after training with the Xiao training data. Similar performance with CFS features is again found to that seen in Table 6.3 using all three of the reduced feature sets. However, unlike for the training set above, the best performance for each metric is seen using CFS-GS features. Excluding the ZeroR classifier, this can be observed for sensitivity and specificity using RF (96.5% and 100% respectively), ACC and MCC using RF or Naive Bayes (97.5% and 0.950 respectively), and auROC using RF (100%). RF using CFS-GS appears to have the best general performance as it matches or beats the other methods for all metrics other than sensitivity. Naive Bayes using CFS-GS obtains the best performance for sensitivity and ties for ACC and MCC.

Table 6.9: Performance on the Xiao testing data is shown for the CFS-FCBF, CFS-BF and CFS-GS feature subsets trained on the Xiao training data using various classification algorithms (shown in rows). The best performance for each metric (shown in columns) is highlighted in bold and numbers separated for CFS-FCBF in parenthesis "( )", CFS-BF in square brackets "[ ]", and CFS-GS in curly brackets "{}" under each metric. The number of features considered to construct trees by the RF classifiers (based on Weka defaults) was 6 for CFS-FCBF and 7 for both CFS-BF and CFS-GS.

Method	Sensitivity (%)	Specificity (%)	ACC (%)	MCC	auROC (%)
ZeroR	(0.00) [0.00] {0.00}	(**100**) [**100**] {**100**}	(47.1) [47.1] {47.1}	(*NA*) [*NA*] {*NA*}	(50.0) [50.0] {50.0}
LR	(83.1) [90.1] {90.0}	(99.5) [98.9] {98.8}	(90.8) [94.2] {94.1}	(0.830) [0.889] {0.886}	(98.8) [99.4] {99.4}
Kernel-LR	(83.1) [90.5] {90.1}	(99.5) [98.9] {99.0}	(90.8) [94.5] {94.3}	(0.830) [0.893] {0.890}	(98.8) [99.4] {99.4}
NaiveBayes	(89.3) [96.9] {**97.2**}	(98.7) [97.9] {97.9}	(93.7) [97.4] {**97.5**}	(0.879) [0.948] {**0.950**}	(99.2) [99.3] {99.3}
J48	(86.3) [88.1] {87.4}	(97.7) [99.3] {98.2}	(91.6) [93.4] {92.4}	(0.840) [0.874] {0.855}	(93.2) [94.2] {90.5}
RF-50Trees	(92.4) [94.2] {94.0}	(99.9) [**100**] {**100**}	(95.9) [96.9] {96.8}	(0.921) [0.941] {0.938}	(99.7) [99.8] {**100**}
RF-100Trees	(93.2) [94.8] {94.9}	(**100**) [**100**] {**100**}	(96.4) [97.2] {97.3}	(0.931) [0.946] {0.947}	(99.8) [**100**] {**100**}
RF-200Trees	(92.9) [95.0] {95.1}	(**100**) [**100**] {**100**}	(96.2) [97.3] {97.4}	(0.928) [0.948] {0.949}	(99.8) [**100**] {**100**}
RF-250Trees	(92.7) [95.0] {94.8}	(**100**) [**100**] {**100**}	(96.1) [97.3] {97.2}	(0.924) [0.948] {0.946}	(99.8) [**100**] {**100**}
RF-500Trees	(93.3) [94.5] {95.2}	(99.9) [**100**] {**100**}	(96.4) [97.1] {**97.5**}	(0.931) [0.944] {**0.950**}	(99.8) [**100**] {**100**}
RF-750Trees	(93.3) [94.8] {95.2}	(99.9) [**100**] {**100**}	(96.4) [97.2] {**97.5**}	(0.931) [0.946] {**0.950**}	(99.8) [**100**] {**100**}
RF-775Trees	(93.6) [94.9] {95.1}	(99.9) [**100**] {**100**}	(96.5) [97.3] {97.4}	(0.933) [0.947] {0.949}	(99.8) [**100**] {**100**}
RF-800Trees	(93.5) [94.9] {95.1}	(99.9) [**100**] {**100**}	(96.5) [97.3] {97.4}	(0.932) [0.947] {0.949}	(99.8) [**100**] {**100**}
SVM-Linear	(85.2) [91.6] {91.6}	(99.4) [98.5] {98.5}	(91.9) [94.9] {94.9}	(0.849) [0.916] {0.900}	(98.8) [99.2] {98.9}
SVM-RBF	(90.5) [93.9] {91.4}	(99.3) [99.9] {99.8}	(94.6) [96.7] {95.3}	(0.897) [0.936] {0.910}	(99.4) [99.6] {99.3}
Earth-Deg2	(92.8) [93.9] {93.9}	(99.4) [99.6] {99.0}	(95.9) [96.6] {96.3}	(0.920) [0.934] {0.928}	(99.4) [99.8] {99.7}
Earth-Deg3	(91.3) [93.6] {94.1}	(99.9) [98.9] {98.4}	(95.3) [96.1] {96.1}	(0.910) [0.923] {0.924}	(99.4) [99.4] {99.3}

6.3.3 Classifier Performance Comparison on the Fernandes Testing Set

Table 6.10 below shows classification performance using all three feature reduction methods on the Fernandes data set after training with the Xiao training data. Similar performance with CFS features is again found to that seen in Table 6.4 using all three of the reduced feature sets. As with the Xiao testing set, best performance under each metric is seen using the CFS-GS features. As in Table 6.4, RF using any number of trees between $50 - 800$ generally outperforms the other classifiers for all metrics other than specificity. However, RF using 100 trees and CFS-GS features tends achieves the best general performance whereas RF using 200 trees did the best with CFS. The difference on this data set between RF using 100 and 200 trees using CFS-GS features is one additional FP observation which happens once the number of trees is increased over 116. Table 6.10 shows RF using 100 trees and CFS-GS obtains the best ACC (94.4%) and MCC (0.888) while matching RF using more trees for best sensitivity (96.5%). RF using between 200 and 800 trees and CFS-GS obtains the best auROC (98.8%).

Table 6.10: Performance on the Fernandes data set is shown for the CFS-FCBF, CFS-BF and CFS-GS feature subsets trained on the Xiao training data using various classification algorithms (shown in rows). The best performance for each metric (shown in columns) is highlighted in bold and numbers separated for CFS-FCBF in parenthesis "( )", CFS-BF in square brackets "[ ]", and CFS-GS in curly brackets "{}" under each metric.

Method	Sensitivity (%)	Specificity (%)	ACC (%)	MCC	auROC (%)
ZeroR	(0.00) [0.00] {0.00}	(**100**) [**100**] {**100**}	(50.2) [50.2] {50.2}	(*NA*) [*NA*] {*NA*}	(50.0) [50.0] {50.0}
LR	(73.9) [82.6] {83.5}	(87.1) [92.2] {93.1}	(80.5) [87.4] {88.3}	(0.615) [0.752] {0.770}	(87.2) [92.6] {93.7}
Kernel-LR	(73.9) [82.6] {83.5}	(87.1) [92.2] {91.4}	(80.5) [87.4] {87.4}	(0.615) [0.752] {0.751}	(87.2) [92.6] {93.6}
NaiveBayes	(80.9) [87.0] {87.8}	(81.0) [86.2] {87.1}	(81.0) [86.6] {87.4}	(0.619) [0.732] {0.749}	(87.7) [94.6] {94.3}
J48	(87.8) [87.8] {88.7}	(82.8) [85.3] {88.8}	(85.3) [86.6] {88.7}	(0.707) [0.732] {0.775}	(87.1) [86.3] {88.5}
RF-50Trees	(93.0) [93.9] {95.7}	(85.3) [89.7] {91.4}	(89.2) [91.8] {93.5}	(0.786) [0.836] {0.871}	(96.9) [98.4] {98.5}
RF-100Trees	(93.9) [94.8] {**96.5**}	(87.6) [92.2] {92.2}	(90.8) [93.5] {**94.4**}	(0.817) [0.870] {**0.888**}	(96.5) [98.5] {98.6}
RF-200Trees	(94.8) [95.7] {**96.5**}	(86.2) [91.4] {91.4}	(90.5) [93.5] {93.9}	(0.813) [0.871] {0.880}	(96.7) [98.6] {**98.8**}
RF-250Trees	(95.7) [95.7] {**96.5**}	(87.1) [92.2] {91.4}	(91.3) [93.9] {93.9}	(0.830) [0.879] {0.880}	(96.9) [98.6] {**98.8**}
RF-500Trees	(94.8) [95.7] {**96.5**}	(85.3) [91.4] {91.4}	(90.0) [93.5] {93.9}	(0.805) [0.871] {0.880}	(97.0) [98.6] {**98.8**}
RF-750Trees	(94.8) [95.7] {**96.5**}	(86.2) [91.4] {91.4}	(90.5) [93.5] {93.9}	(0.813) [0.871] {0.880}	(97.0) [98.6] {**98.8**}
RF-775Trees	(94.8) [95.7] {**96.5**}	(85.3) [91.4] {91.4}	(90.0) [93.5] {93.9}	(0.805) [0.871] {0.880}	(96.9) [98.6] {**98.8**}
RF-800Trees	(94.8) [95.7] {**96.5**}	(85.3) [91.4] {91.4}	(90.0) [93.5] {93.9}	(0.805) [0.871] {0.880}	(96.9) [98.6] {**98.8**}
SVM-Linear	(63.5) [65.2] {72.2}	(94.8) [97.4] {97.4}	(79.2) [81.4] {84.8}	(0.615) [0.662] {0.720}	(90.1) [94.3] {94.7}
SVM-RBF	(49.6) [80.9] {86.1}	(95.7) [81.9] {87.9}	(72.7) [81.4] {87.0}	(0.510) [0.628] {0.740}	(87.9) [91.3] {92.9}
Earth-Deg2	(87.0) [86.1] {87.0}	(85.3) [90.5] {90.5}	(86.1) [88.3] {88.7}	(0.723) [0.767] {0.775}	(91.6) [94.1] {93.8}
Earth-Deg3	(85.2) [87.0] {88.7}	(85.3) [91.4] {91.4}	(85.3) [89.2] {90.0}	(0.706) [0.784] {0.801}	(92.6) [92.0] {95.7}

Averaging the results between the Fernandes and Xiao testing sets shows RF with 100 trees and CFS-GS features to be the overall best performer. It obtains an average ACC of 95.9% and MCC of 0.918 which is 0.2% and 0.004 higher than with CFS features respectively. MARS using third-order interactions and CFS-GS features is the second best general performing method with an average ACC of 93.1% and MCC of 0.863. This is respectively 2.4% and 0.046 higher than the method obtained with CFS features. This demonstrates that a reduced set of only 74 CFS-GS features is needed to achieve good general AMP recognition performance. In the next section we show how both of these methods compare to other currently available AMP prediction methods in the field.

6.3.4 Performance Comparison with Other Prediction Servers

Table 6.11 lists a performance comparison on the Xiao testing and Fernandes data sets using our top methods and the `AntiBP2`, `CAMP`, and `iAMP-2L` prediction servers. Results are shown for the Xiao testing set in parenthesis "( )", for the Fernandes data set in square brackets "[ ]", and average performance between both sets is given in curly brackets "{}." Our methods comprise of RF using 100 trees, as well as, multivariate adaptive regression

splines with third-order interactions. Both are paired with the CFS-GS subset of features so are respectively listed as RF-CFS-GS and Earth-CFS-GS.

Since our features require sequences with $\geq$ 10 AA and the `AntiBP2` server requires sequences $\leq$ 100 AA, we submit only sequences that fall into this range for all tests. For the Xiao testing set, this leaves 916 AMPs and 815 non-AMPs. The Fernandes set remains unchanged at 115 AMPs and 116 non-AMPs since all sequences are between $10 - 100$ AA in length.

Best average performance between both data sets is obtained with RF-CFS-GS according to specificity (96.0%), ACC (95.5%), MCC (0.911%) and auROC (97.6%). The CAMP-RF method obtains the best average sensitivity at 97.8% but is one of the weaker performers according to other metrics. Earth-CFS-GS obtains second-best average performance with comparable but slightly lower values compared to RF-CFS-GS.

As for specific data sets, the iAMP-2L method does best on the Xiao testing set with a MCC of 0.965. This testing set was taken from the iAMP-2L paper in [41] and it is not clear from the web site if the online implementation uses just the training set or if it has been re-trained using both the training and testing data sets provided with the paper. Table 4 in [41] lists a MCC on the testing set of 0.845 but this is using all 920 non-AMPs. If the server model is trained only using the training set, the large discrepancy could come from removing sequences due to the length limits mentioned above, as well as, additional sequences labeled as non-AMP but with potential antimicrobial activity detailed in Chapter 3. For the Fernandes data set, RF-CFS-GS obtains the best performance according to all metrics with a MCC of 0.888.

We next highlight how the RF-CFS-GS and Earth-CFS-GS methods ranked features in terms of importance.

6.3.5 Random Forest Feature Rankings

Table 6.12 below show feature rankings generated by RF using CFS-GS features on the Xiao training data set and 100 trees (each generated by 7 randomly selected features resulting

Table 6.11: Shown below is a performance comparison between our RF and MARS models using the CFS-GS features and other publicly-available methods are shown below. RF is listed as "RF-CFS-GS" and is implemented with 100 trees while MARS is listed as "Earth-CFS-GS" and is implemented allowing third-order interactions. Analysis was performed on both the Xiao and Fernandes testing data sets using only peptides with lengths between 10 and 100 AA. Performance under each metric is shown for the Xiao testing set in parenthesis "( )", for the Fernandes data set in square brackets "[ ]", and average performance across both sets in curly brackets "{}." The best average performance is highlighted in bold under each metric. The CAMP server RF implementation obtains best average sensitivity (97.8). Our RF-CFS-GS model is shown to have the best performance for specificity (96.9), ACC (95.5), MCC (0.911) and auROC (97.6). Earth-CFS-GS obtains (or ties for) second-best average performance for all metrics other than sensitivity.

Method	Sensitivity (%)	Specificity (%)	ACC (%)	MCC	auROC (%)
AntiBP2	(89.6) [87.0] {88.3}	(88.5) [83.6] {86.1}	(89.1) [85.3] {87.2}	(0.781) [0.706] {0.744}	(89.0) [85.3] {87.2}
CAMP-SVM	(96.1) [95.7] {95.9}	(35.7) [58.6] {47.2}	(67.7) [77.1] {72.4}	(0.405) [0.584] {0.495}	(75.8) [81.4] {78.6}
CAMP-RF	(97.4) [98.3] {**97.8**}	(28.7) [61.2] {45.0}	(65.1) [79.7] {72.4}	(0.366) [0.640] {0.503}	(75.6) [84.4] {80.0}
CAMP-ANN	(86.2) [87.0] {86.6}	(74.5) [67.2] {70.9}	(82.3) [77.1] {79.7}	(0.647) [0.553] {0.600}	(82.8) [78.2] {80.5}
CAMP-DA	(94.3) [93.0] {93.7}	(45.8) [61.2] {53.5}	(71.5) [77.1] {74.3}	(0.465) [0.572] {0.519}	(77.0) [80.1] {78.6}
iAMP-2L	(97.7) [91.3] {94.5}	(98.9) [84.5] {91.7}	(98.3) [87.9] {93.1}	(0.965) [0.759] {0.862}	(98.2) [88.1] {93.1}
RF-CFS-GS	(93.7) [96.5] {95.1}	(98.4) [92.2] {**96.0**}	(96.1) [94.4] {**95.5**}	(0.924) [0.888] {**0.911**}	(99.3) [98.6] {**97.6**}
Earth-CFS-GS	(94.1) [88.7] {91.4}	(98.4) [91.4] {94.9}	(96.1) [90.0] {93.1}	(0.924) [0.801] {0.863}	(99.3) [95.7] {97.5}

in an out-of-bag error rate of 9.33%). Ranking is based on the decrease in ACC which is incurred when a feature is removed. This was generated using the `importance` function in the `randomForest` package in R after ensuring results generated are comparable to the Weka implementation. It is interesting to note that the EFC features occur at the ranks of least importance. Since these features are binary and RF was run in the traditional manner to sample with replacement, this could follow observations by Strobl that RF may impart a selection bias against binary features [94].

Top features appear to make biological sense in the context of AMP activity. The top-ranked feature relates to the loss of side chain hydropathy by helix formation which is a conformation many cationic AMPs transition to upon contact with a bacterial membrane [18]. The feature at rank 9 corresponds to net charge which was discussed in Chapter 1 for its importance for attracting an AMP to a membrane surface. Many other of the top features relate to having helix secondary structure which is quite common for many of the AMPs in the data set. We note of the top 10 features, 6 are full global-average features, 3 are for the NT functional region and only 1 is for the CT functional region and is one of

the new SASA features for apolar side-chains.

Table 6.12: Shown below are the CFS-GS features ranked by RF on the Xiao training data set according to the decrease in ACC incurred when a feature is removed. Columns from left-to-right show the rank (ordered from 1 as most important to 74 as least-important), feature name, and the decrease in ACC incurred when the feature is removed. Additional descriptions and sources of these features are available in Table 6.3.2.

Rank	**Name**	**Decrease in ACC**
1	FULL_ROSM880103	10.3574802
2	NT_DAYM780201	9.3814786
3	NT_BUNA790103	8.5353082
4	FULL_AcidicMolPerc	8.1422359
5	NT_CHOP780207	7.7491338
6	FULL_FINA910101	7.657823
7	CT_SCL_A	7.594106
8	FULL_GEOR030101	7.5725872
9	FULL_KLEP840101	7.4724213
10	FULL_YUTK870103	7.1932824
11	FULL_AURR980105	7.1194716
12	FULL_DAYM780201	7.0940486
13	FULL_CHOP780204	7.0752303
14	NT_SCL_A	7.0159069
15	NT_MAXF760105	6.9045459
16	FULL_OOBM850104	6.8144778
17	NT_BBU_A	6.6410383
18	FULL_ROBB760107	6.5996323
19	FULL_CHOP780207	6.5850119
20	FULL_in_vitro_Aggregation	6.5130744
21	NT_HUTJ700101	6.5062518
22	CT_SCL_All	6.4194938
23	NT_GEIM800104	6.2780921
24	NT_KHAG800101	6.1084665
25	FULL_QIAN880136	6.0700374
26	FULL_AURR980106	6.0466086
27	Helix_Propensity	6.0351803
28	FULL_SCL_A	6.033634
29	NT_CHOP780215	6.013576
30	CT_AURR980102	5.9735267
31	NT_GEOR030101	5.9219024
32	FULL_RACS820104	5.9208739
33	FULL_FINA910104	5.872415
34	NT_SCL_All	5.7807886
35	FULL_Beta_Propensity	5.7706932
36	FULL_BBU_A	5.7606794
37	FULL_AURR980107	5.6624518
38	FULL_SCL_All	5.6074702
39	FULL_RACS820101	5.5372257
40	FULL_FUKS010	5.5276168
41	Length_x_in_vitro_Aggregation	5.4414818
42	CT_HOPA770101	5.3585378

43	AS_FINA910101	5.328133
44	NT_WERD780102	5.3267919
45	FULL_SCL_P	5.3055493
46	AS_DAYM780201	5.2159369
47	NT_RICJ880101	5.2134909
48	AS_FAUJ880110	5.1984117
49	NT_NAKH900102	5.1935546
50	FULL_KARP850103	5.0676063
51	FULL_MolWeight	5.0109137
52	FULL_BUNA790103	4.9937246
53	AS_ROSM880103	4.9596397
54	CT_FINA910102	4.8615348
55	FULL_Turn_Propensity	4.7130941
56	FULL_NAKH900102	4.6946247
57	CT_CHOP780204	4.366797
58	NT_NAKH900109	4.0824516
59	FULL_BBU_All	4.0379129
60	FULL_RACS820102	4.0016598
61	FULL_CHAM830102	3.8038285
62	NT_KARP850103	3.5849561
63	CT_WILM950104	3.1582087
64	NT_QIAN880137	2.6645796
65	EFC_33	2.4553139
66	EFC_185	2.3228409
67	EFC_121	2.1949489
68	EFC_134	1.6247798
69	EFC_195	1.6208426
70	EFC_192	1.6014912
71	EFC_37	1.034299
72	EFC_31	0.9944085
73	EFC_59	0.2605553
74	EFC_187	0

6.3.6 MARS Feature Analysis

MARS selects only 30 of the 74 available CFS-GS features provided to it on the Xiao training set. Table 6.13 below shows the rankings for these top 30 features based on the number of times they were selected out of 80 pruning passes run by the algorithm. Being selected more times equates to be higher importance. Table 6.14 shows the terms generated by the method and their coefficients used to build the model after the pruning procedure. Figure 6.2 shows the hinge functions for the 10 first-order terms given by the model.

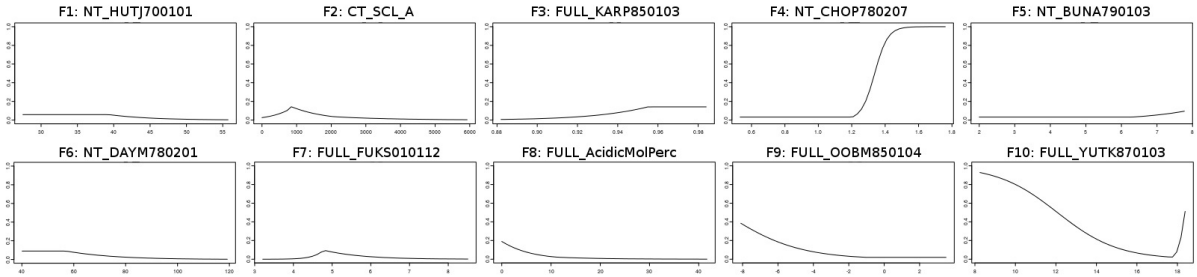

Figure 6.2: Shown are plots of the hinge functions for 10 of the top first-order terms selected by MARS to build our model. Additional descriptions and sources of these features are available in Table 6.3.2.

A few terms such as "CT_SCL_A" and "FULL_AcidicMolPerc" appear at the top of the rankings for both this and the RF method discussed above. However, may of the top features for RF are assigned lower rankings here. The top feature for RF, "FULL_ROSM880103" is in last place here after being selected in only 7 of 80 pruning passes. We can see from Table 6.14 that the feature only appears in terms with feature interactions. While RF may achieve slightly higher performance, this illustrates how MARS provides more transparency in terms of how features are being used. Net charge can be seen dropping to rank 11, with most of the top 10 feature being related to SASA, chain flexibility, helical secondary structure, and aggregation propensity.

Table 6.14: Shown below are the CFS-GS features selected by multivariate adaptive regression slines to build the model. Features are shown as hinge functions and are presented with their coefficients. Interaction terms are separated by a "*" character. Additional descriptions and sources of these features are available in Table 6.3.2.

Feature	**Coefficient**
(Intercept)	1.71226
h(NT_Heat_capacity_.Hutchens_1970._HUTJ700101-39.351)	−0.31245
h(848.6-CT_SCL_A)	−0.00300
h(CT_SCL_A-848.6)	−0.00143
h(CT_SCL_A-2080.4)	0.00062
h(0.955-FULL_Flexibility_parameter_for_two_rigid_neighbors_.Karplus.Schulz_1985._KARP850103)	−43.26888
h(NT_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.204)	22.71642
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29)	3.08162
h(NT_Relative_mutability_.Dayhoff_et_al_1978b._DAYM780201-57.2)	−0.06068

h(4.782-FULL_Entire_chain_compositino_of_amino_acids_in_nuclear_proteins_.percent._.Fukuchi.Nishikawa_2001. FUKS010112)	−4.04626
h(FULL_Entire_chain_compositino_of_amino_acids_in_nuclear_proteins_.percent._.Fukuchi.Nishikawa_2001. FUKS010112-4.782)	−1.11159
h(11.268-FULL_AcidicMolPerc)	0.52454
h(FULL_AcidicMolPerc-11.268)	−0.11352
h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104)	1.79436
h(17.747-FULL_Activation_Gibbs_energy_of_unfolding_pH70_.Yutani_et_al_1987._YUTK870103)	0.62410
h(FULL_Activation_Gibbs_energy_of_unfolding_pH70_.Yutani_et_al_1987._YUTK870103-17.747)	6.08710
h(-0.002-FULL_A_parameter_defined_from_the_residuals_obtained_from_the_best_correlation_of_the_Chou.Fasman _parameter_of_beta.sheet_.Charton.Charton_1983._CHAM830102) * h(17.747-FULL_Activation_Gibbs_energy_of _unfolding_pH70_.Yutani_et_al_1987._YUTK870103)	137.93549
h(FULL_A_parameter_defined_from_the_residuals_obtained_from_the_best_correlation_of_the_Chou.Fasman _parameter_of_beta.sheet_.Charton.Charton_1983._CHAM830102- -0.002) * h(17.747-FULL_Activation_Gibbs_energy_of_unfolding_pH70_.Yutani_et_al_1987._YUTK870103)	6.15644
h(1.22-NT_Relative_preference_value_at_Ntt_.Richardson.Richardson_1988._RICJ880101) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104)	−2.38326
h(NT_Relative_preference_value_at_Ntt_.Richardson.Richardson_1988._RICJ880101-1.22) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104)	−23.63644
h(NT_Heat_capacity_.Hutchens_1970._HUTJ700101-39.351) * h(FULL_Normalized_positional _residue_frequency_at_helix_termini_N2_.Aurora.Rose_1998._AURR980107-0.864)	1.03810
h(NT_Heat_capacity_.Hutchens_1970._HUTJ700101-39.351) * h(0.864-FULL_Normalized_positional_residue _frequency_at_helix_termini_N2_.Aurora.Rose_1998._AURR980107)	−4.57315
h(36.843-FULL_in_vitro_Aggregation) * h(11.268-FULL_AcidicMolPerc)	−0.00630
h(FULL_in_vitro_Aggregation-36.843) * h(11.268-FULL_AcidicMolPerc)	−0.00037
h(1.518-FULL_Information_measure_for_extended_without_H.bond_.Robson.Suzuki_1976._ROBB760107) * h(17.747-FULL_Activation_Gibbs_energy_of_unfolding_pH70_.Yutani_et_al_1987._YUTK870103)	−0.31162
h(1.029-AS_Helix_initiation_parameter_at_posision_i.minus1_.Finkelstein_et_al_1991._FINA910101) * h(NT_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.204)	−261.70492
h(AS_Helix_initiation_parameter_at_posision_i.minus1_.Finkelstein_et_al_1991._FINA910101-1.029) * h(NT_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.204)	−109.94941
h(FULL_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.08) * h(CT_SCL_A-848.6)	0.02488
h(0.932-NT_Linker_propensity_from_all_dataset_.George.Heringa_2003._GEOR030101) * h(FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104- -0.965)	−701.90671
h(1845.7-CT_SCL_All) * h(11.268-FULL_AcidicMolPerc)	−0.00060
h(CT_SCL_All-1845.7) * h(11.268-FULL_AcidicMolPerc)	−0.00007
h(NT_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.204) * h(FULL_SCL_P-749.8)	0.00502
h(NT_Normalized_frequency_of_C.terminal_non_helical_region_.Chou.Fasman_1978b._CHOP780207-1.204) * h(749.8-FULL_SCL_P)	0.10945
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(NT_Normalized_frequency_of_zeta_L_.Maxfield.Scheraga_1976._MAXF760105-1.214)	−12.74806
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(1.214-NT_Normalized_frequency_of_zeta_L_.Maxfield.Scheraga_1976._MAXF760105)	−4.45668
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_AcidicMolPerc-6.25)	−0.17181
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(6.25-FULL_AcidicMolPerc)	−0.09352
h(0.125-FULL_Net_charge_.Klein_et_al_1984._KLEP840101) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b. DAYM780201-57.2)	−0.17058
h(FULL_Net_charge_.Klein_et_al_1984._KLEP840101-0.125) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b. DAYM780201-57.2)	−1.21361
h(0.333-FULL_Net_charge_.Klein_et_al_1984._KLEP840101) * h(11.268-FULL_AcidicMolPerc)	−0.07498
h(FULL_Net_charge_.Klein_et_al_1984._KLEP840101-0.333) * h(11.268-FULL_AcidicMolPerc)	−1.54106
h(1935.4-FULL_SCL_P) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985. OOBM850104)	−0.00068
h(FULL_SCL_P-1935.4) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985. OOBM850104)	−0.00044

h(11.268-FULL_AcidicMolPerc) * h(FULL_Lengthx_FULL_in_vitro_Aggregation-44874)	0.00000
h(1.22-NT_Relative_preference_value_at_Ntt_.Richardson.Richardson_1988._RICJ880101) * h(1917-CT_SCL_All) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104)	0.00347
h(NT_Heat_capacity_.Hutchens_1970._HUTJ700101-39.351) * h(FULL_Linker_propensity_from_all_dataset George.Heringa_2003._GEOR030101-0.998) * h(0.864-FULL_Normalized_positional_residue_frequency_at _helix_termini_N2_.Aurora.Rose_1998._AURR980107)	1480.29783
h(0.797-CT_Normalized_frequency_of_N.terminal_helix_.Chou.Fasman_1978b._CHOP780204) * h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_AcidicMolPerc-6.25)	14.16869
h(736.5-CT_SCL_A) * h(0.125-FULL_Net_charge_.Klein_et_al_1984._KLEP840101) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b._DAYM780201-57.2)	0.00055
h(CT_SCL_A-736.5) * h(0.125-FULL_Net_charge_.Klein_et_al_1984._KLEP840101) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b._DAYM780201-57.2)	0.00012
h(1845.7-CT_SCL_All) * h(2.11-CT_Helix_initiation_parameter_at_posision_ii.plus1i.plus2_.Finkelstein_et_al_1991. _FINA910102) * h(11.268-FULL_AcidicMolPerc)	0.00015
h(CT_SCL_All-1845.7) * h(0.986-FULL_Linker_propensity_from_all_dataset_.George.Heringa_2003._GEOR030101) * h(11.268-FULL_AcidicMolPerc)	0.00489
h(-0.623-CT_Hydrophobicity_coefficient_in_RP.HPLC_C18_with_01percentTFA.2.PrOH.MeCN.H2O_.Wilce_et_al _1995._WILM950104) * h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_AcidicMolPerc-6.25)	1.86818
h(NT_SD_of_AA_composition_of_total_proteins_.Nakashima_et_al_1990._NAKH900102-2.978) * h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_AcidicMolPerc-6.25)	9.70940
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_Loss_of_Side_chain_hydropathy_by_helix_formation_.Roseman_1988._ROSM880103-0.565) * h(FULL_AcidicMolPerc-6.25)	1.70500
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(0.565-FULL_Loss_of_Side_chain_hydropathy_by_helix_formation_.Roseman_1988._ROSM880103) * h(FULL_AcidicMolPerc-6.25)	6.26744
h(NT_Spin.spin_coupling_constants_3JHalpha.NH_.Bundi.Wuthrich_1979._BUNA790103-6.29) * h(FULL_AcidicMolPerc-6.25) * h(65.281-FULL_Relative_mutability_.Dayhoff_et_al_1978b._DAYM780201)	−0.22506
h(0.936-FULL_Linker_propensity_from_all_dataset_.George.Heringa_2003._GEOR030101) * h(FULL_SCL_P-1935.4) * h(-0.965-FULL_Optimized_average_non.bonded_energy_per_atom_.Oobatake_et_al_1985._OOBM850104)	−0.40596
h(0.125-FULL_Net_charge_.Klein_et_al_1984._KLEP840101) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b. DAYM780201-57.2) * h(645.5-FULL_SCL_P)	0.00214
h(FULL_Net_charge_.Klein_et_al_1984._KLEP840101-0.125) * h(NT_Relative_mutability_.Dayhoff_et_al_1978b. _DAYM780201-57.2) * h(FULL_Entire_chain_compositino_of_amino_acids_in_nuclear_proteins _.percent._.Fukuchi.Nishikawa_2001._FUKS010112-5.804)	1.96454

6.4 Conclusion and Chapter Summary

In this chapter we have successfully improved the recognition performance on our model through the inclusion of both SASA and regional features. Comparing LR results on the Xiao testing set using the new CFS-GS feature set to the EFC+307-FCBF model in the previous chapter shows an improvement in MCC of 0.03. These features are also more descriptive and allow us to determine where on an AMP they are more, or less, important. Moreover, using non-linear classifiers such as RF and MARS show dramatic improvements

in terms of ACC and MCC on all of the data sets considered. Exhaustive testing of three different feature selection methods demonstrates how we can maintain excellent recognition using only 74 features with the CFS-GS subset. Finally, we can see in Table 6.11 that our two top models presented in this chapter have better average performance compared to other publicly available AMP predictors.

Table 6.15: Summary of RF-CFS-GS Performance with other AMP Prediction Algorithms and Data Sets

Algorithm	**MCC**			**Database Source**
	Training Data Set	Validation Data Set	Testing Data Set	
HMM [50]		0.98		AMPer
ANN [51]		0.88		RANDOM
DA [44]	0.75		0.74	CAMP
RF [44]	0.86		0.86	CAMP
SVM [44]	0.88		0.82	CAMP
SVM [87]			0.84	AntiBP2
NNA [40]			0.73	CAMP
SVM [52]			0.80	APD
ANFIS [47]		0.94		APD
ANN [47]		0.85		APD
FKNN [41]	0.73		0.84	APD
LR (Ch.4 Model 1) [48]			0.78	APD
LR (Ch.4 Model 4) [49]	0.79		0.82	APD,CAMP
LR (ECB+307-FCBF) [115, 116]			0.86	APD,CAMP
Earth-CFS-GS			0.86	APD,CAMP
RF-CFS-GS			**0.91**	APD,CAMP

This chapter concludes the design of our models for basic AMP recognition. In the following chapter, we reapply some of the basic methods used here in the context of predicting AMP-selectivity.

Table 6.13: Shown are CFS-GS features ranked top-to-bottom in order of importance as determined by the `earth` package `evimp` function. Columns from left-to-right show the feature rank, name, the number of times the feature were selected, the GCV, and RSS attributed to the feature. The times selected value relates to the number of times the feature was selected out of 80 pruning passes performed by the method. A higher number of selections signifies a more important feature. GCV and RSS represent decreases in the respective values between selections containing the feature vs. those not containing it. Both GCV and RSS values are scaled so that the top feature has values of 100. Further details on these metrics are available in the `earth` package documentation [127]. Descriptions and sources of these features are available in Table 6.3.2.

Rank	**Feature**	**Times Selected**	**GCV**	**RSS**
1	CT_SCL_A	58	100	100
2	FULL_AcidicMolPerc	55	57.6	59.6
3	FULL_in_vitro_Aggregation	54	50.9	53.3
4	CT_SCL_All	53	48.8	51.4
5	FULL_OOBM850104	52	46.7	49.3
6	NT_DAYM780201	51	44.1	46.9
7	NT_HUTJ700101	50	42.5	45.4
8	FULL_GEOR030101	49	39.6	42.7
9	FULL_Length_x_*in_vitro*_Aggregation	47	36.3	39.6
10	FULL_YUTK870103	46	34.6	38.1
11	FULL_KLEP840101	44	33	36.5
12	CT_CHOP780204	43	32	35.4
13	NT_BUNA790103	43	32	35.4
14	NT_CHOP780207	42	30.8	34.3
15	NT_NAKH900102	41	29.6	33.3
16	NT_RICJ880101	40	28.8	32.4
17	FULL_FUKS010112	38	27.1	30.8
18	CT_WILM950104	35	24.6	28.4
19	FULL_KARP850103	34	23.8	27.6
20	FULL_SCL_P	32	22.6	26.3
21	FULL_AURR980107	31	21.8	25.5
22	FULL_CHOP780207	23	16.6	20.2
23	FULL_ROBB760107	20	14.8	18.3
24	FULL_DAYM780201	19	14.1	17.6
25	NT_MAXF760105	18	13.4	16.9
26	NT_GEOR030101	17	12.7	16.2
27	CT_FINA910102	15	11.2	14.7
28	AS_FINA910101	12	9.6	12.8
29	FULL_CHAM830102	8	6.8	9.8
30	FULL_ROSM880103	7	5.9	8.9

Chapter 7: Building a Classifier for AMP Selectivity

7.1 Introduction

In this chapter we present models to target general AMP selectivity against Gram-positive (GP) and Gram-negative (GN) bacteria in two experimental settings. The first uses a similar approach to that of Chapter 5 to generate EFC features for separate Gram-specific sets of AMPs. Since there are many AMPs known to target both classes of bacteria [44, 73], we also consider an additional "Gram-both" (GB) model. After demonstrating that these Gram-specialized EFC features are relevant to recognizing antimicrobial activity, we pair them with the global features from Chapter 6 in a new experimental setting. Here we take first steps in building a model to predict if an AMP may perform better against a GN or GP species of bacteria. To accomplish this we compile a carefully constructed data set of AMPs which have been tested against both the GB bacteria *E. coli* and GP bacteria *S. aureus*. These two taxa are some of the most-studied in the AMP literature and currently over 100 AMPs with response data listed for both are currently (as of August 2015) available in the DBAASP [29]. Due to the limited number of peptides with species-specific activity available, in this chapter we only design and evaluate models in the context of CV as we do not have enough observations for separate testing sets.

When selecting features for AMP-selectivity, it is important to consider some of the biological differences AMPs encounter from specific bacterial classes. One major differences between GP and GN bacteria with a direct impact on AMP specificity is the presence of an outer plasma membrane in GN bacteria. The outer membrane contains lipopolysaccharide (LPS) and acts as an additional physical barrier that an AMP must pass through before it can reach the cytoplasmic inner membrane to exert an attack. AMPs and other molecules can typically traverse the LPS layer via porin proteins which physically exclude molecules

greater than approximately 600 Da in size [142]. Since LPS is negatively charged, it can attract and trap small cationic AMPs as observed with temporins [143]. AMPs that do make it to the periplasmic space between membranes can accumulate in high concentrations, potentially promoting AMP aggregation and cooperative attack mechanisms such as the carpet model [18]. An additional barrier that must be crossed and is present in both GP and GN bacteria, is the mesh-like peptidoglycan layer. While this varies in structure by species, GP bacteria typically exhibit wider openings through which AMPs can more easily navigate [18]. Another important biological factor in determining AMP selectivity, particularly for peptides that induce lipid clustering, are differences in the composition of the inner cytoplasmic plasma membrane [55–57]. While GN bacteria usually contain more of the phospholipid phosphatidylethanolamine (PE) compared to GP bacteria, the amount can also vary between species of the same Gram-class [18]. In regards to representative bacteria used in this chapter, the membranes of *E. coli* have a typical composition of 80% PE, 15% phosphatidylglycerol (PG), and 5% cardiolipin. This contrasts greatly with *S. aureus* which typically has 0% PE, 58% PG, and 42% cardiolipin [57]. Differences in the amount of negative or zwitterionic lipids present can strengthen or attenuate the attractive forces which draw a cationic AMP to the membrane surface. Even subtle differences in the lipid ratio between species can result in an AMP being active against one, yet ineffective against another [56, 57].

The above examples highlight the importance for AMP physicochemical properties to match the unique challenges presented by any GP or GN target. Accordingly, in this chapter we perform feature selection separately for each of the GP, GN and GB models to capture unique feature subsets for each. We must also accept limitations in the data we have available to us when designing an experiment to predict AMP selectivity. Many additional factors influence both AMP activity and reported responses, such as: environmental pH or salt concentration, the type and manner in which AMP lethality is calculated (e.g. MIC, EC50, LD50, etc.), and even the basic experimental design used to test AMPs [10,12,13,18].

In all of the AMP databases we consider (APD, CAMP and DBAASP), the availability of such environmental information is both rare and highly inconsistent across peptide records [18, 29, 44, 73]. While an alternative approach like molecular dynamics is capable of controlling for some environmental variable like salt concentration, it is currently computationally prohibitive beyond only a few peptides and verification from the wet lab would still be required. Consequently, for experiments in this chapter which tend to have smaller sample sizes we model general relationships of activity and avoid trying to predict an exact response (i.e. an exact MIC value). As a starting point for this new type of experimental setting we continue with binary responses. For the design of the Gram-specialized EFC features we utilize separate data sets and continue using non-AMP (0) vs. AMP (1) responses. For predicting better AMP activity between our representative bacteria, we assign 0 to peptides with lower (better) median MIC values against *E. coli* and 1 for those against *S. aureus*.

We begin with a description of methods and implementation details for each of the experimental settings. This is followed by performance results and an analysis of feature rankings for each of the separate models. We note work on the design of Gram-specialized EFC features and related models was first presented in [116] and co-authored with Dr. Uday Kamath and Dr. Amarda Shehu. Code specific for the EFC framework is written by UK. Selection of data sets and features, feature representation, biological analysis and interpretation is performed by the author. Definitions for the new Gram-specific EFC features presented in this chapter are provided in Appendix C.

7.2 Methods

We begin by describing the generation of Gram-specialized EFC features. This is performed separately for the GP, GN and GB data sets which are detailed in Chapter 3. The FCBF algorithm introduced in Chapter 5 is again used to select small subsets of features sharing low redundancy. Selected features are paired with a variety of classifiers to build models

for AMP recognition. After assessing general model performance, we present visual representations of feature importance through the use of decision trees. A comparative analysis at the end of the chapter highlights some EFC features which are shared between the GP, GN and GB feature subsets. The new EFC features are then combined with others from the previous chapter for use in the next experimental setting. These experiments utilize a carefully-constructed data set (also detailed in Chapter 3) based on relative MIC values for use in predicting if an AMP may demonstrate better activity against *E. coli* vs. *S. aureus*. We refer to this as the "ECvSA" experimental setting throughout the rest of the chapter. FCBF is again used for feature selection, and model performance and feature rankings are reported as in the first setting. For the final ECvSA model, we also perform one additional feature ranking procedure using the R implementation of RF as in Chapter 6. While the EFC features generated in the first setting are designed to discriminate between AMPs and non-AMPs, we try them with the ECvSA setting to see if GN and/or GP features may be more prevalent in AMPs better against *E. coli* or *S. aureus* respectively. As all AMPs in this data set are verified against at least one GP and GN bacteria, in the ECvSA setting we would expect GB features to be prevalent but, not necessarily, discriminatory.

All models presented in this chapter are evaluated using supervised classification in the context of 10-fold CV. Feature reduction is applied separately on each training fold and we report average performance obtained on the respective testing folds. Feature analysis and ranking is based on final selections made after running the FCBF algorithm on the full data set. All references to Weka [117] are for Vr. 3.7. The follow-up analysis of features using RF is performed in R Vr. 3.2.2. Discussions about EFC features refer to the GBMR4 alphabet introduced in Chapter 5, and outlined in Table 5.1.

7.2.1 Gram-specialized Evolutionary Feature Construction

For the first experimental setting, we reapply the EFC feature generation method introduced in Chapter 5 separately for the GP, GN and GB data sets detailed in Chapter 3. AMPs from the three Gram-specific data sets are separately paired with the same 2368 non-AMPs from

the Xiao training data to generate EFC features. This is done to generate distal features which are representative of Gram-specific AMPs. The positive examples in the GN, GP and GB sets are each comprised of 128, 271 and 1103 AMP observations. Compared to the > 2500 total records available in the APD, these numbers reflect how relatively few GP- and, in particular, GN-specific AMPs have been characterized. Due to the small number of positive samples available, we run the EFC generation method 3 times for each data set (as opposed to 10 in Chapter 5) since generating too many specialized features may risk overfitting the data. We note the large imbalance in sample size between AMP and non-AMPs is not a major concern for EFC feature generation as the fitness function only tracks the occurrence of features in AMPs. Non-AMP sequences are only used to penalize features which are found non-discriminating.

7.2.2 Evaluation of Features and Performance Measurements for Gram-Based Models

To ensure that the new EFC features are relevant to AMP activity, we assess AMP recognition performance separately for each of the three data sets in the context of 10-fold CV and pair them with the same 307 global (whole-peptide) features used in Chapter 5. We also now match the AMPs observations with the 897 non-AMPs in the Xiao testing data set. This allows for a more direct comparison to LR results from Chapter 5 and has the added benefit of making the class sizes more balanced. Using Weka's `AttributeSelectedClassifier` option, we run FCBF feature selection on the 9 training folds and assess performance on the remaining testing fold and repeat until each fold is used once for testing.

Classification is performed with LR, NB, J48 (C4.5) and RF classifiers as in previous chapters. We take advantage of the single decision tree output by the J48 algorithm to produce interpretable visualizations of features for each of the Gram-based models. Direct comparison between models of LR coefficients for individual features is avoided, as this approach is known to be influenced by hidden heterogeneity in the underlying data [144,145]. Finally, we present an IG analysis of features selected by FCBF and compare the new EFC

features generated for the GP, GN and GB data sets at the end of the chapter.

In addition to the auROC and MCC measurements used previously, in this chapter we also report auPRC values since classes for some data sets are unbalanced. All experiments are replicated 100 times and standard deviations (SD) are reported.

7.2.3 Predicting AMP Selectivity for *E. coli* vs. *S. aureus*

The ECvSA experimental setting focuses on models to predict if an AMP may perform better against *E. coli* or *S. aureus*. We use a data set of 124 AMPs taken from the DBAASP with reported experimental MIC values for both taxa. As detailed in Chapter 3, responses for this data set are encoded as 0 if a peptide performs better against *E. coli* ($n = 62$), and 1 if they perform better against *S. aureus* ($n = 62$). Sequences are encoded using the CFS features from Chapter 6, however, we replace the original 208 EFC features with 208 new ones combined from the new GB, GN and GP models described above. Feature selection with FCBF and model generation using LR, NB, J48 and RF is then performed in the context of supervised 10-fold CV. Selected features are again visualized using a J48 decision tree, however, in addition to ranking features by IG, we also assess feature importance using RF based on the mean decrease in ACC when an individual feature is removed. Performance metrics in this setting also use auPRC, auROC and MCC along with SD values after replicating experiments 100 times.

7.2.4 Implementation Details

All experiments in this chapter are performed on an Intel i5 quad-core machine with 3.2 Ghz CPU and 8GB of RAM. EFC is run 3 times for each of the Gram-based data sets and average results with standard deviations are reported. One run of EFC takes about 1 hour of CPU time. As in Chapter 5, the maximum motif length in EFC is set to $k = 8$ with other parameters set as follows: $n = 10,000$, $D = 5$, $r = 10$, $G = 3$, $\ell = 500$, and $m = 100$. The mutation and crossover operators are performed with probability 0.3 and 0.7, respectively. Weka is used to apply FCBF to EFC-obtained features in the hall of

fame and also to select a subset of features after each EFC run. The method is run with *numToSelect*=−1 and using the *SymmetricalUncertAttributeSetEval* option. Using FCBF for EFC feature selection typically takes $2-5$ minutes of CPU time. The final GP, GN and GB predictive models are built using the EFC+307-FCBF method introduced in Chapter 5.

In order to evaluate the models in the context of 10-fold CV, the FCBF method is applied using the *AttributeSelectedClassifier* option so that features are selected based on the training folds and average performance is reported on the testing folds. SD values are provided as experiments are repeated 100 times.

Predictive models are built in Weka using the following classifiers and default settings: LR with $ridge = 1.0\text{E-}8$, J48 with $confidenceFactor = 0.25$ and $minNumObj = 2$, RF with $numTrees = 50, 100, 200, 500$ and $numFeatures = log_2(n \times Features) + 1$. Each method typically takes $5-10$ min of CPU time. J48 is run once on each data set after feature selection with FCBF to generate the decision tree figures.

For the ECvSA experimental setting, models are built and tested using the same classifiers and parameters. The additional feature ranking procedure employs the `importance` function (with $type = 1$, $scale = FALSE$) built into the `randomForest` package Vr. 4.6-10 in R. We use the RF parameters $numTrees = 100$ and $mtry = 4$ and testing between implementations in Weka and R show similar results.

7.3 Results

7.3.1 Gram-specific AMP Feature Sets

Our first experimental setting investigates the ability of our EFC+307-FCBF method from Chapter 5 to extend recognition beyond AMPs in general, and apply them to AMPs active against Gram-specific classes of bacteria. We employ three separate positive data sets. These are first paired with the Xiao non-AMP training set to generate EFC features and later paired with the Xiao non-AMP testing set to build and evaluate models using LR,

NB, J48 and RF classifiers. Our experiments below are performed in the context of 10-fold CV and replicated 100 times.

Performance Summary of Gram-based Models

For each of the three separate Gram-based settings, the positive (GP, GN, or GB) data set is paired with the Xiao negative (non-AMP) training data set. All unique hall-of-fame features obtained after repeating the EFC method 3 times (per Gram-based data set) are combined and paired with the 307 physicochemical features described in Chapter 5. FCBF is applied to reduce the number of features and average performance is reported using LR in a 10-fold CV setting as summarized in Figure 7.1. When the FCBF algorithm is applied to the full-size data sets using the Xiao negative training data, 82 features are selected for GP, 91 for GN, and 54 for GB. Using LR, resulting auROC values range from 90.3%-92.6%, MCC values from 0.58-0.69, and auPRC values from 80.5%-92.4%.

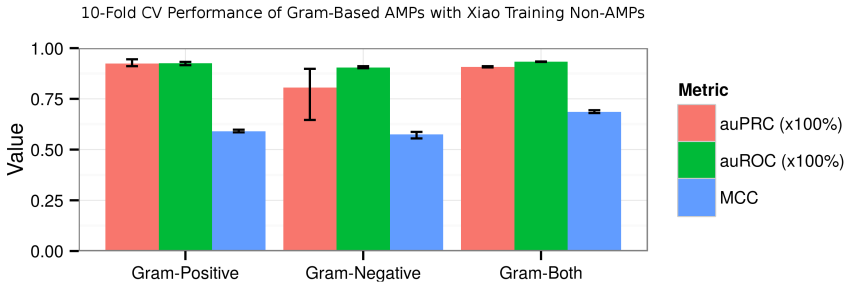

Figure 7.1: Average recognition performance of EFC+307-FCBF features using LR and 10-fold CV is shown for three separate Gram-specific AMP data sets combined with the Xiao training non-AMPs. Performance values are as follows (SD in parenthesis), **GP**: $auPRC = 92\%$ (0.02), $auROC = 93\%$ (0.01), $MCC = 0.59$ (0.01). **GN**: $auPRC = 81\%$ (0.16), $auROC = 90\%$ (0.01), $MCC = 0.58$ (0.02). **GB**: $auPRC = 91\%$ (0.03), $auROC = 93\%$ (0.00), $MCC = 0.69$ (0.01).

The performance of the features in each Gram-based setting is similarly high, as seen

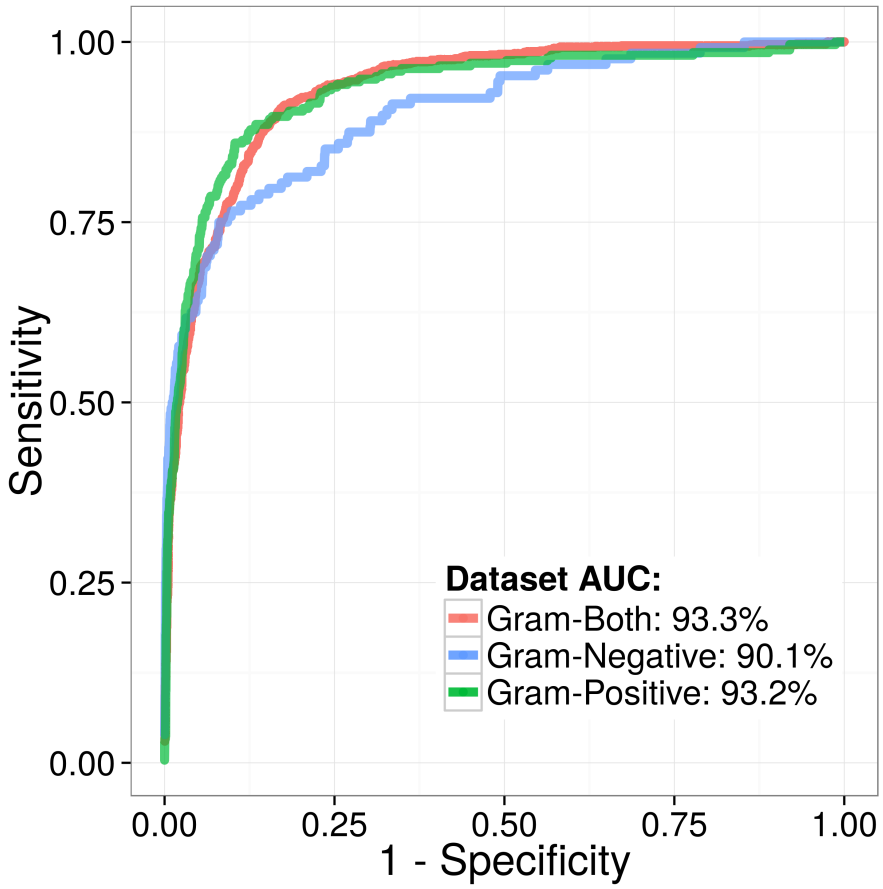

Figure 7.2: An ROC curve depicting 10-Fold CV average performance with LR for each of the Gram-specific models when paired with the Xiao training non-AMPs.

in Table 7.1, when the Xiao negative training set is replaced by the negative testing set. For the GP setting, the 10-fold CV performance by LR and the RF classifier (using 500 trees) yields auPRC values of 99.0% and 99.6%, auROC values of 98.0% and 99.0% and MCC values of 0.89 and 0.92 for LR and RF respectively. With the GN setting, auPRC values of 99.4% and 99.6%, auROC values of 97.1% and 98.1%, and MCC values of 0.82 and 0.86 are obtained for LR and RF. Finally, on the GB setting, auPRC values of 96.9% and 98.6%, auROC values of 98.0% and 99.0%, and MCC values of 0.87 and 0.91 are obtained for LR and RF. As results using LR show MCCs ranging between 0.82-0.89, performance appears comparable to the MCC of 0.86 achieved by EFC+307-FCBF on generalized AMPs in Chapter 5.

Table 7.1: Average recognition performance of EFC+307-FCBF features using different classifiers is shown when Gram-specific AMPs are paired with the Xiao negative testing data. Results are reported using auPRC, auROC, and MCC in the context of 10-fold CV with SD values for 100 experimental replications shown in parenthesis. Weka defaults for RF were used to selected the following number of features to generate trees: GP= 5, GN= 4 and GB= 5. Bold font is used to highlight best performance on a specific metric per column.

Method	**auPRC (%)** *GP* | *GN* | *GB*	**auROC (%)** *GP* | *GN* | *GB*	**MCC** *GP* | *GN* | *GB*
LR	99.0(0.01) | 99.4(0.01) | 96.9(0.02)	98.0(0.02) | 97.1(0.03) | 98.0(0.01)	0.887(0.051) | 0.822(0.084) | 0.870(0.034)
NB	98.8(0.01) | 98.3(0.01) | 96.6(0.02)	97.1(0.02) | 91.8(0.05) | 97.4(0.01)	0.814(0.060) | 0.656(0.119) | 0.843(0.038)
J48	95.7(0.02) | 96.2(0.02) | 89.2(0.04)	92.2(0.04) | 86.2(0.08) | 92.9(0.03)	0.867(0.057) | 0.778(0.093) | 0.869(0.036)
RF-50	99.4(0.01) | 99.4(0.01) | 98.3(0.01)	98.8(0.01) | 97.7(0.03) | 98.9(0.01)	0.916(0.042) | 0.851(0.076) | 0.909(0.029)
RF-100	99.5(0.01) | 99.5(0.01) | 98.5(0.01)	98.9(0.01) | 98.0(0.03) | **99.0**(0.01)	0.920(0.041) | 0.854(0.076) | 0.911(0.028)
RF-150	**99.6**(0.01) | 99.5(0.01) | 98.5(0.01)	**99.0**(0.01) | 98.0(0.02) | **99.0**(0.01)	0.921(0.041) | 0.854(0.076) | **0.912**(0.028)
RF-200	**99.6**(0.00) | 99.5(0.01) | 98.5(0.01)	**99.0**(0.01) | 98.0(0.02) | **99.0**(0.01)	**0.922**(0.041) | **0.855**(0.076) | **0.912**(0.029)
RF-500	**99.6**(0.00) | **99.6**(0.01) | **98.6**(0.01)	**99.0**(0.01) | **98.1**(0.02) | **99.0**(0.01)	**0.922**(0.041) | **0.855**(0.074) | **0.912**(0.028)

Results in Table 7.1 show that EFC+307-FCBF features can achieve good recognition performance when using Gram-specific AMPs. Applying FCBF to entire feature sets with Xiao negative testing data reduces the number of features to: 21 for GP, 10 for GN, and 16 for GB. We next conduct a detailed analysis of the reduced features obtained for each of the target-specific AMP data sets. First, the features are analyzed in terms of their information gain. Second, the importance of features for each data set is visualized through

a decision tree (generated by the J48 implementation in Weka). Common features found between Gram-based subsets are discussed at the end of the chapter.

Detailed Analysis of Gram-Positive Reduced Feature Set

The information gain of each of the features in the GP reduced feature set is shown in Table 7.2. The importance of features can also be visualized through a decision tree generated by the J48 algorithm in Weka for all GP AMPs and Xiao testing non-AMPs. Figure 7.3 shows that EFC-based features dominate the top of the tree, recognizing 221 of 271 AMPs in the data set before a physicochemical feature (AAIndex: KARP850103) first appears at level 9. Specifically, the first feature (ranked second by IG) looks for the motif **AC** at position 8 ± 3 *and* motif **CGA** anywhere else on the peptide and accounts for 111 AMPs in the data set. A survey of AMPs captured by this feature in the APD reveals that they are broad in scope in terms of both family and structure. 16 out of the 25 temporin-family AMPs in the *GP* set are accounted for, but the remaining 95 AMPs are listed under 76 different AMP names or families. In terms of peptide structure, 78 are listed as 'unknown' while the remaining fall under 'helix' (17), disulfide bond 'bridge' (10), 'beta' (5) and 'combine helix and beta structure' (1). Overall, physicochemical features tend to promote helical or flexible structures. While the J48 tree only utilizes 12 out of the 21 selected features listed in Table 7.2, the unused features appear important for recognition by other classifiers.

Detailed Analysis of Gram-Negative Reduced Feature Set

Initially 19 features were selected after running FCBF to reduce the size of the full GN feature set. However, further analysis showed that 9 of these features could be removed without impacting classification performance using LR in the context of 10-fold CV. The information gain of these remaining 10 features in the GN reduced feature set is shown in Table 7.3.

The importance of features is additionally visualized through a J48 decision tree using all GN AMPs and the Xiao testing non-AMPs as shown in Figure 7.4. The overall structure

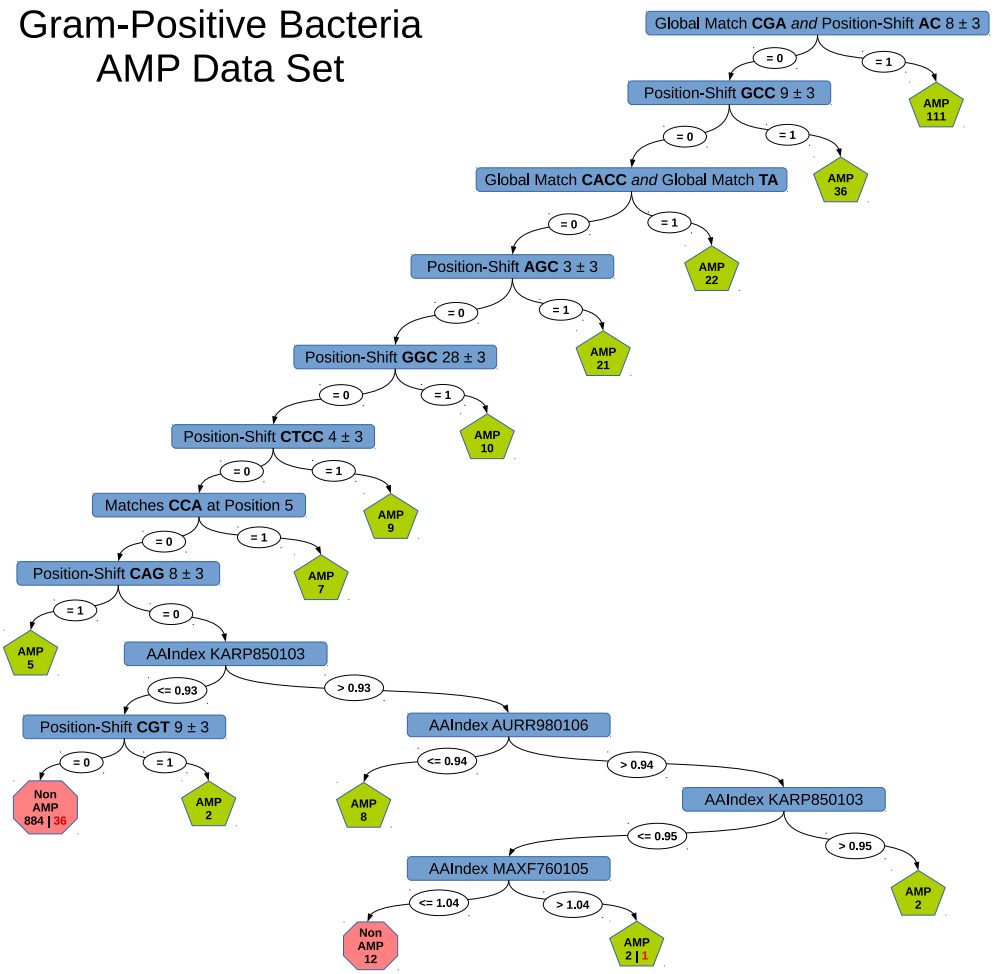

Figure 7.3: A binary decision tree based on a reduced set of 21 features generated from a data set of AMPs active against Gram-positive bacteria is shown here. The tree was generated using the J48 (C4.5) classifier in Weka using all GP AMPs and the Xiao testing non-AMPs ($TP = 271, TN = 897$); some features were not utilized by the classifier. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown in white ovals). Terminating leaves in green classify a query as an AMP, while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label.

of the decision tree appears similar to the tree for the *GP* data set, with EFC features dominating the upper levels of the tree and accounting for 65 out of 128 AMPs before the first physicochemical feature (AAIndex: RACS820101) is encountered at level 6. The top feature (ranked third by IG) looks for the motif **CGA** at position 10 ± 3 and a check of the AMPs captured by this feature in the APD shows they are broad in scope for family with structure mostly unknown. 4 out of the 9 microcin-family AMPs in the *GN* set are recognized but the remaining 20 AMPs are listed under 19 different names or families. A survey of peptide structure shows 19 listed as 'unknown,' 2 labeled disulfide bond 'bridge' and 2 having a 'beta' structure. Similar to the *GP* case, physicochemical features again promote helical or non-rigid structures. While the J48 tree only utilizes 9 out of the 10 selected features listed in Table 7.3, the unused feature appears important for recognition using other classifiers. While the letter **T** (proline, a helix-breaker) occurs the least in motifs across all data sets, it is interesting to note it occurs only once in the *GN* reduced feature set.

Detailed Analysis of Gram-Both Reduced Feature Set

The information gain of each of the 16 features in the GB reduced feature set is shown in Table 7.4. The importance of features is also visualized through a decision tree. The J48 algorithm in Weka was used to produce a single decision tree using all of the GB AMPs and Xiao testing non-AMPs and can be seen in Figure 7.5.

Unlike the previous two cases, a non-EFC feature (peptide length) appears as the first feature and splits the nodes into two major subtrees. If a query peptide is > 51 AA long, it encounters a subtree more similar to those produced by the *GP* and *GN* data sets, with EFC features dominating higher levels and physicochemical features at lower ones. For peptides ≤ 51 residues long, the subtree is a mixture of non-EFC features and EFC features which target residues at the N-terminus. While the J48 tree only utilizes 13 out of the 16 selected features listed in Table 7.4, the unused features appear important for recognition using other classifiers. We caution that length is a feature that needs to be considered carefully,

Table 7.2: The 21 EFC+307-FCBF features used in the *GP* reduced feature set are ranked here by their information gain, shown in column 2. The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 5.1.

Rank	IG	Source	Description
1	0.260	AAIndex: AURR980106	Normalized positional residue frequency at helix termini N1 [146]
2	0.227	EFC: Global Motif AND Position-Shift	CGA at any position **and** AC at position 8 ± 3
3	0.171	AAIndex: YUTK870103	Activation Gibbs energy of unfolding pH 7.0 [147]
4	0.127	One of 8 features from [46]	*In vitro* peptide aggregation from Tango Server [101]
5	0.099	AAIndex: MAXF760105	Normalized frequency of ζ_L [148]
6	0.097	EFC: Position-Shift	GCC at position 9 ± 3
7	0.095	EFC: Motif AND Motif	CACC **and** TA at any positions
8	0.085	EFC: Position-Shift	AGC at position 3 ± 3
9	0.071	EFC: Position-Shift	CAG at position 8 ± 3
10	0.058	EFC: Match Position	CCA at position 5
11	0.058	EFC: Position-Shift	CTCC at position 4 ± 3
12	0.039	EFC: Position-Shift	GGC at position 28 ± 3
13	0.037	EFC: Position-Shift	AAAA at position 33 ± 3
14	0.028	AAIndex: KARP850103	Flexibility parameter for two rigid neighbors [149]
15	0.015	EFC: Position-Shift	CGT at position 9 ± 3
16	0.002	EFC: Global Motif	GTACACA at any position
17	0.002	EFC: Position-Shift	GCCTGA at position 1 ± 3
18	0.002	EFC: Position-Shift	TATCAT at position 10 ± 3
19	0.002	EFC: Position-Shift	TTATT at position 7 ± 3
20	0.002	EFC: Position-Shift	TGCCAA at position 44 ± 3
21	0.002	EFC: Global Motif	TCTCAT at any position

Table 7.3: The 10 EFC+307-FCBF features used in the *GN* reduced feature set are ranked here by their information gain, shown in column 2. The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino-acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 5.1.

Rank	IG	Source	Description
1	0.097	AAIndex: RICJ880104	Relative preference value at N1 of alpha helix [150]
2	0.086	AAIndex: RACS820101	Average relative fractional occurrence in A0(i) [151]
3	0.073	EFC: Position-Shift	CGA at position 10 ± 3
4	0.065	AAIndex: KARP850103	Flexibility parameter for two rigid neighbors [149]
5	0.064	EFC: Position-Shift	GCC at position 13 ± 3
6	0.054	EFC: Position-Shift	GCA at position 9 ± 3
7	0.027	EFC: Match Position	CAC at position 3
8	0.021	EFC: Position-Shift	GGC at position 13 ± 3
9	0.012	EFC: Match Position	TC at position 3
10	0.009	EFC: Global Motif	GGGGGG at any position

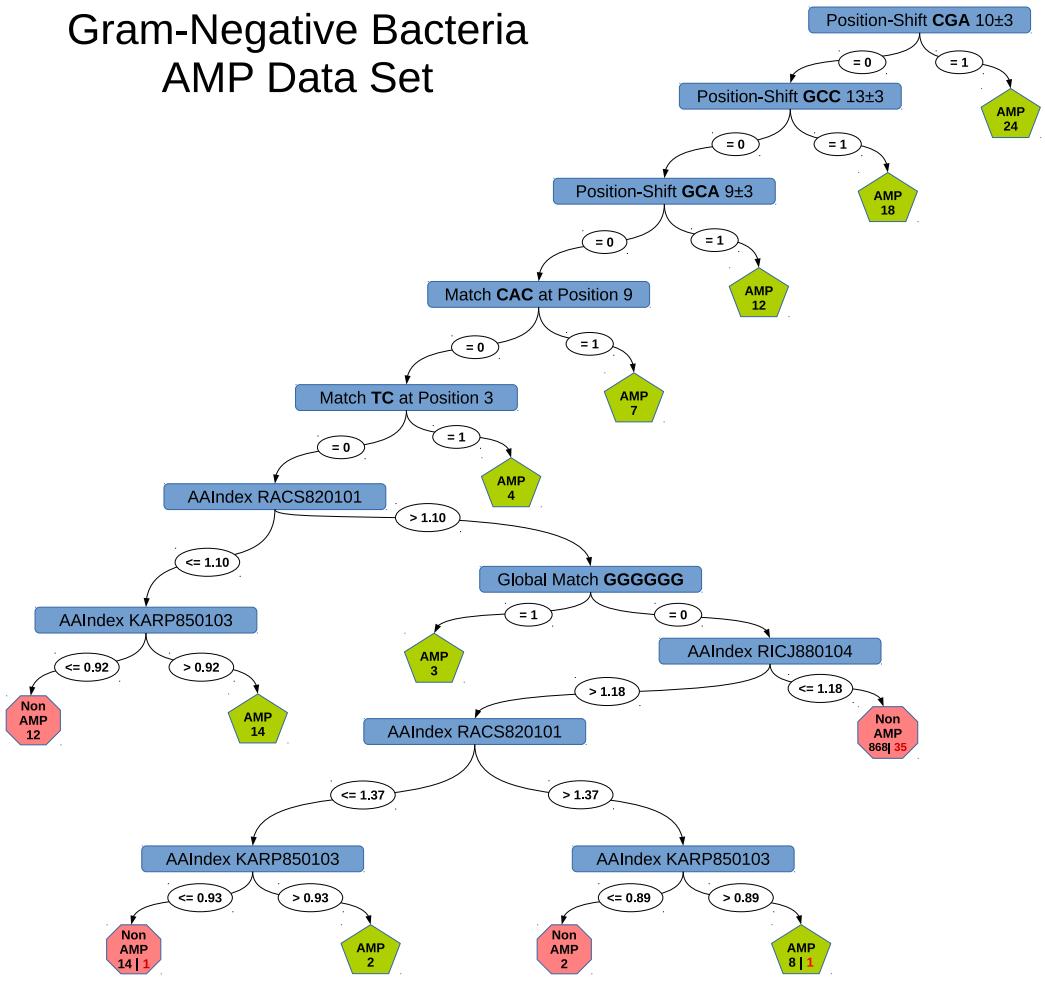

Figure 7.4: A binary decision tree based on a reduced set of 10 features generated from a data set of AMPs active against Gram-negative bacteria. The tree was generated using the J48 classifier in Weka using all GN AMPs and the Xiao testing non-AMPs ($TP = 128$, $TN = 897$); some features were not utilized by the classifier to build the tree. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown as white ovals). Terminating leaves in green classify a query as an AMP while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label.

Table 7.4: The 16 EFC+307-FCBF features used in the *GB* reduced feature set are ranked here by their information gain, shown on column 2. The source of the feature is shown in column 3, and a description of the feature is provided in column 4. Amino-acid position numbers start with 0 for the first residue, and motifs are shown in GBMR4 format as detailed in Table 5.1.

Rank	IG	Source	Description
1	0.650	One of 8 features from [46]	Peptide Length
2	0.306	AAIndex: AURR980107	Normalized positional residue frequency at helix termini N2 [146]
3	0.298	AAIndex: KLEP840101	Net charge [152]
4	0.222	One of 8 features from [46]	*In vitro* peptide aggregation from Tango Server [101]
5	0.210	EFC: Position-Shift	GC at position 0 ± 3
6	0.167	AAIndex: GEOR030101	Linker propensity from all data set [153]
7	0.159	EFC: Position-Shift	CG at position 5 ± 3
8	0.155	EFC: Position-Shift	CCAA at position 5 ± 3
9	0.077	EFC: Position-Shift	AGC at position 8 ± 3
10	0.035	EFC: Position-Shift	AGCC at position 15 ± 3
11	0.028	EFC: Motif AND Motif	AAAA **and** TACA at any positions
12	0.021	EFC: Position-Shift	GGC at position 11 ± 3
13	0.013	EFC: Match Position	TC at position 3
14	0.011	EFC: Position-Shift	CGA at position 44 ± 3
15	0.006	EFC: Correlate Positions	AT at position 1 **and** TC within 3 positions before/after
16	0.001	EFC: Global Motif	TGCCG at any position

as an AMP peptide can contain shorter fragments which are themselves antimicrobial [154]. Removing *length* as a feature generates a decision tree with a topology more similar to the other data sets and EFC features at higher levels. In this case the feature **CG** at position 0 ± 3 takes the top position and recognizes almost half of the 1103 AMPs. The feature **GGC** at position 11 ± 3, absent from the tree in Figure 7.5, also gets incorporated if *length* is removed.

7.3.2 Summary of *E. coli* vs. *S. aureus* Models

For the ECvSA setting, a total of 1484 CFS features using the new Gram-based EFC features are used to encode 124 AMPs tested against both *E. coli* and *S. aureus* as detailed in Chapter 3. Binary responses based on having a lower median MIC value are evenly split with $n = 62$ for each taxa. In terms of new EFC features with at least one match in the dataset, 17 hits are found for GN, 24 hits for GP, and 17 for GB. Table 7.5 below shows leave-one-out CV performance when using FCBF on a number of classifiers. Performance in this setting is slightly lower than seen above for recognizing AMPs from non-AMPs. LR and RF (using 200 trees) obtain ACCs of 71.0% and 83.1%, auPRCs of 72.9% and 83.8%,

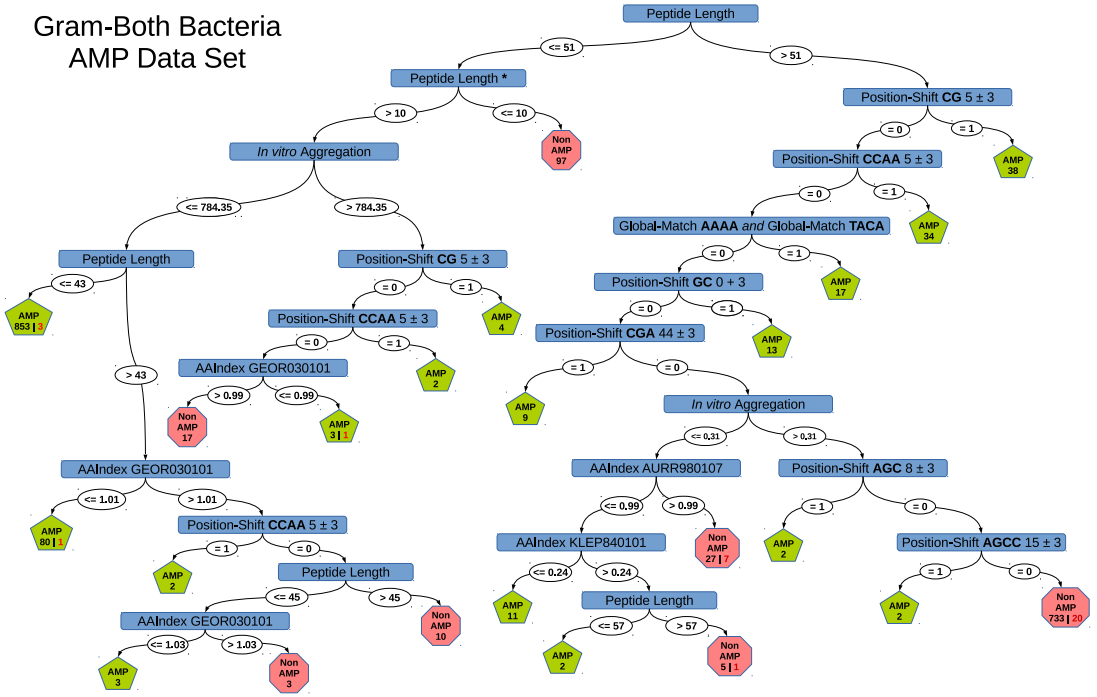

Figure 7.5: A binary decision tree based on a reduced set of 16 features generated from a data set of AMPs active against both Gram-positive and Gram-negative bacteria. The tree was generated using the J48 classifier in Weka using all GB AMPs and the Xiao testing non-AMPs ($TP = 1103$, $TN = 897$). Some features were not utilized by the classifier to build the tree. Passing a GBMR4-encoded query peptide through the tree starting at the top will classify it as either an AMP or Non-AMP. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown as white ovals). Terminating leaves in green classify a query as an AMP while those in red represent a Non-AMP. Numbers in black and red represent the number of instances which stop at a given leaf and respectively assign it a correct or incorrect label. We note, while no AMPs with a length less than 11 were included in this data set, some are listed in the APD and CAMP databases. Accordingly, adding additional rules to the ≤ 10 branch of the "Peptide Length" node denoted by $*$ may further improve classification performance on data sets considering shorter AMPs.

and MCCs of 0.420 and 0.662, respectively. As a first step in this new experimental setting performance looks encouraging, however, models should be reevaluated when a larger sample size of AMPs tested against both taxa becomes available. We next look at features selected by FCBF and their relative importance.

Table 7.5: Results are reported for the ECvSA setting using leave-one-out CV and classifiers implemented in Weka using FCBF feature selection. Performance is shown in terms of ACC, auPRC, auROC, and MCC. The number of features selected by RF to build trees is 4 and out of bag error ranges between 16.1-19.5%. Bold font is used to highlight best performance on a specific metric per column.

Method	**ACC (%)**	**auPRC (%)**	**auROC (%)**	**MCC**
LR	71.0	72.9	74.5	0.420
NB	72.6	78.8	79.0	0.452
J48	79.0	79.5	74.7	0.592
RF-50	81.5	83.7	85.3	0.631
RF-100	81.5	83.2	85.5	0.631
RF-150	82.3	**83.8**	**86.0**	0.647
RF-200	**83.1**	**83.8**	85.7	**0.662**
RF-500	81.5	83.2	85.5	0.631

FCBF selected 11 features which are listed in Table 7.6 in order from high-to-low IG value. Relative ranks based on the mean decrease in ACC (using the `scale` option) when these features are used with RF (using $mtry = 4$ and $ntrees = 200$) are also shown in column 3. Features a ranks 1, 4, 10 and 11 are in agreement between the two methods while other ranks can differ by as many as 3 positions. All features selected by FCBF come from the AAIndex [119], with the exception of the SASA feature at IG rank 5 from [128]. Reoccurring topics amongst theses features are hydropathy (ranked most important), folding, and phase transitions which are known important for AMP interactions with membranes [12, 18, 34, 35, 56, 58]. We note none of the new EFC features from the Gram-specific experiment above were selected by FCBF for this data set. Considering the small sample size and the fact

these EFC features are generated for a different context (comparing AMPs vs. non-AMPs), this is not necessarily surprising. Given more samples to work with in the future, it will be interesting to see if certain EFC features related to the GN set occur more frequently in AMPs better against *E. coli* or if those in the GP set occur more often in peptides better against *S. aureus*.

Table 7.6: The 11 Features selected for the ECvSA setting are shown and ranked by IG and relative importance for RF. Relative rank based on IG is shown in column 1, the IG score in column 2, the relative rank based on RF importance in column 3, the mean decrease in ACC when the feature is removed in column 4, the feature name in column 5 and a feature description in column 6. The two ranking methods are only in agreement for ranks 1, 4, 10 and 11. Values in column 2 were calculated with Weka, those in column 4 were generated by the *importance* function in the R `randomForest` package with `scale=True`.

IG Rank	**IG**	**RF Rank**	**Mean Decrease in ACC**	**Feature**	**Description**
1	0.262	1	10.45	NT_NADH010101	Hydropathy scale based on self-information values in the two-state model (5% accessibility) [155] at the N-terminus
2	0.206	5	5.94	NT_HUTJ700102	Absolute entropy (Hutchens 1970) at the N-terminus
3	0.188	2	8.84	NT_RICJ880114	Relative preference value at C1 (Richardson-Richardson 1988) at the N-terminus
4	0.165	4	7.29	NT_FASG760104	pKN (Fasman 1976) at the N-termini at the N-terminus
5	0.133	3	8.74	FULL_BBU_Polar	Average SASA for polar backbone atoms (upper-bound) [128] for the whole peptide
6	0.133	7	4.92	NT_PALJ810111	Normalized frequency of beta-sheet in $\alpha + \beta$ class (Palau et al. 1981) at the N-terminus
7	0.114	8	4.92	AT_WILM950103	Hydrophobicity coefficient in RP-HPLC C4 with 01%TFAMeCNH2O (Wilce et al. 1995) about the most amphipathic window
8	0.114	9	3.71	NT_CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman 1978b) about the N-terminus
9	0.114	6	5.10	CT_NAKH900110	Normalized composition of membrane proteins [156] about the C-terminus
10	0.104	10	2.38	CT_YUTK870103	Activation Gibbs energy of unfolding pH70 (Yutani et al. 1987) about the C-terminus
11	0.095	11	0.75	NT_HUTJ700101	Heat capacity (Hutchens 1970) about the N-terminus

Figure 7.6 below shows a decision tree generated by J48 for the features shown in Table 7.6. Nodes near the top of the tree appear to select for *S. aureus*. The top-ranked N-terminal feature NADH010101 appears as the top node in the tree. It assigns an *S. aureus* label to 27 AMPs with a value > 45.4 using a hydropathy scale developed by Naderi-Manesh et al. in [155] to predict surface accessibility. It is interesting the J48 places the C-terminal feature NAKH900110 at node two despite its lower ranking seen in Table 7.6. However,

it identifies 9 *S. aureus* peptides based on a lower composition score from Nakashim et al. in [156] to describe membrane proteins. The fourth node, N-terminal feature RICJ880114, appears to correctly classify 55 AMPs better against *E. coli*, but misclassifies 11 AMPs actually better against *S. aureus*. It may be worth considering different features in the future which also describe membrane protein composition to see if they can help lower the number of misclassified AMPs.

7.4 Conclusion and Chapter Summary

In this chapter we have introduced two new contributions to the field. The first, comprises sets of new EFC features specialized for AMPs active against Gram-specific bacteria. Reduced feature sets to separately model GP, GN and GB cases were identified from larger groups of initial features. We show that good performance is maintained even using fewer features in the context of supervised classification using LR, NB, and a number of tree-based methods. The provided decision trees illustrate how a peptide may be classified for each model. It can be observed that in all cases (particularly if the *length* feature is removed for the *GB* case) that EFC features quickly identify a majority of AMPs at the upper levels of the tree. Physicochemical features help discriminate at the lower levels of the tree and tend to classify peptides as AMPs when they are helical and/or flexible. While our analysis in section 7.3.1 focused on understanding the features employed by the decision trees in each setting, in Table 7.7 we show features that seem to be shared amongst the three settings. It is unsurprising that most overlapping features occur with *GB*, as AMPs in this set act on a broader range of bacterial targets.

We also contribute a new predictive "ECvSA model" to discriminate between AMPs more active against *E. coli* or *S. aureus*. A data set is constructed of AMPs with experimental responses available for both taxa. Peptides are then placed into one of two groups based on having better activity against *E. coli* or *S. aureus* based on median MIC values taken from the DBAASP. Using CFS-GS features from the previous chapter with EFC features replaced with the new ones from the the setting described above, FCBF is applied

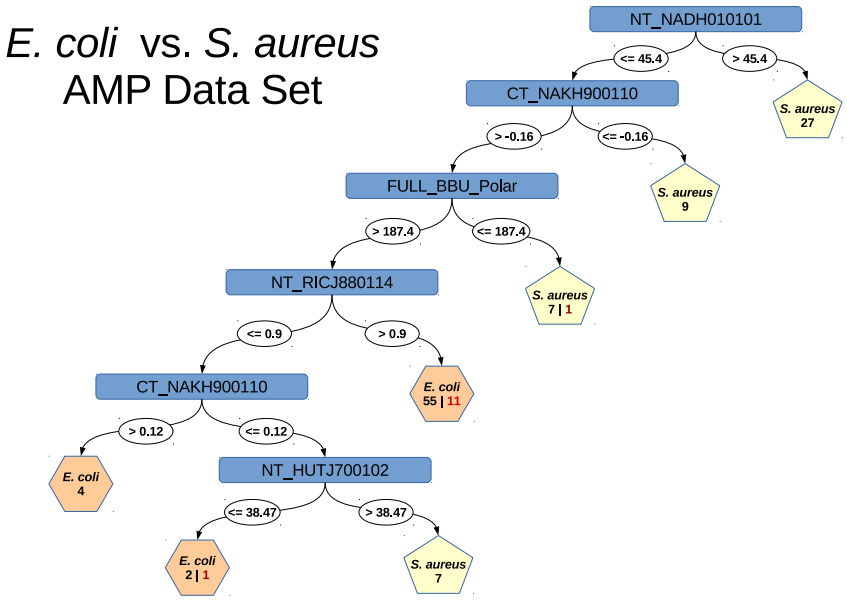

Figure 7.6: Shown is a decision tree based on a reduced set of 11 features generated by FCBF for the ECvSA data set. Feature descriptions are listed in Table 7.6. The tree was generated using the J48 (C4.5) classifier in Weka using 124 AMPs from the DBAASP tested against both *E. coli* and *S. aureus*. 62 observations show better median MIC activity against *E. coli* while another 62 observations show better median MIC activity against *S. aureus*. The figure shows only 6 of the 11 features were used for generating the decision tree. Passing a query peptide through the tree starting at the top will classify it as either better against *E. coli* or *S. aureus*. Nodes in blue represent feature evaluations directing the query to a left or right branch based on a cutoff value (shown in white ovals). Terminating leaves shown as orange hexagons are predicted to be better against *E. coli*, while leaves shown as red pentagons are predicted to be better against *S. aureus*. Numbers in black and red represent the respective number of instances which arrive at a given leaf and receive a correct or incorrect label.

Table 7.7: Mutual features found between the GP, GN and GB models are shown. Features are listed as rows and described in column 1, while the presence or absence in a data set is indicated in columns 2-4. The presence of two • symbols in the same row indicate features between data sets are either identical, or have shifted positions that overlap. A box filled with ∘ indicates an identical motif but with non-overlapping positions. For example, the motif **GGC** in row 6 occurs at overlapping positions for GN and GB but at a non-overlapping position in the GP feature set. Motifs joined by boolean operators in Tables 7.2, 7.3 and 7.4 are considered here as separate features to allow this simplified analysis. EFC features are based on the GBMR4 alphabet introduced in Chapter 5 Table 5.1.

Feature	**GP**	**GN**	**GB**
AAAA	•		•
AGC	•		•
CGA	∘	∘	∘
GCC	•	•	
GGC	∘	•	•
TC		•	•
In vitro peptide aggregation	•		•
AAIndex: KARP850103	•	•	

to select a subset of 11 features. These are used with RF (using 200 trees) to build a model with reasonable classification performance, obtaining an ACC of 83% and MCC of 0.662 in the context of leave-one-out CV. This is an important first step for recognizing AMP-selectivity against certain species of bacteria. To the best of the author's knowledge, it is the first time a ML-based predictive model has been reported in this setting. As more AMPs are deposited in public databases with reported activity against these two taxa, larger sample sizes should make further improvements in recognition possible. So far, the new Gram-specific EFC features from above have not been shown helpful in an ECvSA setting. Provided larger data sets in the future, EFC features can be constructed to specifically discriminate between AMPs showing better activity against any two species of bacteria.

There are many opportunities to further assist wet-laboratory researchers interested in directing the design or modification of novel AMP targets of interest. As more AMPs are characterized and added to databases in the future, we hope that larger sample sizes will aid the computational community in designing more advanced models capable of assisting with

specific bacterial threats. We now proceed to the next chapter where we incorporate models from Chapter 6 and the ECvSA model shown here into a freely-available AMP prediction web server.

Chapter 8: AMP Scanner: A Predictive Web Server

To make our methods reproducible and accessible to the greater AMP research community, we have created "AMP Scanner," a free AMP prediction web server available at `http://www.ampscanner.com`. The server is easy-to-use and simply requires the user to upload a FASTA file of sequences to scan. Two search options are currently available to predict AMPs sequences ≥ 10 AA in length. If the user selects "Classify Sequences Mode" (CSM), up to 2000 peptides can be submitted and classified as AMP or non-AMP based on the entire sequence. By selecting "Proteome Scan Mode" (PSM), a proteome file up to 100MB in size is first scanned using a homology-based search (described below) to identify AMP-like segments to be further classified as AMP or non-AMP. For both modes, results are reported for the top RF and Earth models presented in Chapter 6. If a peptide is classified as a potential AMP by at least one model, it is also checked with the ECvSA model from Chapter 7 to predict if it may work better against the representative GN bacteria *E. coli* or the representative GP bacteria *S. aureus*.

During the screening operation, updates are provided to the user informing them about the current processing stage their query sequences are undergoing. Final results are displayed to the user in a table formatted as in Figure 8.2. The first column contains sequence identifiers which are concatenated in PSM if identical hits are found in multiple sequences. Columns two and three respectively contain prediction results for the RF and Earth classifiers. If PSM mode is used, column four will provide the Earth prediction score (used to order sequences) and the final column contain ECvAS predicted probabilities for *E. coli* vs. *S. aureus* activity. CSM results are sorted alphabetically by identifier while PSM results show the top 100 predicted AMP-like segments when sorted by predicted AMP probability.

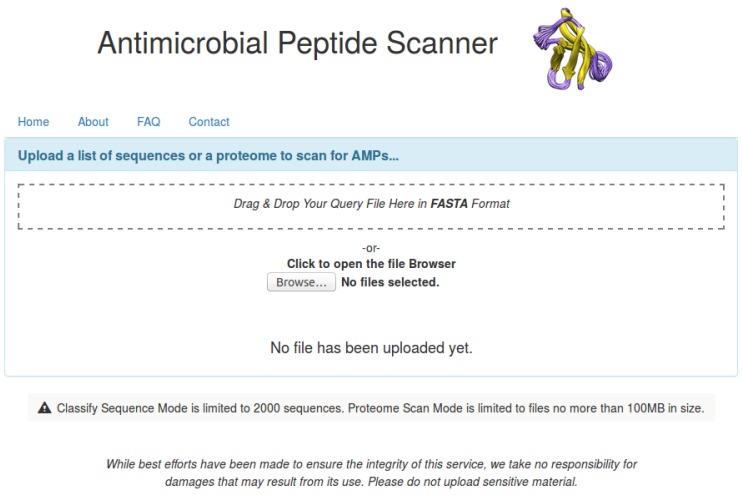

Figure 8.1: A screen shot of the main page for the AMP Scanner prediction server. The user can upload a list of query sequences to check for AMPs with "Classify Sequences Mode" or upload a proteome file to search for potential AMP-like fragments with "Proteome Scan Mode."

We next discuss methods for server implementation including how PSM identifies AMP-like segments. This is followed by results, where CSM is first evaluated using eight experimentally-verified AMPs found recently in *Alligator mississippiensis* [67,68]. PSM is next applied to report the top 10 AMP predictions for the following proteomes: *Alligator mississippiensis*, *Drosophila melanogaster*, *Danio rerio*, *Fusarium graminearum*, *Glycine max*, *Homo sapiens*, *Mus musculus*, *Sus scrofa* and *Xenopus tropicalis*. We conclude the chapter with a brief discussion.

8.1 Methods

Details for how the RF and Earth classifiers are implemented are available in Chapter 6, while the ECvAS predictor is detailed in Chapter 7. We note performance results for RF-based classifiers in previous chapters have been reported using Weka. For logistical reasons, we use `R` and the `randomForest` package Vr.4.6-10 [157] and identical parameter

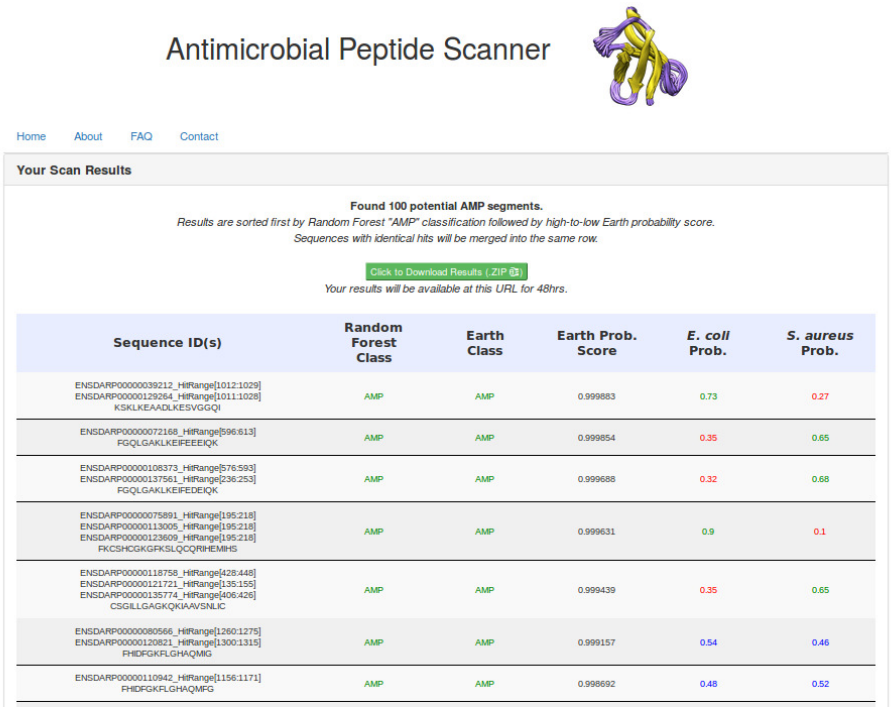

Figure 8.2: Results produced by PSM for the *D. rario* proteome (GenBank: GCA_000002035.3). Using this search mode, up to 100 top potential AMP-like segments are returned by the server and the user has the option to download their results locally. Prediction results are shown for both the RF and Earth models. The probability score used by Earth is also displayed in addition to a prediction for if the sequence might work better against *E. coli* or *S. aureus*. Sequences are ranked higher if both RF and Earth assign an AMP label, with further sorting based on the Earth probability score. Currently, PSM limits the user to uploading files no more than 100MB in size. Results are color coded, with green and red representing AMP and non-AMP classifications for RF and Earth respectively. Predicted probabilities for *E. coli* and *S. aureus* in the same row add up to 1. Probabilities are colored green or red based on higher or lower values. If both categories fall between $0.4 - 0.6$, both values are colored blue to indicate that activity may be similar between taxa.

settings for the server. Testing on a number of data sets show similar performance between implementations. A workflow diagram for the server can be seen in Figure 8.3. We now proceed to describe how PSM detects potential AMP-like segments.

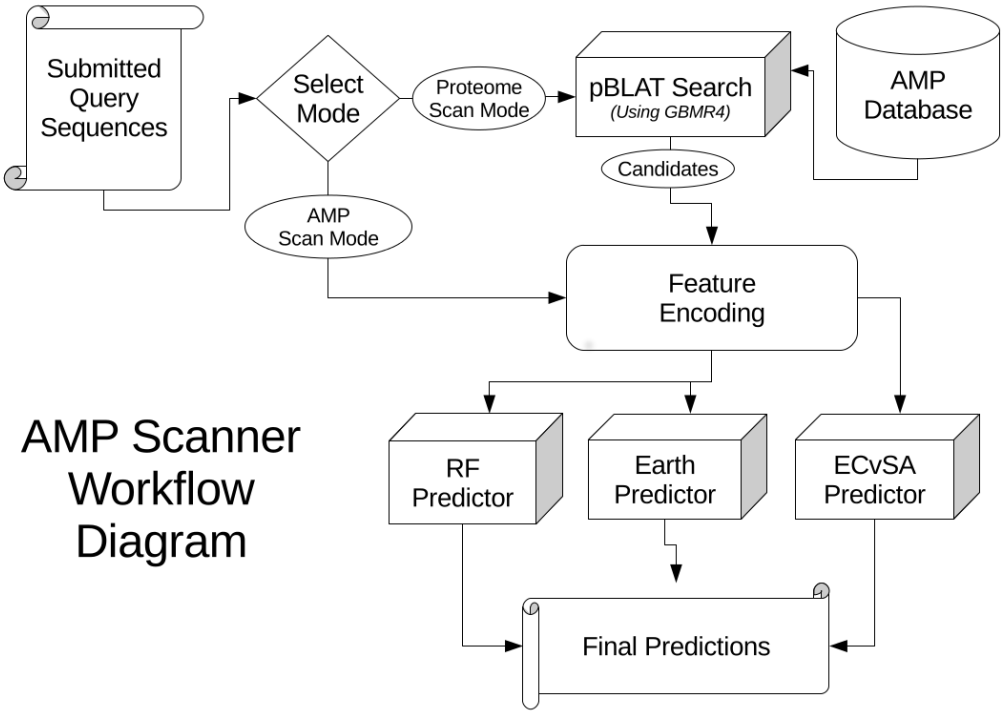

Figure 8.3: A workflow diagram detailing the structure of the AMP Scanner prediction server. The pipeline starts at the top left when the user uploads their query sequences. After selecting the desired mode (CSM or PSM), the pipeline diverges. CSM immediately encodes sequences into feature vectors and submits them to the RF, Earth and ECvSA predictors before reporting results. To generate a diverse range of candidates, PSM first converts the query sequences and a list of known AMPs collected from the APD and CAMP databases into GBMR4 format. These are submitted to pBLAST and hits are collected and merged if tiles overlap. The final list of candidates are converted back to the standard 20 AA alphabet to start the CSM workflow. Sequences identical to already known AMPs are filtered into a separate file and PSM results are sorted and filtered to show the top 100 hits.

8.1.1 Detecting AMP-like Segments with pBLAT

For PSM, we use a homology-based search to identify a list of candidate peptides which are then submitted to our RF and Earth classifiers for further evaluation as shown in Table 8.3. We balance our approach to identify candidates novel enough to be of interest to the research community (i.e. not simply changing a few residues in a known AMP), but similar enough to likely originate from the same natural population from which our training samples were chosen. The method must also be fast as a user may wish to test thousands of large proteins in a reasonable time frame. Keeping these factors in mind, we employ searches with `pBLAT`, a multi-threaded implementation of the `BLAT` program [97] (Kent Informatics; Santa Cruz, CA) written by Wang Meng (`http://icebert.github.io/pblat`). Our testing of the two programs find `BLAT` and `pBLAT` to produce identical results.

Using `pBLAT`, we map the AMPs from the Server Verification Set (detailed in Chapter 3) onto the submitted proteome sequences to find matching hits (which `BLAT` refers to as "tiles"). A custom script in Ruby is used to merge overlapping tiles for the same query and join them together. Rather than simply map known AMPs and adjust parameters to produce near identical matches, we generate a more diverse range of candidates by first changing all query and known AMP sequences into the 4-letter GBMR4 alphabet introduced in Chapter 5. This tends to produce hits sharing between 30-40% sequence identity with known AMPs (based on the standard 20 AA alphabet). However, even non-matching positions may still share similar properties encoded by GBMR4 as detailed in Chapter 5 Table 5.1. Comparing the start and end positions for a GBMR4 hit with the original sequence allows candidates to be converted back to the standard 20 AA alphabet before being checked by classifiers for AMP-activity as in the CSM workflow.

This method generates diverse candidates for consideration with enough speed to process hundreds of thousands of queries in a few minutes. A test using the human proteome GRCh38.p3 (GenBank: GCA_000001405.18) with $101,933$ peptides maps and merges hits to form 462 non-redundant candidates in under a minute using 22 Intel(R) Xeon(R) 2.60GHz CPU cores. As most researchers can only synthesize and test a handful of peptides at a

time, we report the top 100 scoring predictions based on our AMP predictors. If desired, more or fewer candidates can be generated by adjusting the `pBLAT` "`minIdentity`" value or adjusting other parameters.

8.1.2 Implementation Details

All tests reported in this chapter were run on the George Mason University School of Systems Biology BINF server which has 24 Intel(R) Xeon(R) 2.60GHz CPU cores and 64GB of RAM. We limit our pipeline to 22 cores and typically utilize between 600-3000MB of RAM during processing. $100,000$ sequences can be processed in $5-10$ minutes. The server interface makes extensive use of `JQuery` (`http://jquery.com`) and the `Bootstrap` framework (`http://getbootstrap.com`). Back-end scripting of the pipeline is coded with `PHP` (`http://www.php.net/`), `Ruby` (`htt://www.ruby-lang.org`) and `BioRuby` [158]. As mentioned above, sequences are processed with `R` version 3.2.2 and `pBLAT` (using parameters: `-prot`, `-stepSize`=1, and `-minIdentity`=90) which is based on `BLAT` version 35. RF and Earth models are implemented with parameters as described in Chapters 6 and 7.

8.2 Results

We begin with an evaluation of CSM using external AMP data recently verified for *Alligator mississippiensis* as presented in [68]. This is followed by some top AMP predictions made by PSM after submitting nine proteomes from a diverse range of taxa. Genomes are downloaded from Ensembl [159] (`http://ensembl.org`), Broad Institute *Fusarium* Comparative Database (`https://www.broadinstitute.org/annotation/genome/fusarium_graminearum`) and the Crocodilian Genomes Project (`http://crocgenomes.org`). We provide some analysis of top or unusual hits using the APD "Antimicrobial Peptide Calculator and Predictor" (`http://aps.unmc.edu/AP/prediction/prediction_main.php`) to align hits with AMPs in the database (APD accessions are denoted by "AP" followed by a five-digit number). We also use `BLASTp` [160] using NCBI's implementation at: `http://blast.ncbi.nlm.nih.`

gov/Blast.cgi with the NR database.

8.2.1 Performance Comparison on 8 *Alligator mississippiensis* Peptides

Table 8.1 below details how our server using CSM compares to other publicly-available prediction servers introduced in Chapter 6. Eight peptides were recently tested against *E. coli* and *S. aureus* in [68] (see Appendix for sequences). If we consider any AMP activity as a TP and an ineffective response as a TN, both of our RF and Earth methods achieve better performance compared to other servers. As seen in the last two columns in row 10 of Table 8.1, RF and Earth both obtain 75% ACC for *E. coli*. The next row shows Earth to obtain best performance for *S. aureus* with an ACC of 87.5% compared to 62.5% for RF.

Table 8.1: Comparison of AMP prediction servers for 8 *A. mississippiensis* peptides experimentally tested against *E. coli* and *S. aureus* as recently reported in [68]. Columns from left-to-right list: peptide name, experimental response against *E. coli*, experimental response against *S. aureus*, CAMP SVM performance, CAMP RF performance, CAMP DA performance, AntiBP2 performance, APD predictor performance, iAMP-2L performance, our RF model and our Earth model. Rows 10 and 11 list ACC performance for *E. coli* and *S. aureus* for each server when considering any AMP response as a TP and an ineffective response as a TN. Row 12 generalizes the bacterial response levels which are detailed in full in [68]. AMP activity is exhibited in 5 peptides against *E. coli* and 4 against *S. aureus*. Both of our methods perform best for *E. coli* with 75% ACC. Our Earth model performs best for *S. aureus* with 87.5% ACC.

Peptide	***Bacterial Response***		***Prediction Method Response***							
List	*E. coli*	*S. aureus*	CAMP SVM	CAMP RF	CAMP DA	AntiBP2	APD	iAMP-2L	Our RF	Our Earth
APOC1-64:88	+	+/-	-	-	-	-	+	-	+	+
APOC1-67:88	+	+/-	-	-	-	-	+	+	+	+
FGG-398:413	+	+	+	-	-	-	-	-	+	+
FGG-401:413	+	-	+	+	+	NA	-	+	+	-
A1P-394:428	++	+	-	-	-	-	+	-	+	+
AVTG2LP	-	-	+	-	+	+	+	+	+	+
ASAP130LP	-	-	-	-	-	NA	+	+	-	-
NOTS-17:38	-	-	+	+	-	+	+	-	+	-
***E. coli* ACC(%):**			37.5	37.5	37.5	0.0	37.5	37.5	**75.0**	**75.0**
***S. aureus* ACC(%):**			25.0	25.0	12.5	0.0	50.0	25.0	62.5	**87.5**

Bacterial Responses: Great AMP Activity (++), Good AMP Activity (+), Poor AMP Activity (+/-), Ineffective/Non-AMP (-)

ECvSA predicted probabilities for better AMP-activity against *E. coli* (*Ec*) or *S. aureus*

(*Sa*) were as follows: APOC1-64:88 (*Ec*: 0.90, *Sa*: 0.10), APOC1-67:88 (*Ec*: 0.90, *Sa*: 0.10), FGG-398:413 (*Ec*: 0.86, *Sa*: 0.14), FGG-401:413 (*Ec*: 0.68, *Sa*: 0.32), A1P-394:428 (*Ec*: 0.43, *Sa*: 0.57), AVTG130LP (*Ec*: 0.84, *Sa*: 0.16), ASAP130LP (Classified as non-AMP by RF and Earth so not predicted), and NOTS-17:38 (*Ec*:0.14 , *Sa*: 0.86). Predictions are in agreement for APOC1-64:88, APOC1-67:88 and FGG-401:413. Predictions for FGG-398:413 suggest better performance against *E. coli* when equivalent activity is shown by experiment. For A1P-394:428, predictions suggest similar performance (slightly in favor of *S. aureus*) when better activity is reported for *E. coli* by experiment.

8.2.2 Top Predicted Peptides for Nine Proteomes

Alligator mississippiensis

Table 8.2 below shows the top ten AMP-like fragments predicted by PSM for *A. mississippiensis*. The pipeline considered 44,487 proteins based on the American alligator genome assembly allMis0.2 (GenBank: GCA_000281125.1) obtained from the Crocodilian Genomes Project [161]. Potential AMPs shown were predicted as AMPs by both the RF and Earth models and peptides listed range between $16 - 24$ AA in length. Four of the sequences are predicted to do better against *E. coli* compared to six for *S. aureus*. The predicted AMP at rank 1 is suggested to have helix-forming potential by the APD prediction server and best local alignment is reported for the AMP Macropin 2 (AP01957; GTGLPMSERRKIMLMMR) with 39% sequence similarity. Macropin 2 is derived from the bee *Macropis fulvipes* and has been shown effective against Gram-positive and Gram-negative bacteria [162].

Figure 8.4: American alligator photo from U.S. Fish and Wildlife Service.

Table 8.2: Top 10 peptide fragments for *A. mississippiensis* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	TGFFFGLERRQAMAKM	amisp024350 [85:100]	1.000	0.60	0.40
2	ALEKAGGKEFLETVKELRKSQ	amisp016059 [231:251]	1.000	0.20	0.8
3	KKLGGKTSNGWPKKTAVQKETP	amisp014735 [1434:1455]	1.000	0.38	0.62
4	NGSHGLINILKPLRSFFSSYLKSLP	amisp005082 [253:277]	0.999	0.33	0.67
5	HIIVTLSLLSPPKKVLP	amisp020684 [7:23]	0.999	0.66	0.34
6	AGGLVDAEALVALKDL	amisp020398 [344:359]	0.999	0.70	0.30
7	FLNFVSFYGTGTNMTRLNRMS	amisp009641 [54:74]	0.999	0.52	0.48
8	FFVGAGLKTSLLKEQYFFSIKKC	amisp025732 [101:123]	0.998	0.25	0.75
9	TIIGSGIFISPKSVLANVAAV	amisp015708 [46:66]	0.998	0.24	0.76
10	KEEGGISTVFHLHKLFSFG	amisp024429 [1845:1863]	0.998	0.28	0.72

Drosophila melanogaster

Table 8.3 below shows the top ten AMP-like fragments predicted by PSM for *D. melanogaster*. The pipeline considered 30,362 proteins based on the fly genome assembly BDGP6 (GenBank: GCA_000001215.4) obtained from EnsemblMetazoa. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $18 - 25$ AA in length. Six of the sequences are predicted to do better against *E. coli* compared to four for *S. aureus*. The predicted AMP at rank 1 shares 95% sequence identity with Drosocin (AP00172; GKPRPYSPRPTSHPRPIRV) according to the APD peptide predictor and differs only by an additional R residue at the C-terminus. Drosocin, also from fly, is active against Gram-positive and Gram-negative bacteria [163]. The peptide predicted at rank 2 is suggested by the APD as a possible helical AMP with S-S bonds and has two top matches with AMPs sharing 36% sequence identity. One is AP00599 (GIWDTIKSMGKVFAGKILQNL) from *Rana septentrionalis* which is known active against Gram-positive and Gram-negative bacteria [164]. The other is a helical peptide AP00399 (HVDKKVADKVLLLKQLRIMRLLTRL)

Figure 8.5: Fruit fly image from U.S. National Aeronautics and Space Administration (NASA).

from *Pseudacanthotermes spiniger* which is also known effective against Gram-positive and Gram-negative bacteria [165].

Table 8.3: Top 10 peptide fragments for *D. melanogaster* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID[Location]	Earth Prob.	*Ec*	*Sa*
1	GKPRPYSPRPTSHPRPIRVR	FBpp0086566 [21:40]	1.000	0.45	0.55
2	HKKCHDKLLGKCSGSVFTSASTILL	FBpp0289213 [182:206]	1.000	0.82	0.18
3	VKSNCRLKHMRGRSFPPNYV	FBpp0080933 [763:782]	0.999	0.85	0.15
4	ICLTTTALGGKVINIWKY	FBpp0075412 [71:88]	0.999	0.56	0.44
5	FGKFAKYFQQLGIPTGKKLLNL	FBpp0310239 [45:66]	0.999	0.76	0.24
6	ILRKGLQGGVILFRRHCMITL	FBpp0079539 [30:50]	0.998	0.20	0.80
7	VVGLLHGLISRFVFTQKRTKK	FBpp0111644 [56:76]	0.998	0.12	0.88
8	IPKANILSLFAKSALPGGKYKNL	FBpp0083144 [49:71]	0.998	0.56	0.44
9	KIGPKSVLLCSGLLQISGWAC	FBpp0077271 [81:101]	0.997	0.72	0.28
10	GLVRKFAGFTTGHVSTPKKKIR	FBpp0079696 [24:44]	0.997	0.10	0.90

Danio rerio

Table 8.4 below shows the top ten AMP-like fragments predicted by PSM for *D. rerio*. The pipeline considered 44,487 proteins based on the zebrafish genome assembly GRCz10 (GenBank: GCA_000002035.3) obtained from Ensembl. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $16-24$ AA in length. Half of the sequences are predicted to do better against *E. coli* compared to *S. aureus* and vice versa. The predicted AMP at rank 1 is suggested to have helix-forming potential by the APD prediction server and best local alignment with sequence AP02392 (VKLEILGSKGGAKI) is reported having 42% sequence similarity. AP02392 has been shown effective against *E. coli* and *M. tetragenus* [166].

Figure 8.6: Zebrafish photo from U.S. National Institute of Health (NIH).

Table 8.4: Top 10 peptide fragments for *D. rerio* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	KSKLKEAADLKESVGGQI	ENSDARP00000039212 [1012:1029]	1.000	0.73	0.27
2	FGQLGAKLKEIFEEEIQK	ENSDARP00000072168 [596:613]	1.000	0.35	0.65
3	FGQLGAKLKEIFEDEIQK	ENSDARP00000108373 [576:593]	1.000	0.32	0.68
4	FKCSHCGKGFKSLQCQRIHEMIHS	ENSDARP00000075891 [195:218]	1.000	0.90	0.10
5	CSGILLGAGKQKIAAVSNLIC	ENSDARP00000118758 [428:448]	0.999	0.35	0.65
6	FHIDFGKFLGHAQMIG	ENSDARP00000080566 [1260:1275]	0.999	0.54	0.46
7	FHIDFGKFLGHAQMFG	ENSDARP00000110942 [1156:1171]	0.999	0.48	0.52
8	KRLKERRRSSVVVSLPGLDVS	ENSDARP00000106435 [260:280]	0.999	0.82	0.18
9	LRRTGRLFGGVIRDVRRR	ENSDARP00000041318 [675:692]	0.999	0.44	0.56
10	AAKNILHAKGGSLLAAYIKVLP	ENSDARP00000048217 [297:318]	0.998	0.69	0.31

Fusarium graminearum

Table 8.5 below shows the top ten AMP-like fragments predicted by PSM for *F. graminearum*. The pipeline considered $13,321$ proteins based on the *F. graminearum* PH-1 (FG3) assembly obtained from the Broad Institute *Fusarium Comparative Database* (`https://www.broadinstitute.org/annotation/genome/fusarium_graminearum`). Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $16-22$ AA in length. Six of the sequences are predicted to do better against *E. coli* compared to four for *S. aureus*. The predicted AMP at rank 1 is suggested to have helix-forming potential by the APD prediction server and shares 46% sequence similarity with AP00885 (FLPILASLAAKLGPKLFCLVTKKC). AP00885 is a brevinin AMP with activity against Gram-positive and Gram-negative bacteria. It also has been shown to have antifungal activity against *Candida* spp. [167].

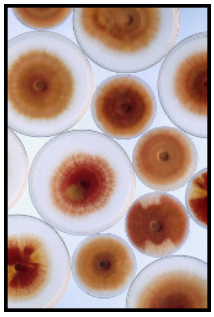

Figure 8.7: Photo of *Fusarium graminearum* cultures from USDA.

Table 8.5: Top 10 peptide fragments for *F. graminearum* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	FIPSFSLNIKLPKFTFVLKYT	FGSG_12311T0 [94:114]	1.000	0.60	0.40
2	IRKVEEHCIYVGRVGF	FGSG_11266T0 [638:653]	1.000	0.81	0.19
3	FDVSIGEIFGTLVAGGCL	FGSG_10702T0 [434:451]	1.000	0.25	0.75
4	SKLTVISSYLPSFFRIITSSRL	FGSG_03683T0 [143:164]	1.000	0.78	0.22
5	VRKVAKLGALDGRKPDVETEV	FGSG_04378T0 [1182:1202]	1.000	0.47	0.53
6	FFSIFYKMKNNLPRVLLVLLL	FGSG_12791T0 [66:86]	1.000	0.59	0.41
7	LCFFVGGLKQASQKFHAIVSEV	FGSG_01802T0 [180:201]	0.999	0.74	0.26
8	GGWFCPFVQRSWITIHEKR	FGSG_04212T0 [43:61]	0.998	0.48	0.52
9	SLSFARFPTLLCVAFRRFN	FGSG_08544T0 [907:925]	0.997	0.25	0.75
10	SVGCVAAFANVFPTLTSKIYAL	FGSG_03069T0 [226:247]	0.995	0.63	0.37

Glycine max

Table 8.6 below shows the top ten AMP-like fragments predicted by PSM for *G. max*. The pipeline considered 73,319 proteins based on the soybean genome assembly V1.0 (GenBank: GCA_000004514.1) obtained from EnsemblPlants. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $18-22$ AA in length. Six of the sequences are predicted to do better against *E. coli* compared to four for *S. aureus*. The predicted AMP at rank 1 is suggested to have helix-forming potential by the APD prediction server and a top alignment with 36% shared sequence identity is reported for PR-bombesin (AP01233; QKKPPRPPQWAVGHFM). PR-bombesin is listed in the APD as having activity against Gram-positive and Gram-negative bacteria in addition to antifungal activity [168].

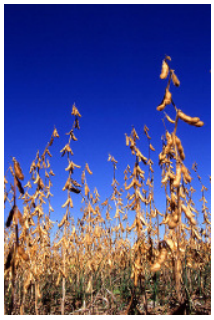

Figure 8.8: Soybean photo from USDA.

Table 8.6: Top 10 peptide fragments for *G. max* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	IKKEFDKRHGPTWHCIVGRNFV	GLYMA15G16060.2 [70:91]	1.000	0.70	0.30
2	AFGGVLLELLSGNVKEKL	GLYMA15G28000.1 [333:350]	0.999	0.07	0.93
3	HGLHPFKFKSLVDFAAFQ	GLYMA12G35100.1 [45:62]	0.999	0.60	0.40
4	RFRNQPLPLKVFTRVAVSTV	GLYMA16G08170.2 [23:42]	0.998	0.34	0.66
5	KRARKSIKGVAPVKRLKK	GLYMA17G06340.1 [430:447]	0.998	0.73	0.27
6	VDFAFGFFLAKNIPGEKKKMI	GLYMA09G32140.1 [1114:1134]	0.998	0.04	0.96
7	IKKEFDKRHGPTWHCIVGRNFG	GLYMA09G04820.1 [70:91]	0.998	0.70	0.30
8	HGLHPFKFKSLVAFAAFQ	GLYMA13G35460.1 [45:62]	0.998	0.60	0.40
9	ASFKVEAKKGEWLPGLAS	GLYMA04G33360.1 [44:61]	0.998	0.57	0.43
10	GVSFFLGLSIPAYFQQYKPQT	GLYMA18G29440.1 [648:668]	0.997	0.03	0.97

Homo sapiens

Table 8.7 below shows the top ten AMP-like fragments predicted by PSM for *Homo sapiens*. The pipeline considered 101,933 proteins based on the human genome assembly GRCh38.p3 (GenBank: GCA_000001405.18) and taken from Ensembl. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $18-24$ AA in length. Half of the sequences are predicted to do better against *E. coli* compared to *S. aureus* and vice versa. An unusual looking sequence can be seen at rank 3 which is H-rich and shares 83% sequence identity when aligned with AP01494 (GHHPHGHHPHGHHPHGHHHPH) using the APD prediction server.

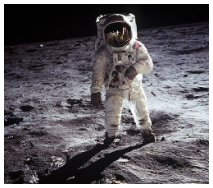

Figure 8.9: Photo of Buzz Aldrin from NASA.

AP01494 is listed as a histidine-rich antifungal protein with activity against *Candida parapsilosis* and *Candida albicans* [169]. A search with `BLASTp` at NCBI (default settings) returns a hit for a human histidine-rich glycoprotein (GI:51476334) as one of the top hits. Another unusual peptide is shown at rank 7 which contains three N-terminal C residues. A viable AMP might require some of these leading C residues to be trimmed as no sequences starting

with more than one sequential C residue at the N-termini are currently listed in the APD. The APD server prediction program suggests the protein has a propensity to form a helix and reports a top alignment with AP00828 (GMFSVLKNLGKVGLGFVACKINKQC) sharing 41% sequence identity. NCBI BLASTp reveals a top hit (GI:767943020) for a predicted hexaprenyldihydroxybenzoate methyltransferase enzyme involved with ubiquinone biosynthesis [170].

Table 8.7: Top 10 peptide fragments for *H. sapiens* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	PPKCPPKCTPKCPPKCPPKCPP	ENSP00000357772 [12:33]	1.000	0.44	0.56
2	PPKCPPKCTPKCPPKCPPKCL	ENSP00000357769 [12:32]	1.000	0.44	0.56
3	GHHPHGHHPHGHHPHGHHPHGHHP	ENSP00000232003 [382:405]	1.000	0.68	0.32
4	ATSGKKGGKKSKAAKPRTSKKS	ENSP00000362352 [138:159]	1.000	0.34	0.66
5	LKPRNHFGVGRSTVTTKVTL	ENSP00000358062 [213:232]	0.999	0.47	0.53
6	LNFQTKGYNKVSPFFVPKI	ENSP00000280701 [164:182]	0.998	0.87	0.13
7	CCCQVLKPGGSLFITTINKTQL	ENSP00000254759 [236:257]	0.998	0.92	0.08
8	PRFGFFTSDFKAKSSIQF	ENSP00000257408 [939:956]	0.995	0.23	0.77
9	SVRISFAKGWGPCYSRQFITSCP	ENSP00000288840 [462:484]	0.992	0.75	0.25
10	WKRPRLTHNGPVRRSTVIDQI	ENSP00000306682 [245:265]	0.989	0.56	0.44

Mus musculus

Table 8.8 below shows the top ten AMP-like fragments predicted by PSM for *M. musculus*. The pipeline considered 55,664 proteins based on the mouse genome assembly GRCm38.p4 (GenBank: GCA_000001635.6) obtained from Ensembl. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $16-30$ AA in length. Six of the sequences

Figure 8.10: Common house mouse photo from NIH.

are predicted to do better against *E. coli* compared to four for *S. aureus*. The sequence at rank 1 with high H-content appears to be an extended version of the predicted human AMP at rank 3. Similar to above, the APD sequence predictor also selects peptide AP01494 (GHHPHGHHPHGHHPHGHHHPH) as having a best alignment with a shared sequence identity of 70%. The peptide at rank 2 is marked by the APD predictor as potentially helical and reports a top alignment with Styelin A (AP00328; GFGKAFHSVSNFAKKHKTA) sharing 41% sequence identity. Styelin A is a helical peptide from *Styela clava* and has been shown to have a broad range of activity against Gram-positive and Gram-negative bacteria [171].

Table 8.8: Top 10 peptide fragments for *M. musculus* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	GHHPHGHHPHGHHPHGHHPHGHHPHGHDFL	ENSMUSP00000023590 [390:419]	1.000	0.62	0.38
2	CGKGFISFAQLTVHIKTH	ENSMUSP00000124075 [712:729]	1.000	0.78	0.22
3	PVYGRRKKCHISRFNLFQVFP	ENSMUSP00000041436 [116:136]	1.000	0.78	0.22
4	IKVLRRVKEENDRRGGFIRIF	ENSMUSP00000039939 [465:485]	1.000	0.41	0.59
5	VILGFIHGAIQTLFMAQL	ENSMUSP00000107132 [143:160]	1.000	0.07	0.93
6	AGGLVDAEALVALKDL	ENSMUSP00000027111 [341:356]	1.000	0.70	0.30
7	FHIDFGKFLGHAQMFG	ENSMUSP00000126092 [1264:1279]	1.000	0.48	0.52
8	DKKRKLKRRESLQDQRSRIKGPF	ENSMUSP00000043570 [950:972]	0.998	0.76	0.24
9	RTSALLGFCGFMFRQTNII	ENSMUSP00000097882 [172:190]	0.998	0.54	0.46
10	QKGHGIVFSEGERWKLLRRFSL	ENSMUSP00000026211 [110:131]	0.988	0.35	0.65

Sus scrofa

Table 8.9 below shows the top ten AMP-like fragments predicted by PSM for *S. scrofa*. The pipeline considered 47,259 proteins based on the pig genome Sscrofa10.2 (GenBank: GCF_000003025.5) obtained from Ensembl. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $16-25$ AA in length. Half of the sequences are predicted to do better against *E. coli* compared to *S. aureus* and vice versa. The peptide at rank 1 is identified by the APD predictor to be rich in R residues and shares a sequence identity of 42.1% to the known AMP scolopendin 2 (AP02447; AGLQFPVGRIGRLLRK). The later peptide has been previously identified in the Chinese red-headed centipede and has been shown active against *S. aureus*, *E. coli*, multiple antibiotic-resistant strains of *P. aeruginosa*, and the fungi *C. albicans* [172]. NCBI `BLASTp` returns numerous hits for a conserved "histone H2A" domain. Scolopendin 2 is also identified by the APD predictor as the top match for the sequence at rank 2, with 42.3% shared sequence identity. The APD predictor suggests the sequence may be helical with membrane-binding potential and NCBI `BLASTp` returns numerous hits for coronin-1B peptides from a variety of eukaryotic organisms.

Figure 8.11: Pig photo from USDA.

Xenopus (Silurana) tropicalis

Table 8.10 below shows the top ten AMP-like fragments predicted by PSM for *X. tropicalis*. The pipeline considered 22,718 proteins based on the Western clawed frog genome assembly JGI 4.2 (GenBank: GCA_000004195.1) obtained from Ensembl. Potential AMPs shown were classified as AMPs by both the RF and Earth models and peptides listed range between $16-22$ AA in length. All but two of the ten sequences are predicted to do better against *E. coli* compared to

Figure 8.12: Western clawed frog photo from Narbonne et al. [173] under a CC-BY license.

Table 8.9: Top 10 peptide fragments for *S. scrofa* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	RGELRGRLARLELERA	ENSSSCP00000013869 [1175:1190]	1.000	0.42	0.58
2	LRALRALVKEQGERIGRLEEQLGRM	ENSSSCP00000013737 [468:492]	1.000	0.60	0.40
3	CFRHHRGLGSFTVSCWASL	ENSSSCP00000025676 [138:156]	1.000	0.85	0.15
4	RGGRLCYCRRRFCVCVGRG	ENSSSCP00000027630 [130:148]	1.000	0.68	0.32
5	LLIAGVGTFAILRKRKKK	ENSSSCP00000004594 [441:458]	1.000	0.08	0.92
6	ATSGKKGGKKSKAAKPRSSKKS	ENSSSCP00000026524 [116:137]	1.000	0.34	0.66
7	SRVACGVHGFPRPCAWRRISS	ENSSSCP00000018254 [1182:1202]	1.000	0.74	0.26
8	KGLGKKKEKLGEVVKKEE	ENSSSCP00000006887 [371:388]	1.000	0.52	0.48
9	IKVLRRVKEENDRRGGFIRIF	ENSSSCP00000002584 [27:47]	0.999	0.41	0.59
10	DAAVRFLEGRGVKIARALV	ENSSSCP00000013919 [281:299]	0.999	0.20	0.80

S. aureus. The sequence at rank 1 is reported as potentially helical by the APD sequence predictor and matches it to the AMP PGLa-St2 (AP02295; EAEKAKIAKAEQAEGKGANAA) with 42% shared sequence identity. PGLa-St2 is an AMP already known in *X. tropicalis* with activity reported against *E. coli*, *P. aeruginosa*, *M. luteus*, and *S. cerevisiae*. It also has activity reported against the parasite *T. brucei* which causes African trypanosomiasis [174]. NCBI `BLASTp` returns hits for "glycine receptor subunit beta precursors." An unusual sequence at rank 10 appears to be enriched with K residues. The APD peptide prediction server suggests the sequence may form a helix and a top alignment shows 41% sequence identity shared with a synthetic AMP (AP01158; ALYKKFKKKLLK-SLKRL) with known activity against *E. coli* [175]. A `BLASTp` search with NCBI reveals a number of RING finger (a zinc finger domain) proteins involved with ubiquitination [176]. If this peptide is an AMP, similarity to a zinc finger motif would suggest it could bind to proteins, lipids or work via DNA and/or RNA interference [177–179].

Table 8.10: Top 10 peptide fragments for *X. tropicalis* predicted as an AMP by both RF and Earth models. The first four columns from left-to-right indicate: rank, the predicted AMP sequence, a protein ID where the fragment was found (first and last AA indicated in brackets), and the Earth probability score used for ranking. The last two columns are ECvSA model probabilities for the predicted AMP sequence working better against *E. coli* (*Ec*) or *S. aureus* (*Sa*), respectively.

Rank	Sequence	Protein ID [Location]	Earth Prob.	*Ec*	*Sa*
1	EAEKAKIAKAEQAEGKGANAA	ENSXETP00000042278 [360:380]	1.000	0.62	0.38
2	ERRELEKKYKIKGGVK	ENSXETP00000044751 [562:577]	1.000	0.58	0.42
3	KVLKAIAEALEEHGAGAG	ENSXETP00000047667 [227:244]	1.000	0.85	0.15
4	ALEKKGGKEFVEAVTELKKKN	ENSXETP00000017330 [228:248]	1.000	0.46	0.55
5	LFFEHGNKCGKLLASALKKKQ	ENSXETP00000059931 [367:387]	1.000	0.88	0.12
6	RLLERREKQKMLTAAGMPLGI	ENSXETP00000050631 [158:178]	1.000	0.78	0.22
7	RRRCLRNLRKGDDCGM	ENSXETP00000047465 [214:229]	1.000	0.73	0.27
8	KGKFGKFKRYVAHSSHVTNLR	ENSXETP00000033211 [1228:1248]	1.000	0.84	0.16
9	GFIGLTSFSVHVLCKTNLF	ENSXETP00000002062 [463:481]	1.000	0.49	0.51
10	AVEKGGKKKKKQQKLLFSTSIV	ENSXETP00000059201 [732:753]	0.999	0.68	0.32

8.3 Conclusion and Chapter Summary

This chapter has presented AMP Scanner, a useful resource for the AMP research community for screening sequences and proteomes for potential new AMP sequences. To the best of our knowledge, it is the first prediction server to make predictions for better activity against *E. coli* and *S. aureus*. The server can handle user-requests with hundreds of thousands of peptides and report results in a reasonable time frame (typically under 10 minutes).

Early promising results can be seen from a small trial using eight experimentally-verified peptides from the American alligator. Both the RF and Earth classifiers from Chapter 6 obtain ACCs of 75.0% for responses against *E. coli*, which, is twice as high as for other methods reported in Table 8.1. The Earth model obtains an ACC of 87.5% for responses against *S. aureus*, which, is 12.5% higher than the RF model and 37.5% better than the next-best-performing APD prediction server. Predictions for better activity against *E. coli* and *S. aureus* agree for three of the five peptides with reported AMP activity. While these results are encouraging, such a small sample size requires caution and the need for additional testing with larger sets of experimentally validated peptides. Hopefully, improvements in

technology will soon make larger-scale AMP testing feasible so that better validation sets generated under the same conditions will become available. Of particular importance, is the need for a large list of confirmed non-AMPs so that the computational community no longer has to rely on negative training sets based on homology or database keyword searches. In the meantime, the server will be continuously tested and improved using the small batches of new AMP results released each year in the literature.

Another contribution from this chapter are the lists of potential AMP-like sequences needing verification which are generated by PSM on a variety of organisms. Based on some of the preliminary analysis presented above, many of these sequences appear promising on their own. Others, such as the human peptide presented at rank 7 in Table 8.7, might require additional modification to be viable based on the expertise and needs of individual researchers. Results from any experiments on these peptides, particularly any incorrect predictions, could help in making improvements to our predictive models.

We hope AMP Scanner will be a helpful tool for hypothesis generation in AMP research. The server is currently available at: `http://www.ampscanner.com`.

Chapter 9: Discussion and Future Directions

This dissertation has presented a number of contributions to the field of computational AMP recognition as outlined in the introduction. Our first contribution is the consideration of feature interactions when building predictive models to classify a sequence as an AMP or non-AMP. In Chapter 4, we start with eight initial physicochemical features used in other AMP recognition work [46,47] and demonstrate superior classification performance using fewer features when interaction terms are included. In Chapter 5, we introduce an evolutionary framework to construct distal "EFC" features which recognize complex patterns contained within an AMP sequence. These novel features go beyond simple sequence composition to build motifs which capture local and non-local relationships between a variable number of AA positions. Furthermore, correlations between motifs are also considered for cases where they are co-expressed or mutually exclusive. In Chapter 6 we contribute two new predictive models which pair a carefully-selected subset of features with RF and multivariate adaptive regression splines. In addition to the above interaction terms and EFC features, we also consider physicochemical features structured about important AMP functional regions. These features provide additional descriptive power to our models as they not only describe which physicochemical feature are important but also in which region they apply. Based on average performance between two publicly available testing sets of AMPs, we show our models outperform other publicly-available AMP prediction servers. In particular, RF beats the next-best iAMP-2L server by 2.4% ACC and 0.05 MCC. In Chapter 7 we examine AMP recognition in a new experimental setting for computational AMP research by considering AMP-selectivity. We produce novel EFC features which embody Gram-specific AMP sequences. For the first time, we also construct model based on RF to predict if an AMP may perform better against the GN bacteria *E. coli* or the GP bacteria *S. aureus* with 83% ACC and 0.66 MCC. Both of these bacteria are important

research models with clinical importance [6, 180–183]. Finally, in accordance with Chou's 5-step rule [60], outlined in the introduction and followed throughout this work, we make the above contributions available online through the web server "AMP Scanner" as detailed in Chapter 8.

AMP Scanner is a free service, available at: `http://www.ampscanner.com`, which makes the methods in this dissertation reproducible and accessible to the greater research community. Two prediction pipelines are offered- one to classify a list of query sequences, and another to screen entire proteomes for potential AMP-like fragments. Using a small validation set of recently discovered peptides in the American alligator, initial results look promising as our method outperforms other servers in correctly discriminating active from inactive sequences. We also correctly predict better activity between *E. coli* and *S. aureus* in three of the five active peptides. However, until many more AMPs predicted by the server can be experimentally validated, researches should be cautious when interpreting AMP Scanner results. We envision this server as helping wet lab researchers with hypothesis generation and cost-reduction; whittling large lists of potential queries down to a manageable number of peptides which have a higher probability of AMP activity.

Aside from improvements to AMP Scanner based on experimental feedback for correct and incorrect predictions, there are a number of exciting new lines of research to pursue. One possibility for future work is the construction of an ensemble learner [184] using features specific to AMP-classes (e.g. temporins, bacteriocins, etc.). While this has been pursued to a limited extent for cathelicidins [52, 88], Figure 9.1 shows an example of how our new EFC features are able to capture class-specific motifs amongst different AMPs. Numerous class-specific predictors could be combined using a boosting technique [185, 186] like AdaBoost [187]. Essentially, weak learners specialized for each AMP class (and trained on the mistakes of complimentary-learners) may perform better collectively so long as each can recognize AMPs with $\geq 50\%$ ACC [188]. As some AMP classes are much better populated in AMP databases than others, a technique such as cascading, where weaker learners are only used when stronger learns are uncertain, may also be worth considering [189].

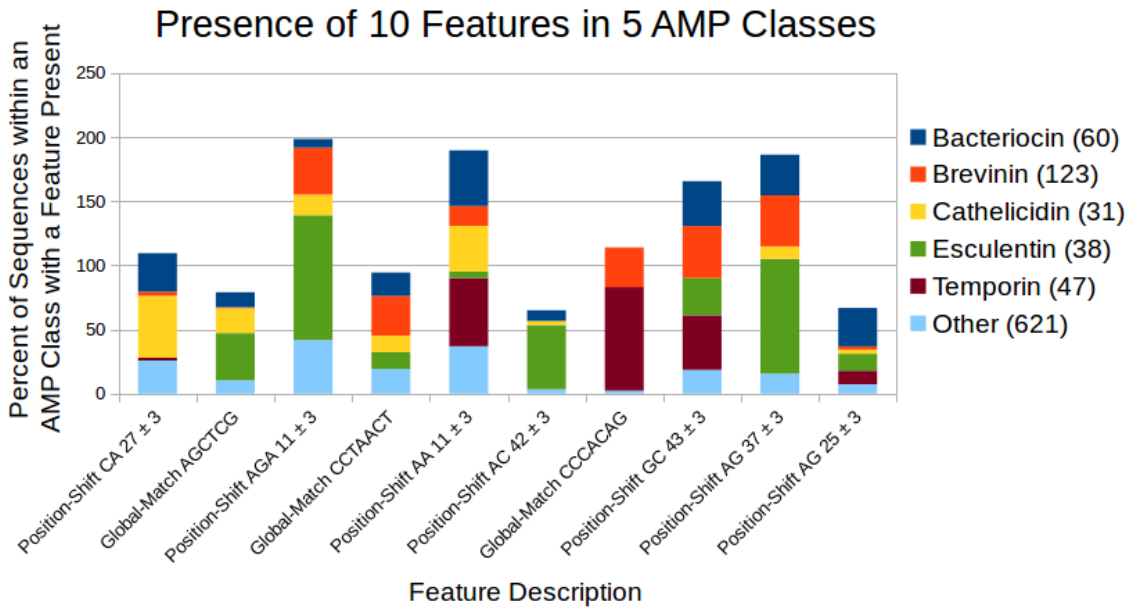

Figure 9.1: This stacked bar graph shows the percentage/proportion of sequences within an AMP class which contain selected EFC features from Chapter 5. We consider the 5 most-populous AMP classes in the Xiao testing data set and group the rest as "Other." The legend shows the assigned color and total number of observations for each group in parenthesis. 10 different EFC features are shown on the x axis using the GBMR4 alphabet detailed in Chapter 2 Table 5.1. The cumulative occurrence of observations containing an EFC feature is shown on the y axis. For each feature (stacked bar), no individual class/color may exceed 100%, but groups can do so collectively. For example, if a feature was found in all observations for all 6 groups, the bar would extend to 600 ($6 \times 100 = 600$). We can see certain features are more broadly employed than others. Feature Position-Shift AA 11 ± 3 (5th bar from the left) is found in all 6 classes. On the other hand, feature Global-Match CCCACAG (7th bar from left) is almost exclusively found in temporins or brevenins, with a few matches for "other."

Another research direction to consider are better predictors for AMP-selectivity using data available in the DBAASP. Going beyond individual bacteria such as *E. coli* and *S. aureus* to form groups based on similar lipid profiles of bacterial membranes may help deal with the shortage of overlapping observations mentioned in Chapter 7.

Finally, a projection of AMPs and non-AMPs from the Xiao training data set seen in Figure 9.2 suggests that there is still room for improving our CFS-GS feature set introduced in Chapter 6. Of particular interest is finding features which better quantify peptide-lipid interaction as this is critical for AMP function [57, 58]. Monte Carlo simulations have been used to model peptide-lipid interactions based on experimental data [190, 191]. Recently, these simulations have been adapted for modeling AMP-membrane interactions and made accessible to the research community through the MCPep server [192] (`http://bental.tau.ac.il/MCPep`). While our testing with MCPep has shown results to vary greatly depending on the AMP submitted, it takes a different and refreshing approach to studying AMPs *in silico*. Brief, coarse-grained molecular dynamics simulations to assess transient binding preferences between AMPs and a small number of lipids of various types may be an avenue to consider until computing costs and power become amenable to handling larger membrane systems in a high-throughput manner.

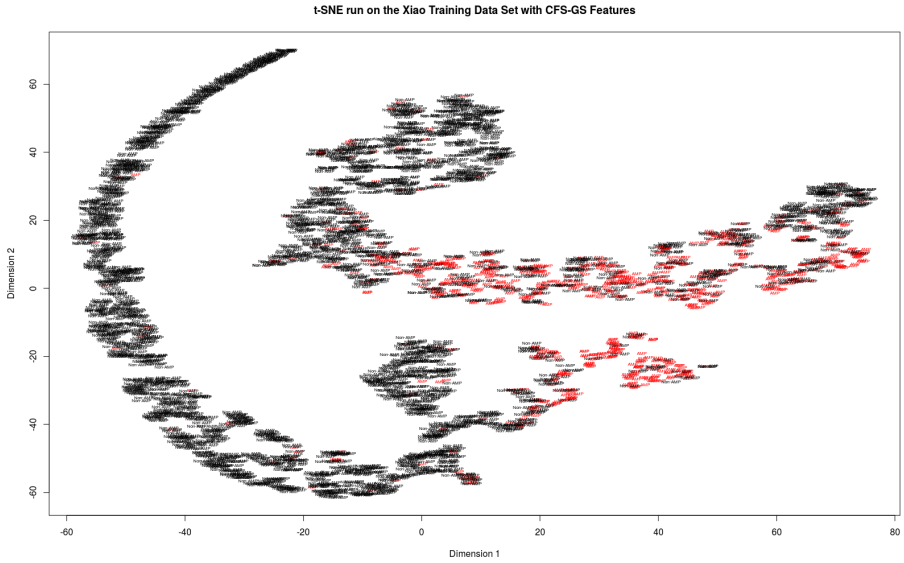

Figure 9.2: Shown is a two-dimensional projection of the Xiao training data set encoded with the 74 CFS-GS features from Chapter 6. The figure was created using the t-distributed stochastic neighbor embedding (t-SNE) non-linear feature reduction algorithm [193] from the `Rtsne` package in R. While the package is designed for using the Barnes-Hut SNE [194] implementation, the theta parameter was set to 0 to run an exact t-SNE calculation. AMPs are shown in red, while non-AMPs are shown in black. A majority of AMPs can be seen to form a large red cluster near the center of the figure. A number of non-AMPs appear to fall into this area as well. The large arc formed by non-AMPs on the left side of the figure can also be seen to contain a few isolated AMPs in red. Investigating why these isolated AMPs appear as outliers compared to the rest may reveal interesting new findings about AMP structure and function. Furthermore, the lack of clear separation between the two classes in the figure suggests improvements can still be made to our feature set.

Appendices

Chapter A: Data Sets

We list here the protein sequences used for each of the data sets listed in Chapter 3.

Table A.1: Fernandes et al. (2012) Data Set AMP Sequences

Definition	Sequence
1AFP:A_PDBID	ATYNGKCYKKDNICKYKAQSGKTAICKCYVKKCPRDGAKCEFDSYKGKCYC
1AYJ:A_PDBID	EKLCERPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
1BNB:A_PDBID	APLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
1BRZ:A_PDBID	EDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
1C01:A_PDBID	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPF GWKSIFIQC
1CIX:A_PDBID	YSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
1CW5:A_PDBID	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGAGSIGRRP
1CZ6:A_PDBID	RSVCRQIKICRRRGGCYYKCTNRPY
1D6X:A_PDBID	VRRFPWWWPFLRR
1D7N:A_PDBID	INLKALAALAKKIL
1DKC:A_PDBID	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
1DQC:A_PDBID	YLAFRCGRYSPCLDDGPNVNLYSCCSFYNCHKCLARLENCPKGLHYNAYLKVCDWPSKAGCTSVNKE CHLWKT
1E4Q:A_PDBID	PVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
1E4S:A_PDBID	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK
1E4T:A_PDBID	NSKRACYREGGECLQRCIGLFHKIGTCNFRFKCCKFQ
1E68:A_PDBID	MAKEFGIPAAVAGTVLNVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVI AW
1ED0:A_PDBID	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK
1EWS:A_PDBID	MPCSCKKYCDPWEVIDGSCGLFNSKYICCREK
1FRY:A_PDBID	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG
1G6E:A_PDBID	MINRTDCNENSYLEIHNNEGRDTLCFANAGTMPVAIYGVNWVESGNNVVTLQFQRNLSDPRLETITLQ KWGSWNPGHIHEILSIRIY
1G89:A_PDBID	ILPWKWPWWPWRR
1HVZ:A_PDBID	GFCRCLCRRGVCRCICTR
1I2U:A_PDBID	DKLIGSCVWGAVNYTSDCNGECKRRGYKGGHCGSFANVNCWCET
1ICA:A_PDBID	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
1IYC:A_PDBID	ELPKLPDDKVLIRSRSNCPKGKVWNGFDCKSPFAFS
1K48:A_PDBID	NGLPVCGETCVGGTCNTPGCTCSWPVCTR
1KFP:A_PDBID	ECRRLCYKQRCVTYCRGR
1KJ6:A_PDBID	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
1KV4:A_PDBID	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKA
1L9L:A_PDBID	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRVCRTGRSRWRDVCRNFMRRYQSRVIQGLVAGETA QQICEDLR
1LFC:A_PDBID	FKCRRWQWRMKKLGAPSITCVRRAF
1M02:A_PDBID	HPLKQYWWRPSI
1MAG:A_PDBID	VGALAVVVWLWLWLW
1MM0:A_PDBID	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
1MMC:A_PDBID	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
1MQZ:A_PDBID	CTFTLPGGGGVCTLTSECI

Table A.1: Fernandes et al. (2012) Data Set AMP Sequences Continued...

Definition	Sequence
1MR4:A_PDBID	RECKTESNTFPGICITKPPCRKACISEKFTDGHCSKILRRCLCTKPC
1MYN:A_PDBID	DCLSGRYKGPCAVWDNETCRRVCKEEGRSSGHCSPSLKCWCEGC
1NKL:A_PDBID	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAI CVDIKICKE
1OF9:A_PDBID	GEILCNLCTGLINTLENLLTTKGADKVKDYISSLCNKASGFIATLCTKVLDFGIDKLIQLIEDKVDANAIC AKIHAC
1OG7:A_PDBID	KYYGNGVHCGKHSCTVDWGTAIGNIGNNAAANWATGGNAGWNK
1P9Z:A_PDBID	ETCASRCPRPCNAGLCCSIYGYCGSGAAYCGAGNCRCQCRG
1PG1:A_PDBID	RGGRLCYCRRRFCVCVGR
1PXQ:A_PDBID	NKGCATCSIGAACLVDGPIPDFEIAGATGLGLWG
1Q3J:A_PDBID	CIKNGNGCQPNGSQGNCCSGYCHKQPGWVAGYCRRK
1Q71:A_PDBID	GGAGHVPEYFVGIGTPISFYG
1R1F:A_PDBID	TFCGETCRVIPVCTYSAALGCTCDDRSDGLCKRNGDP
1RKK:A_PDBID	RRWCFRVCYRGFCYRKCR
1RPB:A_PDBID	CLGIGSCNDFAGCGYAVVCFW
1S6W:A_PDBID	GCRFCCNCCPNMSGCGVCCRF
1T51:A_PDBID	ILGKIWEGIKSLF
1TI5:A_PDBID	RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC
1UT3:A_PDBID	SFGLCRLRRGFCARGRCRFPSIPIGRCSRFVQCCRRVW
1VM5:A_PDBID	GLFDIIKKIAESF
1XC0:A_PDBID	GFFALIPKIISSPLFKTLLSAVGSALSSSGGQE
1XKM:A_PDBID	ENREVPPGFTALIKTLRKCKII
1XKM:B_PDBID	NLVSGLIEARKYLEQLHRKLKNCKV
1XV3:A_PDBID	HSSGYTRPLRKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHL
1YP8:A_PDBID	CGESCFLGTCYTKGCSCGEWKLCYGTNGGTIFD
1YTR:A_PDBID	KSSAYSLQMGATAIKQVKKLFKKWGW
1Z64:A_PDBID	GWGSFFKKAAHVGKHVGKAALTHYL
1Z6V:A_PDBID	GRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQA
1Z99:A_PDBID	YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG
1ZA8:A_PDBID	CGESCAMISFCFTEVIGCSCKNKVCYLNSIS
1ZFU:A_PDBID	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
1ZMM:A_PDBID	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
1ZMP:A_PDBID	ATCYCRTGRCATRESLSGVCEISGRLYRLCCR
1ZMQ:A_PDBID	AFTCHCRRSCYSTEYSYGTCTVMGINHRFCCL
1ZRV:A_PDBID	HVDKKVADKVLLLKQLRIMRLLTRL
1ZRX:A_PDBID	RGFRKHFNKLVKKVKHTISETAHVAKDTAVIAGSGAAVVAAT
2A2B:A_PDBID	ARSYGNGVYCNNKKCWVNRGEATQSIIGGMISGWASGLAGM
2AMN:A_PDBID	RVKRVWPLVIRTVIAGYNLYRAIKKK
2B5B:A_PDBID	EKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
2B68:A_PDBID	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
2B9K:A_PDBID	AIKLVQSPNGNFAASFVLDGTKWIFKSKYYDSSKGYWVGIYEVWDRK
2CRD:A_PDBID	EFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
2DCV:A_PDBID	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
2DCX:A_PDBID	ALWKTLLKKVLKA
2E2F:A_PDBID	AVRIGPCDQVCPRIVPERHECCRAHGRSGYAYCSGGGMYCN
2EEM:A_PDBID	SCASRCKGHCRARRCGYYVSVLYRGRCYCKCLRC
2ERI:A_PDBID	CGESCVFIPCISTLLGCSCKNKVCYRNGVIP
2G9L:A_PDBID	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
2G9P:A_PDBID	GLFGKLIKKFGRKAISYAVKKARGKH
2GDL:A_PDBID	LVQRGRFGRFLRKIRRFRPKVTITIQGSARF

Table A.1: Fernandes et al. (2012) Data Set AMP Sequences Continued...

Definition	Sequence
2GL1:A_PDBID	KTCENLANTYRGPCFTTGSCDDHCKNKEHLRSGRCRDDFRCWCTRNC
2GW9:A_PDBID	GLLCYCRKGHCKRGERVRGTCGIRFLYCCPRR
2HFR:A_PDBID	KRFWPLVPVAINTVAAGINLYKAIRRK
2JNI:A_PDBID	RWCVYAYVRIRGVLVRYRRCW
2JOS:A_PDBID	FFHHIFRGIVHVGKTIHRLVTG
2JPJ:A_PDBID	GTWDDIGQGIGRVAYWVGKALGNLSDVNQASRINRKKKH
2JPY:A_PDBID	FLSLIPHAINAVSTLVHHF
2JR3:A_PDBID	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
2JS9:A_PDBID	MSGSHHHHHHSSGIEGRGRSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECD HYVNSKVDPIIHELEGGTAPKDVCTKLNECP
2K10:A_PDBID	GILSSFKGVAKGVAKDLAGKLLETLKCKITGC
2K1I:A_PDBID	RRTCHCRSRCLRRESNSGSCNINGRIFSLCCR
2K35:A_PDBID	QIVDCWETWSRCTKWSQGGTGTLWKSCNDRCKELGRKRGQCEEKPSRCPLSKKAWTCICY
2K38:A_PDBID	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
2K6O:A_PDBID	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
2K9B:A_PDBID	GLWSKIKAAGKEAAKAAAKAAGKAALNAVSEAV
2KCN:A_PDBID	AKYTGKCTKSKNECKYKNDAGKDTFIKCPKFDNKKCTKDNNKCTVDTYNNAVDCD
2KEF:A_PDBID	DTHFPICIFCCGCCHRSKCGMCCKT
2KEG:A_PDBID	RRSRKNGIGYAIGYAFGAVERAVLGGSRDYNK
2KFE:A_PDBID	GRGREFMSNLKEKLSGVKEKMKNS
2KJF:A_PDBID	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL
2KNJ:A_PDBID	HHQELCTKGDDALVTELECIRLRISPETNAAFDNAVQQLNCLNRACAYRKMCATNNLEQAMSVYFTN EQIKEIHDAATACDPEAHHEHDH
2KSG:A_PDBID	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSVL
2KUY:A_PDBID	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGIKHHSSGSSSYHC
2L2R:A_PDBID	GSGRGSCRSQCMRRHEDEPWRVQECVSQCRRRRGGGD
2MAG:A_PDBID	GIGKFLHSAKKFGKAFVGEIMNS
2MLT:A_PDBID	GIGAVLKVLTTGLPALISWIKRKRQQ
2PCO:A_PDBID	SMWSGMWRRKLKKLRNALKKKLKGEK
2RLG:A_PDBID	ALYKKFKKKLLKSLKRLG
2RNG:A_PDBID	NPLIPAIYIGATVGPSVWAYLVALVGAAAVTAANIRRASSDNHSCAGNRGWCRSKCFRHEYVDTYYSA VCGRYFCCRSR
3HJ2:A_PDBID	CGGACYCRIPACIAGERRYGTCIYQGRLWAFCC
8TFV:A_PDBID	GSKKPVPIIYCNRRTGKCQRM

Table A.2: Fernandes et al. (2012) Data Set Non-AMP Sequences

Definition	Sequence
1AFP:A_PDBID	ATYNGKCYKKDNICKYKAQSGKTAICKCYVKKCPRDGAKCEFDSYKGKCYC
1AYJ:A_PDBID	EKLCERPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
1BNB:A_PDBID	APLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
1BRZ:A_PDBID	EDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
1C01:A_PDBID	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPF GWKSIFIQC
1CIX:A_PDBID	YSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
1CW5:A_PDBID	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGAGSIGRRP
1CZ6:A_PDBID	RSVCRQIKICRRRGGCYYKCTNRPY
1D6X:A_PDBID	VRRFPWWWPFLRR
1D7N:A_PDBID	INLKALAALAKKIL

Table A.2: Fernandes et al. (2012) Data Set Non-AMP Sequences Continued...

Definition	Sequence
1DKC:A_PDBID	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
1DQC:A_PDBID	YLAFRCGRYSPCLDDGPNVNLYSCCSFYNCHKCLARLENCPKGLHYNAYLKVCDWPSKAGCTSVNKE CHLWKT
1E4Q:A_PDBID	PVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
1E4S:A_PDBID	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK
1E4T:A_PDBID	NSKRACYREGGECLQRCIGLFHKIGTCNFRFKCCKFQ
1E68:A_PDBID	MAKEFGIPAAVAGTVLNVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVI AW
1ED0:A_PDBID	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK
1EWS:A_PDBID	MPCSCKKYCDPWEVIDGSCGLFNSKYICCREK
1FRY:A_PDBID	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG
1G6E:A_PDBID	MINRTDCNENSYLEIHNNEGRDTLCFANAGTMPVAIYGVNWVESGNNVVTLQFQRNLSDPRLETITLQ KWGSWNPGHIHEILSIRIY
1G89:A_PDBID	ILPWKWPWWPWRR
1HVZ:A_PDBID	GFCRCLCRRGVCRCICTR
1I2U:A_PDBID	DKLIGSCVWGAVNYTSDCNGECKRRGYKGGHCGSFANVNCWCET
1ICA:A_PDBID	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
1IYC:A_PDBID	ELPKLPDDKVLIRSRSNCPKGKVWNGFDCKSPFAFS
1K48:A_PDBID	NGLPVCGETCVGGTCNTPGCTCSWPVCTR
1KFP:A_PDBID	ECRRLCYKQRCVTYCRGR
1KJ6:A_PDBID	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
1KV4:A_PDBID	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKA
1L9L:A_PDBID	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRVCRTGRSRWRDVCRNFMRRYQSRVIQGLVAGETA QQICEDLR
1LFC:A_PDBID	FKCRRWQWRMKKLGAPSITCVRRAF
1M02:A_PDBID	HPLKQYWWRPSI
1MAG:A_PDBID	VGALAVVVWLWLWLW
1MM0:A_PDBID	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
1MMC:A_PDBID	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
1MQZ:A_PDBID	CTFTLPGGGGVCTLTSECI
1MR4:A_PDBID	RECKTESNTFPGICITKPPCRKACISEKFTDGHCSKILRRCLCTKPC
1MYN:A_PDBID	DCLSGRYKGPCAVWDNETCRRVCKEEGRSSGHCSPSLKCWCEGC
1NKL:A_PDBID	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAI CVDIKICKE
1OF9:A_PDBID	GEILCNLCTGLINTLENLLTTKGADKVKDYISSLCNKASGFIATLCTKVLDFGIDKLIQLIEDKVDANAIC AKIHAC
1OG7:A_PDBID	KYYGNGVHCGKHSCTVDWGTAIGNIGNNAAANWATGGNAGWNK
1P9Z:A_PDBID	ETCASRCPRPCNAGLCCSIYGYCGSGAAYCGAGNCRCQCRG
1PG1:A_PDBID	RGGRLCYCRRRFCVCVGR
1PXQ:A_PDBID	NKGCATCSIGAACLVDGPIPDFEIAGATGLGLWG
1Q3J:A_PDBID	CIKNGNGCQPNGSQGNCCSGYCHKQPGWVAGYCRRK
1Q71:A_PDBID	GGAGHVPEYFVGIGTPISFYG
1R1F:A_PDBID	TFCGETCRVIPVCTYSAALGCTCDDRSDGLCKRNGDP
1RKK:A_PDBID	RRWCFRVCYRGFCYRKCR
1RPB:A_PDBID	CLGIGSCNDFAGCGYAVVCFW
1S6W:A_PDBID	GCRFCCNCCPNMSGCGVCCRF
1T51:A_PDBID	ILGKIWEGIKSLF
1TI5:A_PDBID	RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC
1UT3:A_PDBID	SFGLCRLRRGFCARGRCRFPSIPIGRCSRFVQCCRRVW
1VM5:A_PDBID	GLFDIIKKIAESF
1XC0:A_PDBID	GFFALIPKIISSPLFKTLLSAVGSALSSSGGQE
1XKM:A_PDBID	ENREVPPGFTALIKTLRKCKII

Table A.2: Fernandes et al. (2012) Data Set Non-AMP Sequences Continued...

Definition	Sequence
1XKM:B_PDBID	NLVSGLIEARKYLEQLHRKLKNCKV
1XV3:A_PDBID	HSSGYTRPLRKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHL
1YP8:A_PDBID	CGESCFLGTCYTKGCSCGEWKLCYGTNGGTIFD
1YTR:A_PDBID	KSSAYSLQMGATAIKQVKKLFKKWGW
1Z64:A_PDBID	GWGSFFKKAAHVGKHVGKAALTHYL
1Z6V:A_PDBID	GRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQA
1Z99:A_PDBID	YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG
1ZA8:A_PDBID	CGESCAMISFCFTEVIGCSCKNKVCYLNSIS
1ZFU:A_PDBID	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
1ZMM:A_PDBID	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
1ZMP:A_PDBID	ATCYCRTGRCATRESLSGVCEISGRLYRLCCR
1ZMQ:A_PDBID	AFTCHCRRSCYSTEYSYGTCTVMGINHRFCCL
1ZRV:A_PDBID	HVDKKVADKVLLLKQLRIMRLLTRL
1ZRX:A_PDBID	RGFRKHFNKLVKKVKHTISETAHVAKDTAVIAGSGAAVVAAT
2A2B:A_PDBID	ARSYGNGVYCNNKKCWVNRGEATQSIIGGMISGWASGLAGM
2AMN:A_PDBID	RVKRVWPLVIRTVIAGYNLYRAIKKK
2B5B:A_PDBID	EKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
2B68:A_PDBID	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
2B9K:A_PDBID	AIKLVQSPNGNFAASFVLDGTKWIFKSKYYDSSKGYWVGIYEVWDRK
2CRD:A_PDBID	EFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
2DCV:A_PDBID	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
2DCX:A_PDBID	ALWKTLLKKVLKA
2E2F:A_PDBID	AVRIGPCDQVCPRIVPERHECCRAHGRSGYAYCSGGGMYCN
2EEM:A_PDBID	SCASRCKGHCRARRCGYYVSVLYRGRCYCKCLRC
2ERI:A_PDBID	CGESCVFIPCISTLLGCSCKNKVCYRNGVIP
2G9L:A_PDBID	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
2G9P:A_PDBID	GLFGKLIKKFGRKAISYAVKKARGKH
2GDL:A_PDBID	LVQRGRFGRFLRKIRRFRPKVTITIQGSARF
2GL1:A_PDBID	KTCENLANTYRGPCFTTGSCDDHCKNKEHLRSGRCRDDFRCWCTRNC
2GW9:A_PDBID	GLLCYCRKGHCKRGERVRGTCGIRFLYCCPRR
2HFR:A_PDBID	KRFWPLVPVAINTVAAGINLYKAIRRK
2JNI:A_PDBID	RWCVYAYVRIRGVLVRYRRCW
2JOS:A_PDBID	FFHHIFRGIVHVGKTIHRLVTG
2JPJ:A_PDBID	GTWDDIGQGIGRVAYWVGKALGNLSDVNQASRINRKKKH
2JPY:A_PDBID	FLSLIPHAINAVSTLVHHF
2JR3:A_PDBID	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
2JS9:A_PDBID	MSGSHHHHHHSSGIEGRGRSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECD HYVNSKVDPIIHELEGGTAPKDVCTKLNECP
2K10:A_PDBID	GILSSFKGVAKGVAKDLAGKLLETLKCKITGC
2K1I:A_PDBID	RRTCHCRSRCLRRESNSGSCNINGRIFSLCCR
2K35:A_PDBID	QIVDCWETWSRCTKWSQGGTGTLWKSCNDRCKELGRKRGQCEEKPSRCPLSKKAWTCICY
2K38:A_PDBID	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
2K6O:A_PDBID	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
2K9B:A_PDBID	GLWSKIKAAGKEAAKAAAKAAGKAALNAVSEAV
2KCN:A_PDBID	AKYTGKCTKSKNECKYKNDAGKDTFIKCPKFDNKKCTKDNNKCTVDTYNNAVDCD
2KEF:A_PDBID	DTHFPICIFCCGCCHRSKCGMCCKT
2KEG:A_PDBID	RRSRKNGIGYAIGYAFGAVERAVLGGSRDYNK
2KFE:A_PDBID	GRGREFMSNLKEKLSGVKEKMKNS
2KJF:A_PDBID	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL

Table A.2: Fernandes et al. (2012) Data Set Non-AMP Sequences Continued...

Definition	Sequence
2KNJ:A_PDBID	HHQELCTKGDDALVTELECIRLRISPETNAAFDNAVQQLNCLNRACAYRKMCATNNLEQAMSVYFTN EQIKEIHDAATACDPEAHHEHDH
2KSG:A_PDBID	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSVL
2KUY:A_PDBID	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGIKHHSSGSSSYHC
2L2R:A_PDBID	GSGRGSCRSQCMRRHEDEPWRVQECVSQCRRRRGGGD
2MAG:A_PDBID	GIGKFLHSAKKFGKAFVGEIMNS
2MLT:A_PDBID	GIGAVLKVLTTGLPALISWIKRKRQQ
2PCO:A_PDBID	SMWSGMWRRKLKKLRNALKKKLKGEK
2RLG:A_PDBID	ALYKKFKKKLLKSLKRLG
2RNG:A_PDBID	NPLIPAIYIGATVGPSVWAYLVALVGAAAVTAANIRRASSDNHSCAGNRGWCRSKCFRHEYVDTYYSA VCGRYFCCRSR
3HJ2:A_PDBID	CGGACYCRIPACIAGERRYGTCIYQGRLWAFCC
8TFV:A_PDBID	GSKKPVPIIYCNRRTGKCQRM

Table A.3: CAMP Database AMP Sequences

Definition	Sequence
AMP1	SIGTAVKKAVPIAKKVGKVAIPIAKAVLSVVGQLVG
AMP2	ELDRICGYGTARCRKKCRSQEYRIGRCPNTYACCLRKWDESLLNRTKP
AMP3	GLKDKFKSMGEKLKQYIQTWKAKF
AMP4	ICIFCCGCCHRSKCGMCCKT
AMP5	FLPIPRPILLGLL
AMP6	GLHKVMREVLGYERNSYKKFFLR
AMP7	INWKKIAEVGGKILSSL
AMP8	INWKGIAAMKKLL
AMP9	FLPMLAGLAANFLPKLFCKITKKC
AMP10	FLPIVGKLLSGLSGLL
AMP11	IDWKKLLDAAKQIL
AMP12	KQATVGDINTERPGILDLKGKAKWDAWNGLKGTSKEDAMKAYINKVEELKKKYGI
AMP13	GLLSVLGSVAKHVLPHVVPVIAEKL
AMP14	FLPLIGRVLSGIL
AMP15	GIGASILSAGKSALKGLAKGLAEHFAN
AMP16	SMWSGMWRRKLKKLRNALKKKLKGE
AMP17	GLFGKLIKKFGRKAISYAVKKARGKH
AMP18	SWKSMAKKLKEYMEKLKQRA
AMP19	GFFGKMKEYFKKFGASFKRRFANLKKRL
AMP20	SWLSKTAKKLENSAKKRISEGIAIAIQGGPR
AMP21	WNPFKELERAGQRVRDAVISAAAVATVGQAAAIARGG
AMP22	KWKIFKKIEKVGRNIRNGIIKAGPAVAVLGEAKAL
AMP23	SGISGPLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
AMP24	SLQGGAPNFPQPSQQNGGWQVSPDLGRDDKGNTRGQIEIQNKGKDHDFNAGWGKVIRGPNKAKPT WHVGGTYRR
AMP25	ATCDALSFSSKWLTVNHSACAIHCLTKGYKGGRCVNTICNCRN
AMP26	SLFSLIKAGAKFLGKNLLKQGACYAACKASKQC
AMP27	GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCYR
AMP28	FLPLLASLFSRLL
AMP29	FLPVILPVIGKLLNGILGK
AMP30	ISDYSIAMDKIRQQDFVNWLLAQKGKKSDWKHNITQ

Table A.3: CAMP Database AMP Sequences Continued...

Definition	Sequence
AMP31	KAVAAKKSPKKAKKPATPKKAAKSPKKVKKPAAAAKKAAKSPKKATKAAKPKAAKPKAAKAKKAA PKKK
AMP32	AERVGAGAPVYL
AMP33	FLPLVRGAAKLIPSVVCAISKRC
AMP34	GVITDALKGAAKTVAAELLRKAHCKLTNSC
AMP35	SIWEGIKNAGKGFLVSILDKVRCKVAGGCNP
AMP36	GLFSKFNKKKIKSGLIKIIKTAGKEAGLEALRTGIDVIGCKIKGEC
AMP37	GFFSLIKGVAKIATKGLAKNLGKMGLDLVGCKISKEC
AMP38	GFISTVKNLATNVAGTVIDTIKCKVTGGC
AMP39	WLNALLHHGLNCAKGVLA
AMP40	GIGKFLHSAGKFGKAFVGEIMKS
AMP41	RSGRGECRRQCLRRHEGQPWETQECMRRCRRRG
AMP42	ACHAHCQSVGRRGGYCGNFRMTCYCY
AMP43	AELRCMCIKTTSGIHPKNIQSLEVIGKGTHCNQVEVIATLKDGRKICLDPDAPRIKKIVQKKLAGD
AMP44	RRWCFRVCYRGFCYRKCR
AMP45	FCTMIPIPRCY
AMP46	RVCFAIPLPICH
AMP47	RRCICTTRTCRFPYRRLGTCLFQNRVYTFCC
AMP48	SLGSFLKGVGTTLASVGKVVSDQFGKLLQAGQ
AMP49	ALWKDILKNVGKAAGKAVLNTVTDMVNQ
AMP50	GWMSKIASGIGTFLSGMQQ
AMP51	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
AMP52	GKLQAFLAKMKEIAAQTL
AMP53	KVNVNAIKKGGKAIGKGFKVISAASTAHDVYEHIKNRRH
AMP54	GKIPVKAIKKGGQIIGKALRGINIASTAHDIISQFKPKKKKNH
AMP55	KGIGSALKKGGKIIKGGLGALGAIGTGQQVYEHVQNRQ
AMP56	GWASKIGQTLGKIAKVGLKELIQPK
AMP57	FLSLIPHAINAVSAIAKHFG
AMP58	VIGSILGALASGLPTLISWIKNR
AMP59	SIITMTKEAKLPQLWKQIACRLYNTC
AMP60	ENFFKEIERAGQRIRDAIISAAPAVETLAQAQKIIKGGD
AMP61	GLLRASSVWGRKYYVDLAGCAKA
AMP62	AALRGALRAVARVGKAILPHVAIANPYVRTPYVHNNP
AMP63	VRRFPWWWPFLRR
AMP64	RRRFPWVCWPFLRRR
AMP65	INLKAIAALAKKLLG
AMP66	RSVCRQIKICRRRGGCYYKCTNRPY
AMP67	FIGPIISALASLFG
AMP68	FLSLALAALPKFLCLVFKKC
AMP69	GLFSVVTGVLKAVGKNVAKNVGGSLLEQLKCKISGGC
AMP70	YVPLPNVPQPGRRPFPTFPGQGPFNPKIKWPQGY
AMP71	LRDLVCYCRTRGCKRRERMNGTCRKGHLMYTLCCR
AMP72	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
AMP73	GRFKRFRKKFKKLFKKLSPVIPLLHLG
AMP74	RIIDLLWRVRRPQKPKFVTVWVR
AMP75	GRFRRLRKKTRKRLKKIGKVLKWIPPIVGSIPLGCG
AMP76	GLLSRLRDFLSDRGRRLGEKIERIGQKIKDLSEFFQS
AMP77	RRRPRPPYLPRPRPPPFFPPRLPPRIPPGFPPRFPPRFP
AMP78	DLRFLYPRGKLPVPTPPPFNPKPIYIDMGNRY
AMP79	VCGETCVGGTCNTPGCTCSWPVCTRNGLP

Table A.3: CAMP Database AMP Sequences Continued...

Definition	Sequence
AMP80	SIPCGESCVFIPCTVTALLGCSCKSKVCYKN
AMP81	FFHHIFRGIVHVGRTIHKLVTGG
AMP82	GWKDWAKKAGGWLKKKGPGMAKAALKAAMQ
AMP83	GLVDVLGKVGGLIKKLLP
AMP84	LLKELWTKMKGAGKAVLGKIKGLL
AMP85	WLGSALKIGAKLLPSVVGLFKKKKQ
AMP86	GIWGTLAKIGIKAVPRVISMLKKKKQ
AMP87	FVQWFSKFLGRIL
AMP88	GLFDIVKKIAGHIVSSI
AMP89	GLFGVLAKVAAHVVPAIAEHF
AMP90	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSV
AMP91	GLNTLKKVFQGLHEAIKLINNHVQ
AMP92	FRGLAKLLKIGLKSFARVLKKVLPKAAKAGKALAKSMADENAIRQQNQ
AMP93	GKFSVFGKILRSIAKVFKGVGKVRKQFKTASDLDKNQ
AMP94	GVLSNVIGYLKKLGTGALNAVLKQ
AMP95	ILPWKWPWWPWRR
AMP96	SHQDCYEALHKCMASHSKPFSCSMKFHMCLQQQ
AMP97	SIGAKILGGVKTFFKGALKELASTYLQ
AMP98	GLLNTFKDWAISIAKGAGKGVLTTLSCKLDKSC
AMP99	GSKKPVPIIYCNRRTGKCQRM
AMP100	AGRGKQGGKVRAKAKTRSSRAGLQFPVGRVHRLLRKGNY
AMP101	GLLSKFGRLARKLARVIPKV
AMP102	AKKVFKRLEKLFSKIQNDK
AMP103	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
AMP104	LTCEIDRSLCLLHCRLKGYLRAYCSQQKVCRCVQ
AMP105	ANTAFVSSAHNTQKIPAGAPFNRNLRAMLADLRQNAAFAG
AMP106	FFPIGVFCKIFKTC
AMP107	FGLPMLSILPKALCILLKRKC
AMP108	GKVWDWIKSAAKKIWSSEPVSQLKGQVLNAAKNYVAEKIGATPT
AMP109	FWGALAKGALKLIPSLFSSFSKKD
AMP110	ILPLVGNLLNDLL
AMP111	GIGGKPVQTAFVDNDGIYD
AMP112	FKLGSFLKKAWKSKLAKKLRAKGKEMLKDYAKGLLEGGSEEVPGQ
AMP113	SDEKASPDKHHRFSLSRYAKLANRLANPKLLETFLSKWIGDRGNRSVK
AMP114	ILGKIWEGIKSLF
AMP115	IFGAIWNGIKS
AMP116	GLVTSLIKGAGKLLGGLFGSVTGGQS
AMP117	MLTLKKSMLLLFFLGLVSVSLADDKREDEAEEGEDKRAAEEERNVEKRCYSAAKYPGFQEFINRKYKS SRFG
AMP118	ALWKTMLKKLGTMALHAGKAALGAAADTISQGTQ
AMP119	GLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
AMP120	YRGGYTGPIPRPPPIGRPPFRPVCNACYRLSVSDARNCCIKFGSCCHLVK
AMP121	QVYKGGYTRPIPRPPPFVRPLPGGPIGPYNGCPVSCRGISFSQARSCCSRLGRCCHVGKGYS
AMP122	VFQFLGKIIHHVGNFVHGFSHVF
AMP123	GWKSVFRKAKKVGKTVGGLALDHYLG
AMP124	RWGKWFKKATHVGKHVGKAALTAYL
AMP125	GWGSIFKHIFKAGKFIHGAIQAHNDG
AMP126	FWGKLLKLGMHGIGLLHQHLG
AMP127	GWKKWFTKGERLSQRHFA
AMP128	FLGLLFHGVHHVGKWIHGLIHGHH

Table A.3: CAMP Database AMP Sequences Continued...

Definition	Sequence
AMP129	FIFHIIKGLFHAGKMIHGLVTRRRH
AMP130	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
AMP131	ACYCRIPACLAGERRYGTCFYLGRVWAFCC
AMP132	RRTCRCRFGRCFRRESYSGSCNINGRIFSLCCR
AMP133	FVPYNPPRPGQSKPFPSFPGHGPFNPKIQWPYPLPNPGH
AMP134	VTCDLLSIKGVAEHSACAANCLSMGKAGGRCENGICLCRKTTFKELWDKRF
AMP135	VTCFCKRPVCDSGETQIGYCRLGNTFYRLCCRQ
AMP136	RQRVEELSKFSKKGAAARRRK
AMP137	GWFGKAFRSVSNFYKKHKTYIHAGLSAATLL
AMP138	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
AMP139	GILSLFTGGIKALGKTLFKMAGKAGAEHLACKATNQC
AMP140	GLWSKIKAAGKEAAKAAAKAAGKAALNAVSEAV
AMP141	GLWSKIKEAAKTAGLMAMGFVNDMV
AMP142	GILDSFKGVAKGVAKDLAGKLLDKLKCKITGC
AMP143	FLGGLMKAFPALICAVTKKC
AMP144	GLFLDTLKGLAGKLLQGLKCIKAGCKP
AMP145	NFLGTLINLAKKIM
AMP146	FLPILINLIHKGLL
AMP147	RLCRIVVIRVCR
AMP148	QRPYTQPLIYYPPPPTPPRIYRA
AMP149	FKCRRWQWRMKKLGAPSITCVRRAF
AMP150	ALWMTLLKKVLKAAAKALNAVLVGANA
AMP151	GRLKKLGKKIEGAGKRVFKAAEKALPVVAGVKALG
AMP152	FIGLLISAGKAIHDLIRRRH
AMP153	IWLTALKFLGKHAAKHLAKQQLSKL
AMP154	KIKWFKTMKSIAKFIAKEQMKKHLGGE
AMP155	GIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
AMP156	NNEAQCEQAGGICSKDHCFHLHTRAFGHCQRGVPCCRTVYD
AMP157	GWLRKAAKSVGKFYYKHKYYIKAAWQIGKHAL
AMP158	IIGLVSKGTCVLVKTVCKKVLKQG
AMP159	FDITKLNIKKLTKATCKVISKGASMCKVLFDKKKQE
AMP160	NRWWQGVVPTVSYECRMNSWQHVFTCC
AMP161	GFCRCICTRGFCRCICTR
AMP162	GNNRPVYIPQPRPPHPRL
AMP163	GKPRPYSPRPTSHPRPIRV
AMP164	LFCKGGSCHFGGCPSHLIKVGSCFGFRSCCKWPWNA
AMP165	VDKGSYLPRPTPPRPIYNRN
AMP166	SIVPIRCRSNRDCRRFCGFRGGRCTYARQCLCGY
AMP167	GVWSTVLGGLKKFAKGGLEAIVNPK
AMP168	GMATKAGTALGKVAKAVIGAAL
AMP169	GFLGSLLKTGLKVGSNLL
AMP170	GFLGPLLKLAAKGVAKVIPHLIPSRQQ
AMP171	INWLKLGKAIIDAL
AMP172	ILGTILGLLKGL
AMP173	GCASRCKAKCAGRRCKGWASASFRGRCYCKCFRC
AMP174	PAQPFRFPKHPQGPQTRPPI
AMP175	DQYKCLQHGGFCLRSSCPSNTKLQGTCKPDKPNCCKS
AMP176	KCWNLRGSCREKCIKNEKLYIFCTSGKLCCLKPK
AMP177	FLRFIGSVIHGIGHLVHHIGVAL
AMP178	GFMKYIGPLIPHAVKAISDLI

Table A.3: CAMP Database AMP Sequences Continued...

Definition	Sequence
AMP179	RFRPPIRRPPIRPPFNPPFRPPVRPPFRPPFRPPFRPPIGPFPGRR
AMP180	AKRGGFWRKVGRKLGKGIRKIGKTIKSQLGKFRPRLQYRYQF
AMP181	VIPFVASVAAEMMPHVYCAASRKC
AMP182	GILLDKLKNFAKTAGKGVLQSLLNTASCKLSGQC
AMP183	QRGSRGQRCGPGEVFNQCGSACPRVCGRPPAQACTLQCVSGCFCRRGYIRTQRGGCIPERQCHQR
AMP184	QGYKSGHTGPYPRPLYGSRPIGLRPITRPDPSCAGCRILTLDDAIACCRRLGRCCSALKG
AMP185	GRKSDCFRKNGFCAFLKCPYLTLISGKCSRFHLCCKRIW
AMP186	FLFSLIPSAISGLISAFK
AMP187	AVDFSSCARMDVPGLSKVAQGLCISSCKFQNCGTGHCEKRGGRPTCVCDRCGRGGGEWPSVPMPKG RSSRGR
AMP188	EVERKHPLGGSRPGRCPTVPPGTFGHCACLCTGDASEPKGQKCCSN
AMP189	HVDKKVADKVLLLKQLRIMRLLTRL
AMP190	GFVDLAKKVVGGIRNALGI
AMP191	RVKRVWPLVIRTVIAGYNLYRAIKKK
AMP192	IIGPVLGMVGSALGGLLKKIG
AMP193	SNMIEGVFAKGFKKASH
AMP194	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
AMP195	QWGRRCCGWGPGRRYCVRWC
AMP196	GIFSSRKCKTPSKTFKGICTRDSNCDTSCRYEGYPAGDCKGIRRRCMCSKPC
AMP197	GWRTLLKKAEVKTVGKLALKHYL
AMP198	LDTIKCLQGNNNCHIQKCPWFLLQVSTCYKGKGRCCQKRRWFARSHVYHV
AMP199	GLLDAIKDTAQNLFANVLDKIKCKFTKC
AMP200	FIGAIARLLSKIF
AMP201	GGLRSLGRKILRAWKKYGPIIVPIIRIG
AMP202	FLIGMTQGLICLITRKC
AMP203	GLFTLIKGAAKLIGKTVPKKQARLGMNLWLVKLPTNVKT
AMP204	AALKGCWTKSIPPKPCFGKR
AMP205	GVLGTVKNLLIGAGKSAAQSVLKTLSCKLSNDC
AMP206	FMPILSCSRFKRC
AMP207	ATAWDFGPHGLLPIRPIRIRPLCGKDKS
AMP208	GLLSGILGAGKHIVCGLSGPCQSLNRKSSDVEYHLAKC
AMP209	GFSPNLPGKGLRIS
AMP210	FLPPSPWKETFRTS
AMP211	TSRCYIGYRRKWCS
AMP212	GCSRWIIGIHGQICRD
AMP213	GLLSGTSVRGST
AMP214	INNWVRVPPCDQVCSRTNPEKDECCRAHGHAFHATCSGGMQCYRR
AMP215	VFVALILAIAIGQSEAGWLKKIGKKIERVGQHTRDATIQGLGIAQQAANVAATAR
AMP216	LKLLKKLLKLLKKLGGGK

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao

Definition	Sequence
NonAMP1	MFTNSIKNLIIYLMPLMVTLMLLSVSFVDAGKKPSGPNPGGNN
NonAMP2	MGLKEEFEEHAEKVNTLTELPSNEDLLILYGLYKQAKFGPVDTSRPGMFSMKERAKWDAWKAVEGK SSEEAMNDYITKVKQLLEVAASKAST
NonAMP3	MAATQEEIIAGLAEIIEEVTGIEPSEVTPEKSFVDDLDIDSLSMVEIAVQTEDKYGVKIPDEDLAGLRTV GDVVAYIQKLEEENPEAAAALREKFAADQ

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao Continued...

Definition	Sequence
NonAMP4	MADTLERVTKIIVDRLGVDEADVKLEASFKEDLGADSLDVVELVMELEDEFDMEISDEDAEKIATVGD AVNYIQNQQ
NonAMP5	MEFKIMSTTDKIEQKVIEMVAEKLNKDKAIITTDSRFIEDLKADSLDTVELMMAIEVEYGIDIPDDEATK IKTVSDVIKYIKERQS
NonAMP6	MDRKEIFERIEQVLAEQLGIPAEQITEEADLREDLGMDSLDLVELVSALEDEVGMRVEQSQLEGIETVG HVMELTLDLVARLATASAADKPEAAS
NonAMP7	MVLRQLSRKASVKVSKTWSGTKKRAQRILIFLLEFLLDFCTGEDSVDGKKRQRHSGLTEQTYSALPEP KAT
NonAMP8	MSSGTPTPSNVVLIGKKPVMNYVLAALTLLNQGVSEIVIKARGRAISKAVDTVEIVRNRFLPDKIEIKEI RVGSQVVTSQDGRQSRVSTIEIAIRKK
NonAMP9	MTEKLNEIVVRKTKNVEDHVLDVIVLFNQGIDEVILKGTGREISKAVDVYNSLKDRLGDGVQLVNVQT GSEVRDRRRISYILLRLKRVY
NonAMP10	MSEPGDLSQTIVEEGGPEQETATPENGVIKSESLDEEEKLELQRRLVAQNQERRKSKSGAGKGKLTRS LAVCEESSARPGGESLQDQTL
NonAMP11	MATESPNSVQKIVVHLRATGGAPILKQSKFKVSGSDKFANVIDFLRRQLHSDSLFVYVNSAFSPNPDES VIDLYNNFGFDGKLVVNYACSMAWG
NonAMP12	MAEIKHYQFNVVMTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVDVYTTLPYDFILEKIKKTGKEVRS GKQL
NonAMP13	MKKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQLTENG AESVLQVFREAKAEGCDITIILS
NonAMP14	MLKAAAKRPELSGKNTISNNSDMAEVKSMFREVLPKQGPLFVEDIMTMVLCKPKLLPLKSLTLEKLEK MHQAAQNTIRQQEMAEKDQRQITH
NonAMP15	MSGRRSRSRQSSGTSRISEDQINDLIIKLQQLLPELRDSRRSDKVSAARVLQDTCNYIRNLHREVDDLSE RLSELLANSDTAQAALIRSLLTQ
NonAMP16	MAKAGGITNAVNVGIAVQADWENREFISHISLNVRRLFEFLVQFESTTKSKLASLNEKLDLLERRLEML EVQVSTATANPSLFAT
NonAMP17	MAGQEDPVQREIHQDWANREYIEIITSSIKKIADFLNSFDMSCRSRLATLNEKLTALERRIEYIEARVTK GETLT
NonAMP18	MASTGEIWLQWFSCCFQQQRSPSRRPHQRLRIDRSMIGNPTNFVHTGHIGSADVELSANRLNAISTQM QSKGGYETNSIHSLHAC
NonAMP19	MSEFWHKLGCCVVEKPQPKKKRRRIDRTMIGEPMNFVHLTHIGSGEMGAGDGLAMTGAVQEQMRS KGNRDRPWSNSRGL
NonAMP20	MAKVNIKPLEDKILVQANEAETTTASGLVIPDTAKEKPQEGTVVAVGPGRWDEDGEKRIPLDVAEGD TVIYSKYGGTEIKYNGEEYLILSARDVLAVVSK
NonAMP21	MKLRPLHDRVVIRRSEEESKTAGGIVLPGSAAEKPNRGEVVAVGTGRILDNGEVRALAVKVGDKVVF GPYSGSNTVKVDGEDLLVMAENEILAVIEG
NonAMP22	MKINQPAVAGTLESGDVMIRIAPLDTQDIDLQINSSVEKQFGDAIRTTILDVLARYNVRGVQLNVDDKG ALDCILRARLEALLARASGIPALPWEDCQ
NonAMP23	MSEILPYGEDKMGRFGADPEGSDLSFSCRLQDTNSFFAGNQAKRPPKLGQIGRAKRVVIEDDRIDDVL KGMGEKPPSGV
NonAMP24	MACKVKAELEAAFKKLDANGDGYVTALELQTFMVTLDAYKALSKDKVKEASAKLIKMADKNSDGKI SKEEFLNANAELLCQLK
NonAMP25	MNREGAPGKSPEEMYIQQKVRVLLMLRKMGSNLTASEEEFLRTYAGVVNSQLSQLPPHSIDQGAEDV VMAFSRSETEDRRQ
NonAMP26	MEASSEPPLDAKSDVTNQLVDFQWKLGMAVSSDTCRSLKYPYVAVMLKVADHSGQVKTKCFEMTIP QFQNFYRQFKEIAAVIETV
NonAMP27	MEQKTLQVEGMSCQHCVKAVETSVGELDGVSAVHVNLEAGKVDVSFDADKVSVKDIADAIEDQGYD VAK
NonAMP28	MKQEFSVKGMSCNHCVARIEEAVGRISGVKKVKVQLKKEKAVVKFDEANVQATEICQAINELGYQAE VI
NonAMP29	MSTAMNFGTKSFQPRPPDKGSFPLDHLGECKSFKEKFMKCLHNNNFENALCRKESKEYLECRMERKL MLQEPLEKLGFGDLTSGKSEAKK
NonAMP30	MSGNPGSSLSALRPTPPERGSFPLDHDGECTKYMQEYLKCMQLVQNENAMNCRLLAKDYLRCRMDH QLMDYDEWSHLGLPEDAPGNNGKTIKDATDNK
NonAMP31	MSSGKKAVKVKTPAGKEAELVPEKVWALAPKGRKGVKIGLFKDPETGKYFRHKLPDDYPI
NonAMP32	MTVTGQVKWFNNEKGFGFIEVPGENDVFVHFSAIETDGFKSLEEGQKVSFEIEDGNRGPQAKNVIKL
NonAMP33	MEKGTVKWFNNAKGFGFICPEGGGEDIFAHYSTIQMDGYRTLKAGQSVQFDVHQGPKGNHASVIVP VEVEAAVA
NonAMP34	MIPGGLTEAKPATIEIQEIANMVKPQLEEKTNETYEEFTAIEYKSQVVAGINYYIKIQTGDNRYIHIKVF KSLPQQSHSLILTGYQVDKTKDDELAGF

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao Continued...

Definition	Sequence
NonAMP35	MSSENMDITENIQNEQNKDFDDIDFERRRRLLTLQISKSMNEVVNLMSALNKNLESINGVGKEFENVAS LWKEFQNSVLQKKDREMLDAP
NonAMP36	MMASTSNDEEKLISTTDKYFIEQRNIVLQEINETMNSILNGLNGLNISLESSIAVGREFQSVSDLWKTLYD GLESLSDEAPIDEQPTLSQSKTK
NonAMP37	MLQARIEEKQKEYELICKLRDSSNDMVQQIETLAAKLETLTDGSEAVATVLNNWPSIFESIQIASQHSGA LVRIPPSTSNTNASATEQGDVEEV
NonAMP38	MEHNLSPLQQEVLDKYKQLSLDLKALDETIKELNYSQHRQQHSQQETVSPDEILQEMRDIEVKIGLVGT LLKGSVYSLILQRKQEQESLGSNSK
NonAMP39	MNNPMEEQQSALLGRIISNVEKLNESITRLNHSLQLINMSNMNVELASQMWANYARNVKFHLEETHTL KDPI
NonAMP40	MAVFHDEVEIEDFQYDEDSETYFYPCPCGDNFSITKEDLENGEDVATCPSCSLIIKVIYDKDQFVCGET VPAPSANKELVKC
NonAMP41	MSTYDEIEIEDMTFEPENQMFTYPCPCGDRFQIYLDDMFEGEKVAVCPSCSLMIDVVFDKEDLAEYYE EAGIHPPEPIAAAA
NonAMP42	MSSTAAKKDVKSSAVPVTAVVEKKEFEEEEFEEFPVQEWAERAEGEEDDVNVWEDNWDDETHESEF SKQLKEELRKSGHQVA
NonAMP43	MSRAALPSLENLEDDDEFEDFATENWPMKDTELDTGDDTLWENNWDDEDIGDDDFSVQLQAELKKK GVAAN
NonAMP44	MSQDFVTLVSKDDKEYEISRSAAMISPTLKAMIEGPFRESKGRIELKQFDSHILEKAVEYLNYNLKYSG VSEDDDEIPEFEIPTEMSLELLLAADYLSI
NonAMP45	MASSKQADPQTDARPLPQDFETALAELESLVSAMENGTLPLEQSLSAYRRGVELARVCQDRLAQAEQ QVKVLEGDLLRPLDPAALDDE
NonAMP46	MPKKNEAPASFEKALSELEQIVTRLESGDLPLEEALNEFERGVQLARQGQAKLQQAEQRVQILLSDNE DASLTPFTPDNE
NonAMP47	MPSTPEEKKKVLTRVRRIRGQIDALERSLEGDAECRAILQQIAAVRGAANGLMAEVLESHIRETFDRND CYSREVSQSVDDTIELVRAYLK
NonAMP48	MEDKYILLSAVETFKSRLEELLMQSAKVQKQTMLRKELASSMNDMASTVQEALNKKKSS
NonAMP49	MPYLLISTQIRMEVGPTMVGDEQSDPELMQHLGASKRRALGNNFYEYYVDDPPRIVLDKLERRGFRV LSMTGVGQTLVWCLHKE
NonAMP50	MLADKVKLSAKEILEKEFKTGVRGYKQEDVDKFLDMIIKDYETFHQEIEELQQENLQLKKQLEEASKK QPVQSNTTNFDILKRLSNLEKHVFGSKLYD
NonAMP51	MQQSLAVKTFEDLFAELGDRARTRPADSTTVAALDGGVHALGKKLLEEAGEVWLAAEHESNDALAE EISQLLYWTQVLMISRGLSLDDVYRKL
NonAMP52	MGELPIAPIGRIIKNAGAERVSDDARIALAKVLEEMGEEIASEAVKLAKHAGRKTIKAEDIELARKMFK
NonAMP53	MPKRKVSSAEGAAKEEPKRRSARLSAKPPAKVEAKPKKAAAKDKSSDKKVQTKGKRGAKGKQAEV ANQETKEDLPAENGETKTEESPASDEAGEKEAKSD
NonAMP54	MLPKATVKRIMKQHTDFNISAEAVDELCNMLEEIIKITTEVAEQNARKEGRKTIKARDIKQCDDERLK RKIMELSERTDKMPILIKEMLNVITSEL
NonAMP55	MNANKQRQYNQLAHELRELQTNLQETTKQLDIMSKQCNENLVGQLGKVHGSWLIGSYIYYMEQMLG KTQ
NonAMP56	MTMDQGLNPKQFFLDDVVLQDTLCSMSNRVNKSVKTGYLFPKDHVPSANIIAVERRGGLSDIGKNTSN
NonAMP57	MSWALEMADTFLDTMRVGPRTYADVRDEINKRGREDREAARTAVHDPERPLLRSPGLLPEIAPNASL GVAHRRTGGTVTDSPRNPVTR
NonAMP58	MQPKFNNQAKQDKLVLTGKILEIIHGDKFRVLLENNVEVDAHLAGKMRMRRLRILPGDLVEVEFSPYD LKLGRIIGRK
NonAMP59	MTKNFIVTLKKNTPDVEAKKFLDSVHHAGGSIVHEFDIIKGYTIKVPDVLHLNKLKEKHNDVIENVEED KEVHTN
NonAMP60	MKNLIAELLFKLAQKEEESKELCAQVEALEIIVTAMLRNMAQNDQQRLIDQVEGALYEVKPDASIPDD DTELLRDYVKKLLKHPRQ
NonAMP61	MSKGKKRSGARPGRPQPLRGTKGKRKGARLWYVGGQQF
NonAMP62	MATTYEEFSAKLDRLGEEFNRKMQEQNAKFFADKPDESTLSPEMKEHYEKFERMIKEHTEKFNKKM HEHSEHFKQKFAELLEQQKAAQYPSK
NonAMP63	METPLDLLKLNLDERVYIKLRGARTLVGTLQAFDSHCNIVLSDAVETIYQLNNEELSESERRCEMVFIR GDTVTLISTPSEDDDGAVEI
NonAMP64	MSGKASTEGSVTTEFLSDIIGKTVNVKLASGLLYSGRLESIDGFMNVALSSATEHYESNNNKLLNKFNS DVFLRGTQVMYISEQKI
NonAMP65	MEATLEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLAQQAAKLTSDPTDIPVVCLE SDNGNIMIQKHDGITVAVHKMAS
NonAMP66	MTVKIAQKKVLPVIGRAAALCGSCYPCSCM

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao Continued...

Definition	Sequence
NonAMP67	MTSSPSTVSTTLLSILRDDLNIDLTRVTPDARLVDDVGLDSVAFAVGMVAIEERLGVALSEEELLTCDT VGELEAAIAAKYRDE
NonAMP68	MEQITLSFPASRALSGRALAGVVGSGDMEVLYTAAQSATLNVQITTSVDNSQARWQALFDRLNLINGL PAGQLIIHDFGATPGVARIRIEQVFEEAAHA
NonAMP69	MGNIMSASFAPECTDLKTKYDSCFNEWYSEKFLKGKSVENECSKQWYAYTTCVNAALVKQGIKPALD EAREEAPFENGGKLKEVDK
NonAMP70	MSESLSWMQTGDTLALSGELDQDVLLPLWEMREEAVKGITCIDLSRVSRVDTGGLALLLHLIDLAKKQ GNNVTLQGVNDKVYTLAKLYNLPADVLPR
NonAMP71	MSTFQIHYFASASTYTGRNTESLPAPLPLSSLFDTLEAKYPGIKEKVLSSCSISLGDEYVDLVSDGEKSG NEGLLIQGGDEVAIIPPVSSG
NonAMP72	MVPLCQVEVLYFAKSAEITGVRSETISVPQEIKALQLWKEIETRHPGLADVRNQIIFAVRQEYVELGDQ LLVLQPGDEIAVIPPISGG
NonAMP73	MGECPHLVDVRLGHRSLATGPEQSDICHTGSEARWTTTWYGSLSFSRHKYKMLADLTPGVEMSCRH WARWLTPVIPALWKAEAGGLPELRSSRPAWTTW
NonAMP74	MGKLVKHCHTSLHSELVILFYAKNRFCISIDHLRPLKSHRTHGHYCLLNFSLRENKNYLIIVYLPIEGFSA NHMCISHGTSFNIK
NonAMP75	MDEEASRTARESLELVFRMSNILDTGLDRHTLSVLIALCDLGVNPEALATVVKELRRESIPDSVTTTPSI H
NonAMP76	MASSSGAGAAAAAAAAANLNAVRETMDVLLEISRILNTGLDMETLSICVRLCEQGINPEALSSVIKELRK ATEALKAAENMTS
NonAMP77	MHTNWQVCSLVVQAKSERISDISTQLNAFPGCEVAVSDAPSGQLIVVVEAEDSETLIQTIESVRNVEGV LAVSLVYHQQEEQGEETP
NonAMP78	MIPALTPEERQKLRSAILHRMQLELETTEKLIENIKEETLKKLNLLQQPDATSAPQSKELIREVLEQEGR RIE
NonAMP79	MVYYPELFVWVSQEPFPNKDMEGRLPKGRLPVPKEVNRKKNDETNAASLTPLGSSELRSPRISYLHFF
NonAMP80	MNHLQQRQLFLENLLVGVNNMFHQMQKRPVNTCCRSLQKILDHLILLQTIHSPAFRLDQMQLRQMQT LACLWIHQYNHDHQVTLGAIKWISPLIKELK
NonAMP81	MDPNQTNVVPPALHLVDPQIQLTITRMEDAMGQGQNSADPKVYPIILRLGSQLSLSMARRNLDSLEAR AFQSTPIVVQMTKLATTEELPDEFVVVTAK
NonAMP82	MGQEQDTPWILSTGHISTQKREDGQQTPRLEHHNSTRLMDHCQKTMNQVVMPKQIVYWKQWLSLRS PTPVSLKTRVLKRWRLFSKHEWTS
NonAMP83	MAKTAAALHILVKEEKLALDLLEQIKNGADFGKLAKKHSICPSGKRGGDLGEFRQGQMVPAFDKVVF SCPVLEPTGPLHTQFGYHIIKVLYRN
NonAMP84	MPHIDIKCFPRELDEQQKAALAADITDVIIRHLNSKDSSISIALQQIQPESWQAIWDAEIAPQMEALIKKP GYSMNA
NonAMP85	MCTTLFLLSTLAMLWRRRFANRVQPEPSGADGAVVGSRSERDLQSSGRKEEPLK
NonAMP86	MMTALETRLSVADGTHAAALRQRLQAALAECRRELARGACPEHFQFLQQQARALEGGLGILSQLTED
NonAMP87	MFQQEVTITAPNGLHTRPAAQFVKEAKGFTSEITVTSNGKSASAKSLFKLQTLGLTQGTVVTISAEGE DEQKAVEHLVKLMAELE
NonAMP88	MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGSDVTITVDGA DEAEGMAAIVETLQKEGLAE
NonAMP89	MAKFSAIITDKVGLHARPASVLAKEASKFSSNITIIANEKQGNLKSIMNVMAMAIKTGTEITIQADGNDA DQAIQAIKQTMIDTALIQG
NonAMP90	MAERRVNVGWAEGLHARPASIFVRAATATGVPVTIAKADGSPVNAASMLAVLGLGAQGGEEIVLASD AEGAEAALERLAKLVAEGLEELPETV
NonAMP91	MKRKIIVACGGAVATSTMAAEEIKELCQSHNIPVELIQCRVNEIETYMDGVHLICTTARVDRSFGDIPL VHGMPFVSGVGIEALQNKILTILQG
NonAMP92	MTVKQTVEITNKLGMHARPAMKLFELMQGFDAEVLLRNDEGTEAEANSVIALLMLDSAKGRQIEVEA TGPQEEEALAAVIALFNSGFDED
NonAMP93	MGTKREAILKVLENLTPEELKKFKMKLGTVPLREGFERIPRGALGQLDIVDLTDKLVASYYEDYAAEL VVAVLRDMRMLEEAARLQRAA
NonAMP94	MASSAELDFNLQALLEQLSQDELSKFKSLIRTISLGKELQTVPQTEVDKANGKQLVEIFTSHSCSYWAG MAAIQVFEKMNQTHLSGRADEHCVMPPP
NonAMP95	MSHTIRDKQKLKARASKIQGQVVALKKMLDEPHECAAVLQQIAAIRGAVNGLMREVIKGHLTEHIVHQ GDELKREEDLDVVLKVLDSYIK
NonAMP96	MTVKTGIAIGLNKGKKVTSMTPAPKISYKKGAASNRTKFVRSLVREIAGLSPYERRLIDLIRNSGEKRA RKVAKKRLGSFTRAKAKVEEMNNIIAASRRH
NonAMP97	MAREITDIKQFLELTRRADVKTATVKINKKLNKAGKPFRQTKFKVRGSSSLYTLVINDAGKAKKLIQSL PPTLKVNRL

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao Continued...

Definition	Sequence
NonAMP98	MSNPPTRPTLYNLVNADGELSYRARDLVQDFSVPIPHELPDPDLVHSIQDFATEYMLATGKKSFECLD ESALLGLGYLVTEWMDAMVTDCLEEFEER
NonAMP99	MAHENVWFSHPRRYGKGSRQCRVCSSHTGLIRKYGLNICRQCFREKANDIGFNKFR
NonAMP100	MVIATDDLEVACPKCERAGEIEGTPCPACSGKGVILTAQGYTLLDFIQKHLNK
NonAMP101	MQKYVCNVCGYEYDPAEHDNVPFDQLPDDWCCPVCGVSKDQFSPA
NonAMP102	MSNKVKTKAMVPPINCIFNFLQQQTPVTIWLFEQIGIRIKGKIVGFDEFMNVVIDEAVEIPVNSADGKE DVEKGTPLGKILLKGDNITLITSAD
NonAMP103	MSKAGAPDLKKYLDRQVFVQLNGSRKVYGVLRGYDIFLNIVLEDSIEEKVDGEKVKIGSVAIRGNSVIM IETLDKMT
NonAMP104	MSYDKVSQAKSIIIGTKQTVKALKRGSVKEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKAC GIEVGAAAVAIIL
NonAMP105	MAKRPTETERCIESLIAIFQKHAGRDGNNTKISKTEFLIFMNTELAAFTQNQKDPGVLDRMMKKLDLD SDGQLDFQEFLNLIGGLAIACHDSFIKSTQK
NonAMP106	MTKLEDHLEGIINIFHQYSVRVGHFDTLNKRELKQLITKELPKTLQNTKDQPTIDKIFQDLDADKDGAV SFEEFVVLVSRVLKTAHIDIHKE
NonAMP107	MAAEPLTELEESIETVVTTFFTFARQEGRKDSLSVNEFKELVTQQLPHLLKDVGSLDEKMKSLDVNQD SELKFNEYWRLIGELAKEIRKKKDLKIRKK
NonAMP108	MSGLRVYSTSVTGSREIKSQQSEVTRILDGKRIQYQLVDISQDNALRDEMRALAGNPKATPPQIVNGDQ YCGDYELFVEAVEQNTLQEFLKLA
NonAMP109	MAQYQTWEEFSRAAEKLYLADPMKARVVLKYRHSDGSLCIKVTDDLVCLVYRTDQAQDVKKIEKFH SQLMRLMVAKESRSVAMETD
NonAMP110	MIIWPSYIDKKKSRREGRKVPEELAIEKPSLKDIEKALKKLGLEPKIYRDKRYPRQHWEICGCVEVDYK GNKLQLLKEICKIIKGKN
NonAMP111	MGRFVVWPSELDSRLSRKYGRIVPRSIAVESPRVEEIVRAAEELKFKVIRVEEDKLNPRLSGIDEELRTF GMIVLESPYGKSKSLKLIAQKIREFRRRSA
NonAMP112	MSANQEEDKKPGDGGAHINLKVKGQDGNEVFFRIKRSTQLKKLMNAYCDRQSVDMNSIAFLFDGRRL RAEQTPDELDMEDGDEIDAMLHQTGGSGGGATA
NonAMP113	MFSNIGIPGLILIFVIALIIFGPSKLPEIGRAAGRTLLEFKSATKSLVSGDEKEEKSAELTAVKQDKNAG
NonAMP114	MFETITPLFPGLPGGPELLVVLLIVVLLFGANKIPKLARSSGQAIGEFQRGREEIEDELQDMTGDDDED DATSESSADSVSTDSVSTESSN
NonAMP115	MDPVDPNLEPWNHPGSQPRTPCNKCYCKKCCYHCQMCFITKGLGISYGRKKRRQRRRPPQGNQAHQ DPLPEQPSSQHRGDHPTGPKE
NonAMP116	MSKVKIELFTSPMCPHCPAAKRVVEEVANEMPDAVEVEYINVMENPQKAMEYGIMAVPTIVINGDVE FIGAPTKEALVEAIKKRL
NonAMP117	MNSVGEACTDMKREYDQCFNRWFAEKFLKGDSSGDPCTDLFKRYQQCVQKAIKEKEIPIEGLEFMGH GKEKPENSS
NonAMP118	MTDLFSSPDHTLDALGLRCPEPVMMVRKTVRNMQPGETLLIIADDPATTRDIPGFCTFMEHELVAKE TDGLPYRYLIRKGG
NonAMP119	MLHTLHRSPWLTDFAALLRLLSEGDELLLLQDGVTAAVDGNRYLESLRNAPIKVYALNEDLIARGLTG QISNDIILIDYTDFVRLTVKHPSQMAW
NonAMP120	MSYIPGQPVTAVVQRVEIHKLRQGENLILGFSIGGGIDQDPSQNPFSEDKTDKVNGWDMTMVTHDQA RKRLTKRSEEVVRLLVTRQSLQKAVQQSMLS
NonAMP121	MLLPATMSDKPDMAEIEKFDKSKLKKTETQEKNPLPSKETIEQEKQAGES
NonAMP122	MENKNLHIIAACGNGMGTSMLIKIKVEKIMKELGYTAKVEALSMGQTKGMEHSADIIISSIHLTSEFNPN AKAKIVGVLNLMDENEIKQALSKVL
NonAMP123	MNLTPREKDKLLIAMAAMVARRRLERGVKLNHPEAIALVSDFVVEGARDGRTVAELMEAGAHVITRE QVMDGVAEMIRDIQVEATFPDGTKLVTVHEPIR
NonAMP124	MVNVKVEFLGGLDAIFGKQRVHKIKMDKEDPVTVGDLIDHIVSTMINNPNDVSIFIEDDSIRPGIITLIND TDWELEGEKDYILEDGDIISFTSTLHGG
NonAMP125	YYVLHLCLAATKYPLLKLLGSTWPTTPPRPIPKPSPWAPKKHRRLSSDQDQSQTPETPATPLSCCTET QWTVLQSSLHLTAHTKDGLTVIVTLHP
NonAMP126	MSSDLRLTLLELVRRLNGNATIESGRLPGGRRRSPDTTTGTTGVTKTTEGPKECIDPTSRPAPEGPQE EPLHDLRPRPANRKGAAVE
NonAMP127	MSNSDLNIERINELAKKKKEVGLTQEEAKEQTALRKAYLESFRKGFKQQIENTKVIDPEGNDVTPEKIK EIQQKRDNKN
NonAMP128	MDFSKMGELLNQVQEKAKNIELELANREFSAKSGAGLVKVSANGKGEIIDVSIDDSLLEDKESLQILLIS AINDVLAMVAQNRSSMANDVLGGFGGMKL
NonAMP129	MSSALYKQSTNFTHSTGSFLQSAPVELTTVSGYQEFLKKQEKKNYEIQTVLSEDKSHGYVLKDGEVIA NIIGEAKDYLLDLAGQA

Table A.4: CAMP Database Paired Non-AMP Sequences from Xiao Continued...

Definition	Sequence
NonAMP130	MRHHQNMHYAPQQQPVYVQQPPPRRESGGCCRTCCHFLCCLCLINLCCDVF
NonAMP131	MLETVPVRCVERKITSLVVDLSGVPIVDTMVAQQLYNLSKTLFLLGVKAVFSGIRPDVAQTSIQLGLDF SEYETYGTLKQALENMGVRCIVEELEENK
NonAMP132	MLGMIRWVVEGTLVAMLLSAIRRETGMIFFYNQYQLGGWIHRYLSWGEMCYTRTLKMVKRSKFFRK QLNEDGFGRINDSGPKRRGRDQSQYSSRFVELD
NonAMP133	MKSSIPITEVLPRAVGSLTFDENYNLLDTSGVAKVIEKSPIAEIIRKSNAELGRLGYSVYEDAQYIGHAFK KAGHFIVYFTPKNKNREGVVPPVGITN
NonAMP134	MSDKPDSQVFCPNCNERLQKCLVQQNYAIIICPSLVCGYPFNQREVLENLTYVDDNDVLKVAKKRLSS RSKP
NonAMP135	MSTEKLEASEEPQAPLANTSETNSIKGDTENIVTVFDLANEIEKSLKDVQRQMKENDDEFSRSIQAIED KLNKMSR
NonAMP136	MQYCELDLSGQWLDTVYCEENFSDFVFIKFLNPSQFEEKIYCYTLHITKRTLENKRLLLYYEDEFKKH GHDINELVGDGIILRSCWNPRQ
NonAMP137	MIERELGNWKDFIEVMLRK
NonAMP138	MKKKPVAQLERQHSLLENPCAYGLLSQFQAAIVVNCFTLNKII
NonAMP139	MSDGKKTKTTVDIYGQHFTIVGEESRAHMRYVAGIVDDKMREINEKNPYLDINKLAVLTAVNVVHDY VKLQEKCEKLERQLKEKD
NonAMP140	MTMSLEVFEKLEAKVQQAIDTITLLQMEIEELKEKNNSLSQEVQNAQHQREELERENNHLKEQQNGW QERLQALLGRMEEV
NonAMP141	MGNKQAKAPESKDSPRASLIPDATHLGPQFCKSCWFENKGLVECNNHYLCLNCLTLLLSVSNRCPICK MPLPTKLRPSAAPTAPPTGAADSIRPPPYSP
NonAMP142	MGNSKSKSKLSANQYEQQTVNSTKQVAILKRQAEPSLYGRHNCRCCWFANTNLIKCSDHYICLKCLNI MLGKSSFCDICGEELPTSIVVPIEPSAPPPED
NonAMP143	MERKKLIAKFVEIASEKMGKDLETVDEENTFKELGFDSIDVIDLVMFFEDEFALRIEDEEISKIRKVKDL IDIVIKKLEEIDDEVSEG
NonAMP144	MSEEIKAQVMESVIGCLKLNDEQKQILSGTTNLAKDFNLDSLDFVDLIMSLEERFSLEISDEDAQKLETV DDICRYIASKSSDA
NonAMP145	MKRQKRDRLERAQSQGYKAGLNGRSQEACPYQQVDARSYWLGGWRDARDEKQSGLYK

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences

Definition	Sequence
AP00001	GLWSKIKEVGKEAAKAAAKAAGKAALGAVSEAV
AP00004	NLCERASLTWTGNCGNTGHCDTQCRNWESAKHGACHKRGNWKCFCYFDC
AP00005	VFIDILDKVENAIHNAAQVGIGFAKPFEKLINPK
AP00006	GNNRPVYIPQPRPPHPRI
AP00008	RLCRIVVIRVCR
AP00020	GLFDIVKKIAGHIAGSI
AP00026	FKCRRWQWRMKKLGAPSITCVRRAF
AP00027	ITPATPFTPAIITEITAAVIA
AP00035	KSSAYSLQMGATAIKQVKKLFKKWGW
AP00050	GIGASILSAGKSALKGLAKGLAEHFAN
AP00064	ILGPVLGLVGNALGGLIKNE
AP00066	IKITTMLAKLGKVLAHV
AP00070	INVLGILGLLGKALSHL
AP00071	FLPAIFRMAAKVVPTIICSITKKC
AP00078	GILLDKLKNFAKTAGKGVLQSLLNTASCKLSGQC
AP00080	GIFSKLGRKKIKNLLISGLKNVGKEVGMDVVRTGIDIAGCKIKGEC
AP00088	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
AP00091	GLLNTFKDWAISIAKGAGKGVLTTLSCKLDKSC
AP00095	LLPIVGNLLKSLL
AP00101	FVQWFSKFLGRIL

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00102	GSKKPVPIIYCNRRTGKCQRM
AP00104	FLPFLAKILTGVL
AP00109	VLPLISMALGKLL
AP00110	NFLGTLINLAKKIM
AP00111	FLPILINLIHKGLL
AP00115	GLFLDTLKGAAKDVAGKLEGLKCKITGCKLP
AP00116	GFLDIINKLGKTFAGHMLDKIKCTIGTCPPSP
AP00126	GGLKKLGKKLEGVGKRVFKASEKALPVAVGIKALGK
AP00134	SWLSKTAKKLENSAKKRISEGIAIAIQGGPR
AP00136	FLPLILRKIVTAL
AP00137	LRDLVCYCRTRGCKRRERMNGTCRKGHLMYTLCCR
AP00140	SQLGDLGSGAGQGGGGGGSIRAAGGAFGKLEAAREEEFFYKKQKEQLERLKNDQIHQAEFHHQQIKE HEEAIQRHKDFLNNLHK
AP00144	GIGKFLHSAKKFGKAFVGEIMNS
AP00145	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGAGSIGRRP
AP00146	GIGAVLKVLTTGLPALISWIKRKRQQ
AP00149	MPCSCKKYCDPWEVIDGSCGLFNSKYICCREK
AP00150	ILPWKWPWWPWRR
AP00152	VRRFPWWWPFLRR
AP00153	RSVCRQIKICRRRGGCYYKCTNRPY
AP00154	YSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
AP00155	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG
AP00156	RCVCRRGVCRCVCRRGVC
AP00157	ALWKTMLKKLGTMALHAGKAALGAAADTISQGTQ
AP00163	ALWKDILKNVGKAAGKAVLNTVTDMVNQ
AP00167	GWMSKIASGIGTFLSGMQQ
AP00168	GRPNPVNNKPTPHPRL
AP00170	VDKGSYLPRPTPPRPIYNRN
AP00171	HRHQGPIFDTRPSPFNPNQPRPGPIY
AP00175	DSHEERHHGRHGHHKYGRKFHEKHHSHRGYRSNYLYDN
AP00179	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
AP00184	RSGRGECRRQCLRRHEGQPWETQECMRRCRRRG
AP00186	GRCVCRKQLLCSYRERRIGDCKIRGVRFPFCCPR
AP00191	ECRRLCYKQRCVTYCRGR
AP00193	DTHFPICIFCCGCCHRSKCGMCCKT
AP00195	RGGRLCYCRRRFCVCVGR
AP00196	WYVKKCLNDVGICKKKCKPEEMHVKNGWAMCGKGRDCCVPAD
AP00197	QLKKCWNNYVQGHCRKICRVNEVPEALCENGRYCCLNIKELEAC
AP00198	MRILYLLFSVLFLVLQVSPGLSLPQRDMFLCRIGSCHFGRCPIHLVRVGSCFGFRSCCKSPWDV
AP00201	INLKALAALAKKIL
AP00204	ITSISLCTPGCKTGALMGCNMKTATCNCSIHVSK
AP00207	TAGPAIRASVKQCQKTLKATRLFTVSCKGKNGCK
AP00208	RADTQTYQPYNKDWIKEKIYVLLRRQAQQAGK
AP00209	GVLSNVIGYLKKLGTGALNAVLKQ
AP00211	RRWCFRVCYRGFCYRKCR
AP00217	GICACRRRFCPNSERFSGYCRVNGARYVRCCSRR
AP00228	LTCEIDRSLCLLHCRLKGYLRAYCSQQKVCRCVQ
AP00233	GWIRDFGKRIERVGQHTRDATIQTIAVAQQAANVAATLKG
AP00234	SDEKASPDKHHRFSLSRYAKLANRLANPKLLETFLSKWIGDRGNRSV
AP00236	KSCCRNTWARNCYNVCRLPGTISREICAKKCDCKIISGTTCPSDYPK

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00239	GWASKIGQTLGKIAKVGLKELIQPK
AP00240	GLLSVLGSVAKHVLPHVVPVIAEHL
AP00257	GLWQKIKSAAGDLASGIVEGIKS
AP00261	GLFVGLAKVAAHNNPAIAEHFQA
AP00262	GFVDFLKKVAGTIANVVT
AP00263	GLLQTIKEKLESLESLAKGIVSGIQA
AP00266	GKREKCLRRNGFCAFLKCPTLSVISGTCSRFQVCCKTLLG
AP00272	DQYKCLQHGGFCLRSSCPSNTKLQGTCKPDKPNCCKS
AP00273	SIVPIRCRSNRDCRRFCGFRGGRCTYARQCLCGY
AP00275	GVIPCGESCVFIPCISTLLGCSCKNKVCYRN
AP00276	VFQFLGKIIHHVGNFVHGFSHVF
AP00283	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
AP00284	SHQDCYEALHKCMASHSKPFSCSMKFHMCLQQQ
AP00288	TCESPSHKFKGPCATNRNCES
AP00294	DVQCGEGHFCHDQTCCRASQGGACCPYSQGVCCADQRHCCPVGF
AP00295	EVERKHPLGGSRPGRCPTVPPGTFGHCACLCTGDASEPKGQKCCSN
AP00301	GILDTIKSIASKVWNSKTVQDLKRKGINWVANKLGVSPQAA
AP00303	FCTMIPIPRCY
AP00304	RVCFAIPLPICH
AP00311	CYSAAKYPGFQEFINRKYKSSRF
AP00315	SLGSFLKGVGTTLASVGKVVSDQFGKLLQAGQG
AP00316	GIVDFAKKVVGGIRNALGI
AP00323	GVLDAFRKIATVVKNVV
AP00324	GVGDLIRKAVSVIKNIV
AP00325	GVIDAAKKVVNVLKNLF
AP00327	GWFDVVKHIASAV
AP00328	GFGKAFHSVSNFAKKHKTA
AP00330	GWLRKAAKSVGKFYYKHKYYIKAAWQIGKHAL
AP00332	GCASRCKAKCAGRRCKGWASASFRGRCYCKCFRC
AP00334	FFHHIFRGIVHVGKTIHKLVTG
AP00338	PDPAKTAPKKGSKKAVTKA
AP00339	FFGWLIKGAIHAGKAIHGLIHRRRH
AP00342	AKCIKNGKGCREDQGPPFCCSGFCYRQVGWARGYCKNR
AP00354	VTCDILSVEAKGVKLNDAACAAHCLFRGRSGGYCNGKRVCVCR
AP00355	ANTAFVSSAHNTQKIPAGAPFNRNLRAMLADLRQNAAFAG
AP00356	QRFIHPTYRPPPQPRRPVIMRA
AP00357	FFPIGVFCKIFKTC
AP00358	FGLPMLSILPKALCILLKRKC
AP00359	DLRFLYPRGKLPVPTPPPFNPKPIYIDMGNRY
AP00364	VDKPDYRPRPWPRNMI
AP00369	RIIDLLWRVRRPQKPKFVTVWVR
AP00370	VGRFRRLRKKTRKRLKKIGKVLKWIPPIVGSIPLGCG
AP00371	GLLSRLRDFLSDRGRRLGEKIERIGQKIKDLSEFFQS
AP00376	GWKDWAKKAGGWLKKKGPGMAKAALKAAMQ
AP00382	GLVDVLGKVGGLIKKLLPG
AP00387	WLGSALKIGAKLLPSVVGLFQKKKK
AP00388	GIWGTLAKIGIKAVPRVISMLKKKKQ
AP00391	FIGTALGIASAIPAIVKLFK
AP00392	YRGGYTGPIPRPPPIGRPPLRLVVCACYRLSVSDARNCCIKFGSCCHLVK
AP00394	QVYKGGYTRPIPRPPPFVRPLPGGPIGPYNGCPVSCRGISFSQARSCCSRLGRCCHVGKGYS

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00395	HSSGYTRPLPKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHR
AP00397	SGFVLKGYTKTSQ
AP00399	HVDKKVADKVLLLKQLRIMRLLTRL
AP00400	YPPKPESPGEDASPEEMNKYLTALRHYINLVTRQRY
AP00401	GFTQGVRNSQSCRRNKGICVPIRCPGSMRQIGTCLGAQVKCCRRK
AP00402	KTCENLANTYRGPCFTTGSCDDHCKNKEHLRSGRCRDDFRCWCTRNC
AP00403	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
AP00405	FISAIASMLGKFL
AP00408	FLFPLITSFLSKVL
AP00409	ATTGCSCPQCIIFDPICASSYKNGRRGFSSGCHMRCYNRCHGTDYFQISKGSKCI
AP00410	PKRKSATKGDEPARRSARLSARPVPKPAAKPKKAAAPKKAVKGKKAAENGDAKAEAKVQAAGDGAGNAK
AP00411	KAVAAKKSPKKAKKPATPKKAAKSPKKVKKPAAAAKKAAKSPKKATKAAKPKAAKPKAAKAKKAAPKKK
AP00413	SLQGGAPNFPQPSQQNGGWQVSPDLGRDDKGNTRGQIEIQNKGKDHDFNAGWGKVIRGPNKAKPTWHVGGTYRR
AP00417	SIGTAVKKAVPIAKKVGKVAIPIAKAVLSVVGQLVG
AP00418	GLRKRLRKFRNKIKEKLKKIGQKIQGFVPKLAPRTDY
AP00424	GFLGPLLKLAAKGVAKVIPHLIPSRQQ
AP00425	GCWSTVLGGLKKFAKGGLEAIVNPK
AP00428	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPFGWKSIFIQC
AP00429	GLICESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKEKTGLI
AP00430	ILGKIWEGIKSLF
AP00431	TWLKKRRWKKAKPP
AP00432	KKKKPLFGLFFGLF
AP00433	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSV
AP00437	EFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
AP00439	VTCFCKRPVCDSGETQIGYCRLGNTFYRLCCRQ
AP00449	SYSMEHFRWGKPV
AP00455	FFPIVAGVAGQVLKKIYCTISKKC
AP00456	VNPIILGVLPKFVCLITKKC
AP00475	GLNTLKKVFQGLHEAIKLINNHVQ
AP00480	VGIGTPIFSYGGGAGHVPEYF
AP00481	FFSASCVPGADKGQFPNLCRLCAGTGENKCA
AP00482	FSFKRLKGFAKKLWNSKLARKIRTKGLKYVKNFAKDMLSEGEEAPPAAEPPVEAPQ
AP00484	RGFRKHFNKLVKKVKHTISETAHVAKDTAVIAGSGAAVVAAT
AP00485	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
AP00489	SGRGKTGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAHRVGAGAPVYL
AP00492	RQRVEELSKFSKKGAAARRRK
AP00493	NLVSGLIEARKYLEQLHRKLKNCKV
AP00496	AKKVFKRLEKLFSKIQNDK
AP00499	VGALAVVVWLWLWLW
AP00501	GIGKHVGKALKGLKGLLKGLGES
AP00502	FLRFIGSVIHGIGHLVHHIGVAL
AP00503	FLGVVFKLASKVFPAVFGKV
AP00504	LAHQKPFIRKSYKCLHKRCR
AP00509	VAIALKAAHYHTHKE
AP00510	ILQKAVLDCLKAAGSSLSKAAITAIYNKIT
AP00511	KIGAKIKIGAKIKIGAKI

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00514	FLGGLMKAFPALICAVTKKC
AP00516	IWLTALKFLGKHAAKHLAKQQLSKL
AP00517	KIKWFKTMKSIAKFIAKEQMKKHLGGE
AP00518	QYRHRCCAWGPGRKYCKRWC
AP00522	INWLKLGKAIIDAL
AP00524	GIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
AP00531	GKQYFPKVGGRLSGKAPLAAKTHRRLKP
AP00535	GLGSVFGRLARILGRVIPKVAKKLGPKVAKVLPKVMKEAIPMAVEMAKSQEEQQPQ
AP00536	SVRTQDNAVNRQIFGSNGPYRDFQLSDCYLPLETNPYCNEWQFAYHWNNALMDCERAIYHGCNRTR NNFITLTACKNQAGPICNRRRH
AP00538	WLNALLHHGLNCAKGVLA
AP00542	ILGTILGLLKSL
AP00543	GVVDILKGAGKDLLAHLVGKISEKV
AP00544	GVLDIFKDAAKQILAHAAEKQI
AP00546	FLSLIPHAINAVSAIAKHN
AP00547	RSNKGFNFMVDMIQALSK
AP00549	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
AP00552	GIGRKFLGGVKTTFRCGVKDFASKHLY
AP00556	GFMKYIGPLIPHAVKAISDLI
AP00558	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
AP00559	ATRVVYCNRRSGSVVGGDDTVYYEG
AP00560	TTLTLHNLCPYPVWWLVTPNNGGFPIIDNTPVVLG
AP00561	GWKIGKKLEHHGQNIRDGLISAGPAVFAVGQAATIYAAAK
AP00562	GKVWDWIKSAAKKIWSSEPVSQLKGQVLNAAKNYVAEKIGATPT
AP00564	FLIGMTHGLICLISRKC
AP00567	VWPLGLVICKALKIC
AP00568	GLFSVVTGVLKAVGKNVAKNVGGSLLEQLKCKKISGGC
AP00569	FLPLLLAGLPLKLCFLFKKC
AP00570	SIITMTKEAKLPQLWKQIACRLYNTC
AP00583	GVITDALKGAAKTVAAELLRKAHCKLTNSC
AP00584	VIDDLKKVAKKVRRELLCKKHHKKLN
AP00588	FLGSIVGALASALPSLISKIRN
AP00589	FLGALAKIISGIF
AP00594	EGTWQHGYGVSSAYSNYHHGSKTHSATVVNNNTGRQGKDTQRAGVWAKATVGRNLTEKASFYYNF W
AP00598	FLSAITSLLGKLL
AP00599	GIWDTIKSMGKVFAGKILQNL
AP00600	GLLRASSVWGRKYYVDLAGCAKA
AP00605	ILPILSLIGGLLGK
AP00611	FIGPIISALASLFG
AP00612	AAEFPDFYDSEEQMGPHQEAEDEKDRADQRVLTEEEKKELENLAAMDLELQKIAEKFSQR
AP00613	RVKRFWPLVPVAINTVAAGINLYKAIRRK
AP00615	ALFSILRGLKKLGNMGQAFVNCKIYKKC
AP00619	GFFSTVKNLATNVAGTVIDTLKCKVTGGCRS
AP00621	GIFPKIIGKGIKTGIVNGIKSLVKGVGMKVFKAGLNNIGNTGCNEDEC
AP00624	ALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
AP00630	GEILCNLCTGLINTLENLLTTKGADKVKDYISSLCNKASGFIATLCTKVLDFGIDKLIQLIEDKVDANAIC AKIHAC
AP00639	GLIGSIGKALGGLLVDVLKPKLQAAS
AP00641	GFFALIPKIISSPLFKTLLSAVGSALSSSGEQE
AP00658	FLPLVGKILSGLI

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00659	FLPIASLLGKYL
AP00661	GILSLFTGGIKALGKTLFKMAGKAGAEHLACKATNQC
AP00662	GLFSILRGAAKFASKGLGKDLTKLGVDLVACKISKQC
AP00664	FLPAIAGILSQLF
AP00667	EPHPDEFVGLM
AP00660	FWGALAKGALKLIPSLFSSFSKKD
AP00671	EPNPNEFFGLM
AP00673	VGSRYLCTPGSCWKLVCFTTTVK
AP00674	ITSVSWCTPGCTSEGGGSGCSHCC
AP00675	FELDRICGYGTARCRKKCRSQEYRIGRCPNTYACCLRKWDESLLNRTKP
AP00677	GLRKKFRKTRKRIQKLGRKIGKTGRKVWKAWREYGQIPYPCRI
AP00684	RRLRPRRPRLPRPRPRPRPRPRPRSLPLPRPKPRPIPRPLPLPRPRPKPIPRPLPLPRPRPRRIPRPLPLPR PRPRPIPRPLPLPQPQPSPIPRPL
AP00686	KRFGRLAKSFLRMRILLPRRKILLAS
AP00687	KRRHWFPLSFQEFLEQLRRFRDQLPFP
AP00688	KRFHSVGSLIQRHQQMIRDKSEATRHGIRIITRPKLLLAS
AP00689	AFPPPNVPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPPFPPPIFPGPWFPPPPPFR PPPFGPPRFP
AP00691	GFFKKAWRKVKHAGRRVLDTAKGVGRHYVNNWLNRYR
AP00693	RICSRDKNCVSRPGVGSIIGRPGGGSLIGRPGGGSVIGRPGGGSPPGGGSFNDEFIRDHSDGNRFA
AP00696	GLFDIIKNIVSTL
AP00714	GYGCPFNQYQCHSHCSGIRGYKGGYCKGTFKQTCKCY
AP00722	GLLNGLALRLGKRALKKIIKRLCR
AP00723	SLLSLIRKLIT
AP00727	RWCVYAYVRVRGVLVRYRRCW
AP00748	DIQIPGIKKPTHRDIIIPNWNPNVRTQPWQRFGGNKS
AP00749	EADEPLWLYKGDNIERAPTTADHPILPSIIDDVKLDPNRRYA
AP00750	EIRLPEPFRFPSPTVPKPIDIDPILPHPWSPRQTYPIIARRS
AP00752	DKLIGSCVWGATNYTSDCNAECKRRGYKGGHCGSFWNVNCWCEE
AP00753	VQETQKLAKTVGANLEETNKKLAPQIKSAYDDFVKQAQEVQKKLHEAASKQ
AP00754	ETESTPDYLKNIQQQLEEYTKNFNTQVQNAFDSDKIKSEVNNFIESLGKILNTEKKEAPK
AP00764	GLRSKIWLWVLLMIWQESNKFKKM
AP00765	MHDFWVLWVLLEYIYNSACSVLSATSSVSSRVLNRSLQVKVVKITN
AP00766	IYWIADQFGIHLATGTARKLLDAMASGASLGTAFAAILGVTLPAWALAAAGALGATAA
AP00767	VAGALGVQTAAATTIVNVILNAGTLVTVLGIIASIASGGAGTLMTIGWATFKATVQKLAKQSMARAIA Y
AP00768	PNWTKIGKCAGSIAWAIGSGLFGGAKLIKIKKYIAELGGLQKAAKLLVGATTWEEKLHAGGYALINLA AELTGVAGIQANCF
AP00772	FRGLAKLLKIGLKSFARVLKKVLPKAAKAGKALAKSMADENAIRQQNQ
AP00773	GKFSVFGKILRSIAKVFKGVGKVRKQFKTASDLDKNQ
AP00779	GRRKRKWLRRIGKGVKIIGGAALDHL
AP00780	GRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQA
AP00784	FFRLLFHGVHHVGKIKPRA
AP00785	GWKSVFRKAKKVGKTVGGLALDHYL
AP00788	AGWGSIFKHIFKAGKFIHGAIQAHND
AP00789	GFWGKLFKLGLHGIGLLHLHL
AP00791	GWKKWLRKGAKHLGQAAIKGLAS
AP00805	RSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECDHYVNSKVDPIIHELEGGT APKDVCTKLNECP
AP00806	HHQELCTKGDDALVTELECIRLRISPETNAAFDNAVQQLNCLNRACAYRKMCATNNLEQAMSVYFTN EQIKEIHDAATACDPEAHHEHDH
AP00807	NRWYCNSAAGGVGGAAGCVLAGYVGEAKENIAGEVRKGWGMAGGFTHNKACKSFPGSGWASG

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00810	QSHLSLCRWCCNCCRSNKGC
AP00812	FAEPLPSEEEGESYSKEPPEMEKRYGGFM
AP00737	GLVTSLIKGAGKLLGGLFGSVTGGQS
AP00741	PITYLDAILAAVRLLNQRISGPCILRLREAQPRPGWVGTLQRRREVSFLVEDGPCPPGVDCRSCEPGAL QHCVGTVSIEQQPTAELRCRPLRPQ
AP00743	RYHMQCGYRGTFCTPGKCPYGNAYLGLCRPKYSCCRWL
AP00744	GLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCCRSRWYS
AP00814	GLGSILGKILNVAGKVGKTIGKVADAVGNKE
AP00817	FLPLLASLFSRLL
AP00824	SILPTIVSFLSKFL
AP00835	GKIPVKAIKKGGQIIGKALRGINIASTAHDIISQFKPKKKKNH
AP00840	KGIGSALKKGGKIIKGGLGALGAIGTGQQVYEHVQNRQ
AP00841	TTHSGKYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAIPGKC
AP00847	KYYGNGLSCNKSGCSVDWSKAISIIGNNAVANLTTGGAAGWKS
AP00849	TSYGNGVHCNKSKCWIDVSELETYKAGTVSNPKDILW
AP00857	SSMKLSFRARAYGFRGPGPQL
AP00876	FLSIIAKVLGSLF
AP00879	GRLRNLIEKAGQNIRGKIQGIGRRIKDILKNLQPRPQV
AP00884	QLKVDLWGTRSGIQPEQHSSGKSDVRRWRSRY
AP00887	GILSTFKGLAKGVAKDLAGNLLDKFKCKITGC
AP00891	IIGLVSKGTCVLVKTVCKKVLKQG
AP00892	PDITKLNIKKLTKATCKVISKGASMCKVLFDKKKQE
AP00893	DVKGMKKAIKGILDCVIEKGYDKLAAKLKKVIQQLWE
AP00894	GLLDFVTGVGKDIFAQLIKQI
AP00895	KRFKKFFKKLKNSVKKRAKKFFKKPRVIGVSIPF
AP00915	QQCGRQAGNRRCANNLCCSQYGYCGRTNEYCCTSQGCQSQCRRCG
AP00928	NKGCATCSIGAACLVDGPIPDFEIAGATGLFGLWG
AP00929	MAKEFGIPAAVAGTVINVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVIA W
AP00930	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGVKHSSGGGGSYHC
AP00931	LAGYTGIASGTAKKVVDAIDKGAAAFVIISIISTVISAGALGAVSASADFIILTVKNYISRNLKAQAVIW
AP00973	LLGMIPLAISAISALSKL
AP00987	SRWPSPGRPRPFPGRPKPIFRPRPCNCYAPPCPCDRW
AP00991	GSNFCDSKCKLRCSKAGLADRCLKYCGICCEECKCVPSGTYGNKHECPCYRDKKNSKGKSKCP
AP00992	YSYKKIDCGGACAARCRLSSRPRLCNRACGTCCARCNCVPPGTSGNTETCPCYASLTTHGNKRKCP
AP00993	GIFSSRKCKTPSKTFKGICTRDSNCDTSCRYEGYPAGDCKGIRRRCMCSKPC
AP00994	GIFSNMYARTPAGYFRGPAGYAAN
AP00996	ISLEICAIFHDN
AP00998	ALPKKLKYLNLFNDGFNYMGVV
AP01001	NRWWQGVVPTVSYECRMNSWQHVFTCC
AP01003	FKSWSFCTPGCAKTGSFNSYCC
AP01004	DWTAWSALVAAACSVELL
AP01005	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
AP01009	DYDWSLRGPPKCATYGQKCRTWSPRNCCWNLRCKAFRCRPR
AP01010	SMWSGMWRRKLKKLRNALKKKLKGEK
AP01011	GLFGKLIKKFGRKAISYAVKKARGKH
AP01012	SWKSMAKKLKEYMEKLKQRA
AP01014	GLKDKFKSMGEKLKQYIQTWKAKF
AP01016	GFFGKMKEYFKKFGASFKRRFANLKKRL
AP01018	QAFQTFKPDWNKIRYDAMKMQTSLGQMKKRFNL

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01019	GETFDKLKEKLKTFYQKLVEKAEDLKGDLKAKLS
AP01131	MSWLNFLKYIAKYGKKAVSAAWKYKGKVLEWLNVGPTLEWVWQKLKKIAGL
AP01142	KPYCSCKWRCGIGEEEKGICHKFPIVTYVCCRRP
AP01148	IATQCRIRGGFCRVGSCRFPHIAIGKCATFISCCGRAY
AP01151	GTWDDIGQGIGRVAYWVGKALGNLSDVNQASRINRKKKH
AP01153	YLAFRCGRYSPCLDDGPNVNLYSCCSFYNCHKCLARLENCPKGLHYNAYLKVCDWPSKAGCT
AP01154	AIKLVQSPNGNFAASFVLDGTKWIFKSKYYDSSKGYWVGIYEVWDRK
AP01156	NKLAYNMGHYAGKATIFGLAAWALLA
AP01157	ERGSRGQRCGPGEVFNQCGSACPRVCGRPPAQACTLQCVSGCFCRRGYIRTQRGGCIPERQCHQR
AP01158	ALYKKFKKKLLKSLKRL
AP01160	QDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
AP01161	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRVCRTGRSRWRDVCRNFMRRYQSRVTQGLVAGET AQQICEDLRLCIPSTGPL
AP01167	LTTKLWSSWGYYLGKKARWNLKHPYVQF
AP01168	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL
AP01169	NRWGDTVLSAASGAGTGIKACKSFGPWGMAICGVGGAAIGGYFGYTHN
AP01171	YSGKDCLKDMGGYALAGAGSGALWGAPAGGVGALPGAFVGAHVGAIAGGFACMGGMIGNKFN
AP01172	KRGPNCVGNFLGGLFAGAAAGVPLGPAGIVGGANLGMVGGALTCL
AP01174	KVSGGEAVAAIGICATASAAIGGLAGATLVTPYCVGTWGLIRSH
AP01175	GMSGYIQGIPDFLKGYLHGISAANKHKKGRLGY
AP01176	TTPACFTIGLGVGALFSAKFC
AP01177	FNRGGYNFGKSVRHVVDAIGSVAGILKSIR
AP01178	GAWKNFWSSLRKGFYDGEAGRAIRR
AP01180	NPKVAHCASQIGRSTAWGAVSGA
AP01182	FTPSVSFSQNGGVVEAAAQRGYIYKKYPKGAKVPNKVKMLVNIRGKQTMRTCYLMSWTASSRTAKY YYYI
AP01183	ATYYGNGLYCNKEKCWVDWNQAKGEIGKIIVNGWVNHGPWAPRR
AP01185	ENDHRMPNNLNRPNNLSKGGAKCGAAIAGGLFGIPKGPLAWAAGLANVYSKCN
AP01186	KTYYGTNGVHCTKKSLWGKVRLKNVIPGTLCRKQSLPIKQDLKILLGWATGAFGKTFH
AP01187	MNFLKNGIAKWMTGAELQAYKKKYGCLPWEKISC
AP01188	MLAKIKAMIKKFPNPYTLAAKLTTYEINWYKQQYGRYPWERPVA
AP01189	APAGLVAKFGRPIVKKYYKQIMQFIGEGSAINKIIPWIARMWRT
AP01192	SDCNINSNTAADVILCFNQVGSCALCSPTLVGGPVP
AP01193	DIDITGCSACKYAAGQVCTIGCSAAGGFICGLLGITIPVAGLSCLGFVEIVCTVADEYSGCGDAVAKEA CNRAGLC
AP01196	GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGGALQTVGQGLIDHGPVNVFIPVLIGPSW NGSGSGYNSATSSSGSGS
AP01198	LSCDEGMLAVGGLGAVGGPWGAAVGVLVGAALYCF
AP01204	GKNGVFKTISHECHLNTWAFLATCCS
AP01205	STPVLASVAVSMELLPTASVLYSDVAGCFKYSAKHHC
AP01206	CTFTLPGGGGVCTLTSECIC
AP01213	EFRGSIVIQGTKEGKSRPSLDIDYKQRVYDKNGMTGDAYGGLNIRPGQPSRQHAGFEFGKEYKNGFIK GQSEVQRGPGGRLSPYFGINGGFRF
AP01215	FVPYNPPRPYQSKPFPSFPGHGPFNPKIQWPYPLPNPGH
AP01227	VGIGGGGGGGGGGSCGGQGGGCGGCSNGCSGGNGGSGGSGSHI
AP01229	GDVNWVDVGKTVATNGAGVIGGAFGAGLCGPVCAGAFAVGSSAAVAALYDAAGNSNSAKQKPEGL PPEAWNYAEGRMCNWSPNNLSDVCL
AP01230	DGNDGQAELIAIGSLAGTFISPGFGSIAGAYIGDKVHSWATTATVSPSMSPSGIGLSSQFGSGRGTSSASS SAGSGS
AP01231	GGAPATSANAAGAAAIVGALAGIPGGPLGVVVGAVSAGLTTGIGSTVGSGSASSSAGGGS
AP01232	MNLNGLPASTNVIDLRGKDMGTYIDANGACWAPDTPSIIMYPGGSGPSYSMSSSTSSANSGS
AP01233	EKKPPRPPQWAVGHFM

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01234	FSKYERQKDKRPYSERKNQYTGPQFLYPPERIPPQKVIKWNEEGLPIYEIPGEGGHAEPAAA
AP01235	FNKLKQGSSKRTCAKCFRKIMPSVHELDERRRGANRWAAGFRKCVSSICRY
AP01240	ALKAALLAILKIVRVIKK
AP01238	NPLIPAIYIGATVGPSVWAYLVALVGAAAVTAANIRRASSDNHSCAGNRGWCRSKCFRHEYVDTYYSA VCGRYFCCRSR
AP01237	GLLDLAKHVIGIASKL
AP01242	GLLSFLPKVIGVIGHLIHPPS
AP01243	KGAPCAKKPCCGPLGHYKVDCSTIPDYPCCSKYGFCGSGPQYCG
AP01260	IIGHLIKTALGMLGL
AP01261	IIEKLVNTALGLLSGL
AP01262	GLADFLNKAVGKVVDFVKS
AP01264	RIGVLLARLPKLFSLFKLMGKKV
AP01268	FLPVILPVIGKLLSGIL
AP01283	MRKEFHNVLSSGQLLADKRPARDYNRK
AP01291	AALKGCWTKSIPPKPCFGKR
AP01294	GLGGAKKNFIIAANKTAPQSVKKTFSCKLYNG
AP01296	FMPILSCSRFKRC
AP01299	GLFTLIKGAAKLIGKTVPKKQARLGMNLWLVKLPTNVKT
AP01300	ATAVDFGPHGLLPIRPIRIRPLCGKDKS
AP01304	GLLSGILGAGKHIVCGLSGPCQSLNRKSSDVEYHLAKC
AP01305	FLPPSPWKETFRTS
AP01306	TSRCYIGYRRKVVCS
AP01307	GCSRWIIGIHGQICRD
AP01308	GLLSGTSVRGSI
AP01315	ARLKKCFNKVTGYCRKKCKVGERYEIGCLSGKLCCAN
AP01316	NPANPLNLKKHHGVFCDVCKALVEGGEKVGDDDLDAWLDVNIGTLCWTMLLPLHHECEEELKKVKK ELKKDIENKDSPDKACKDVDLC
AP01319	DPVTYIRNGGICQYRCIGLRHKIGTCGSPFKCCK
AP01323	LPVNEAQCRQVGGYCGLRICNFPSRFLGLCTRNHPCCSRVWV
AP01324	GPDSCNHDRGLCRVGNCNPGEYLAKYCFEPVILCCKPLSPTPTKT
AP01325	QPFIPRPIDTCRLRNGICFPGICRRPYYWIGTCNNGIGSCCARGWRS
AP01326	SKGKKANKDVELARG
AP01331	IFGAILPLALGALKNLIK
AP01339	FLSFPTTKTYFPHFDLSHGSAQVKGHGAK
AP01340	DAECEICKFVIQQVEAFIESNHSQAEIQKELNKLCSSVPSITQTCLSIARMVPYIIKKLEEHNSPGQVCQG LHLCKSS
AP01343	TESYFVFSVGM
AP01347	FIITGLVRGLTKLF
AP01348	SLSRFLSFLKIVYPPAF
AP01350	FLSLLPSIVSGAVSLAKKL
AP01353	FWGHIWNAVKRVGANALHGAVTGALS
AP01354	GFWKKVGSAAWGGVKAAAKGAAVGGLNALAKHIQ
AP01355	RESPSSRMECYEQAERYGYGGYGGGRYGGGYGSGRGQPVGQGVERSHDDNRNQPR
AP01356	KTCMTKKEGWGRCLIDTTCAHSCRKYGYMGGKCQGITRRCYCLLNC
AP01365	AAKPMGITCDLLSLWKVGHAACAAHCLVLGDVGGYCTKEGLCVCKE
AP01367	VTCNIGEWVCVAHCNSKSKKSGYCSRGVCYCTN
AP01372	SKCKCSRKGPKIRYSDVKKLEMKPKYPHCEEKMVIITTKSVSRYRGQEHCLHPKLQSTKRFIKWYNA WNEKRRVYEE
AP01374	NLAKGKEESLDSDLYAELRCMCIKTTSGIHPKNIQSLEVIGKGTHCNQVEVIATLKDGRKICLDPDAPRI KKIVQKKLAGDES
AP01376	YENPYGCPTDEGKCFDRCNDSEFEGGYCGGSYRATCVCYRT
AP01377	FFRHLFRGAKAIFRGARQGWRAHKVVSRYRNRDVPETDNNQEEP

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01378	AREASKSLIGTASCTCRRAWICRWGERHSGKCIDQKGSTYRLCCRR
AP01379	ILENLLARSTNEDREGSIFDTGPIRRPKPRPRPRPEG
AP01380	YDLSKNCRLRGGICYIGKCPRRFFRSGSCSRGNVCCLRFG
AP01381	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
AP01382	EKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
AP01400	RPKHPIKHQGLPQEVLNENLLRF
AP01407	HNKQEGRDHDKSKGHFHRVVIHHKGGKAH
AP01434	FFGSVLKLIPKIL
AP01447	FLPGLIAGIAKML
AP01449	FLGAIAAALPHVINAVTNAL
AP01454	IIPLPLGYFAKKT
AP01455	FFPLALLCKVFKKC
AP01456	VGKTWIKVIRGIGKSKIKWQ
AP01457	GLKDIFKAGLGSLVKGIAAHVAN
AP01465	VNWKKVLGKIIKVAK
AP01471	RPKPQQFFGLM
AP01473	LYENKPRRPYIL
AP01476	ACDTATCVTHRLAGLLSRSGGVVKNNFVPTNVGSKAF
AP01477	HSDAVFTDNYTRLRKQMAVKKYLNSILN
AP01479	YRQSMNNFQGLRSFGCRFGTCTVQKLAHQIYQFTDKDKDNVAPRSKISPQGY
AP01491	QWGYGGYGRGYGGYGGYGRGYGGYGGYGRGYGGYGRGMYGGYGRPYGGYGWGK
AP01496	IPPFIKKVLTTVF
AP01509	FLPKMSTKLRVPYRRGTKDYH
AP01510	GILKKFMLHRGTKVYKMRTLSKRSH
AP01511	TITLSTCAILSKPLGNNGYLCTVTKECMPSSCN
AP01513	GKNPTLQCMGNRGFCRPSCKKGEQAYFYCRTYQICCLQSHVRISLTGVEDNTNWSYEKHWPRIP
AP01514	GVNMYIRQIYDTCWKLKGHCRNVCGKKEIFHIFCGTQFLCCIERKEMPVLFVK
AP01515	AACSDRAHGHICESFKSFCKDSGRNGVKLRANCKKTCGLC
AP01516	LNLKGIFKKVASLLT
AP01517	INLLKIAKGIIKSL
AP01520	SSFSPPRGPPGWGPPCVQQPCPKCPYDDYKCPTCDKFPECEECPHISIGCECGYFSCECPKPVCEPCE SPIAELIKKGGYKG
AP01521	RFRLPFRRPPIRIHPPPFYPPFRRFL
AP01522	TYMPVEEGEYIVNISYADQPKKNSPFTAKKQPGPKVDLSGVKAYGPG
AP01523	AVDFSSCARMDVPGLSKVAQGLCISSCKFQNCGTGHCEKRGGRPTCVCDRCGRGGGEWPSVPMPKG RSSRGRRHS
AP01529	GAARKSIRLHRLYTWKATIYTR
AP01530	GSCSCSGTISPYGLRTCRATKTKPSHPTTKETHPQTLPT
AP01531	GKWGWIYITILFADVGGFKSSRHPEERRVQERRFKRITRGPD
AP01533	KRRGSVTTRYQFLMIHLLRPKKLFA
AP01540	AGANDLCQECEDIVHLLTKMTKEDAFQDTIRKFLEQECDILPLKLLVPRCRQVLDVYLPLVIDYFQGQI KPKAICSHVGLC
AP01543	KPWRFRRAIRRVRWRKVAPYIPFVVKTVGKK
AP01545	FFGHLFKLATKIIPSLFQ
AP01548	ADTLACRQSHQSCSFVACRAPSVDIGTCRGGKLKCCKWAPSS
AP01549	VLLFLFQAAPGSADAPFADTAACRSQGNFCRAGACPPTFAASGSCHGGLLNCCAK
AP01550	SAVGRHGRRFGLRKHRKH
AP01552	QIVDCWETWSRCTKWSQGGTGTLWKSCNDRCKELGRKRGQCEEKPSRCPLSKKAWTCICY
AP01555	TCRYWCKTPENQTYCCEDEREIPSKVGLKPGKCPPVRPVCPPTRGFFEPPKTCSNDGSCYGADKCCF DRCLGEHVCKPIQTRG

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01556	GCFEDWSRCSPSTSRGTGVLWRDCDSYCKVCFKADRGECFDSPSLNCPQRLPNNKQCRCINARTAKD NRNPTCWA
AP01557	DHHHDHGHDDHEHEELTLEKIKEKIKDYADKTPVDQLTERVQAGRDYLLGKGARPSHLPARVDRHLS KLTAAEKQELADYLLTFLH
AP01559	ATYDGKCYKKDNICKYKAQSGKTAICKCYVKVCPRDGAKCEFDSYKGKCYC
AP01565	DDMTMKPTPPPQYPLNLQGGGGGGSGDGFGFAVQGHQKVWTSDNGRHEIGLNGGYGQHLGGPYGN SEPSWKVGSTYTYRFPNF
AP01566	QRPYTQPLIYYPPPPTPPRIYRA
AP01569	SNDSLWYGVGQEMGKQANCITNHPVKHMIIPGYCSKILG
AP01570	GNAACVIGCIGSCVISEGIGSLVGTAFTLG
AP01571	IFGSIYHRKCVVKNRCETVSGHKTCKDLTCCRAVIFRHERPEVCRPQT
AP01572	WNPFKKIANRNCYPKTTCETAGGKKTCKDFSCCQIVLFGKKTRAKCTVVTS
AP01573	GWFKKTFHKVSHAVKSGIHAGQRGCSALGF
AP01575	NLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGE GCDRDGKYGFYTHVFRLKKWIQKVIDQFGE
AP01576	RVPPYLGRDCKHWCRDNNQALYCCGPPGITYPPFIRKHPGKCPSVRSTCTGVRSSRPKFCPHDDACE FRSKCCYDACVKHHVCKTVEFY
AP01577	FLLFPLMCKIQGKC
AP01578	GIHDILKYGKPS
AP01579	FVLPLVMCKILRKC
AP01580	AQEPVKGPVSTKPGSCPIILIRCAMLNPPNRCLKDTDCPGIKKCCEGSCGMACFVPQ
AP01583	GWANTLKNVAGGLCKITGAA
AP01585	KSCCRSTQARNIYNAPRFAGGSRPLCALGSGCKIVDDKKTPPND
AP01589	KDRPKKPGLCPPRPQKPCVKECKNDDSCPGQQKCCNYGCKDECRDPIFVG
AP01592	GIRNTVCFMQRGHCRLFMCRSGERKGDICSDPWNRCCVSSSIKNR
AP01593	CKQSCSFGPFTFVCDGNTK
AP01599	WNPFRKLYRKECNDVTSCDTVSGVKTCTKKNCCHRKFFGKTILKAPECTVIS
AP01600	RARAPHKAWYNCMTDAGISGAIAGAVAGCAATIEIGCVEGAIAGIGPSGIASMIAALWTCRSKY
AP01601	YVPPVQKPHPNGPKFPTFP
AP01602	LVLKYCPKIGYCSNTCSKTQIWATSHGCKMYCCLPASWKWK
AP01603	SSSGWLCTLTIECGTIICACR
AP01604	DAPGHPGKHYLQVNVPSDVRTIGVAGGGVQQCFRVTPGAWNDTRALVSNGAQVEVWGYTVADCAN RTTANQKYYDKAAAPSDSSTYFWFTLKNLRV
AP01606	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR
AP01607	ADRGWIKTLTKDCPNVISSICAGTIITACKNCA
AP01609	KCKWWNISCDLGNNGHVCTLSHECQVSCN
AP01610	CSTNTFSLSDYWGNKGNWCTATHECMSWCK
AP01613	LPRDTSRCVGYHGYCIRSKVCPKPFAAFGTCSWRQKTCCVDTTSDFHTCQDKGGHCVSPKIRCLEEQ LGLCPLKRWTCCKEI
AP01614	WRSLGRTLLRLSHALKPLARRSGW
AP01615	SASVLKTSIKVSKKYCKGVTLTCGCNITGGK
AP01616	SLGPAIKATRQVCPKATRFVTVSCKKSDCQ
AP01618	STPACAIGVVGITVAVTGISTACTSRCINK
AP01619	AANFGPSVFTPEVHETWQKFLNVVVAALGKQYH
AP01622	GLGSLLGKAFKIGLKTVGKMMGGAPREQ
AP01623	GFKLKGMARISCLPNGQWSNFPPKCIRECAMVSS
AP01624	HAEHKVKIGVEQKYGQFPQGTEVTYTCSGNYFLM
AP01625	LQDAALGWGRRCPQCPRCPSCPSCPRCPRCPRCKCNPK
AP01632	ATPATPTVAQFVIQGSTICLVC
AP01633	RRWVRRVRRWVRRVVRVVRRWVRR
AP01637	INWKKIASIGKEVLKAL

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01642	LCLDQKPEMEPFRKDAQQALEPSRQRRWLHRRCLSGRGFCRAICSIFEEPVRGNIDCYFGYNCCRRMF SHYRTS
AP01646	MTPLWRIMGTKPHGAYCQNHYECSTGICRKGHCSYSQPINS
AP01647	RCTCTTIISSSSTF
AP01648	GKLNLFLSRLEILKLFVGAL
AP01650	YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG
AP01651	LVATGMAAGVAKTIVNAVSAGMDIATALSLFSGAFTAAGGIMALIKKYAQKKLWKQLIAA
AP01652	LIDHLGAPRWAVDTILGAIAVGNLASWVLALVPGPGWAVKAGLATAAAIVKHQGKAAAAAW
AP01657	VIPFVASVAAEMMHHVYCAASKRC
AP01658	NALSMPRNKCNRALMCFG
AP01663	RRTCRCRFGRCFRRESYSGSCNINGRIFSLCCR
AP01676	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
AP01693	KFCEKPSGTWSGVCGNSGACKDQCIRLEGAKHGSCNYKPPAHRCICYYEC
AP01696	SPAGCRFCCGCCPNMRGCGVCCRF
AP01701	MKTFSVAVAVAIVLAFICTQESSALPVTGVEELVELVSSDDPVADHQELPVELGERLFNIRKKRASPKC TPYCYPTRDGVFCGVRCDF
AP01713	SRSGRGSGKGGRGGSRGSSGSRGSKGPSGSRGSSGSRGSKGSRGGRSGRGSTIAGNGNRNNGGTRTA
AP01715	PSCVCSGFETSGIHFC
AP01718	FKVQNQHGQVVKIFHH
AP01724	GTPGFQTPDARVISRFGFN
AP01729	GSQLVYREWVGHSNVIKGPP
AP01745	ERILDLRKTKKSCKNGEVLGCVSGHGPPGCSENECGMGPRPKACFFDCHYGCWCTGKLYRRKRDRK CVPKHECLL
AP01746	FLGGILNTITGLL
AP01747	SFPFFPPGICKRLKRC
AP01749	LVQRGRFGRFLKKVRRFIPKVIIAAQIGSRFG
AP01752	VTCELLMFGGVVGDSACAANCLSMGKAGGSCNGGLCDCRKTTFKELWDKRFG
AP01753	GIWSSIKNLASKAWNSDIGQSLRNKAAGAINKFVADKIGVTPSQAASMTLDEIVDAMYYD
AP01754	GGYYCPFFQDKCHRHCRSFGRKAGYCGGFLKKTCICV
AP01756	PDPGQPWQVKAGRPPCYSIPCRKHDECRVGSCSRCNNGLWGDRTCR
AP01757	SPRVSRRYGRPFGGRPFVGGQFGGRPGCVCIRSPCPCANYG
AP01758	IPAMEPAARVKRSPGYGGCSPRWACGGYG
AP01762	SPPNQPSIMTFDYAKTNK
AP01763	SPPSEQLGKSFNF
AP01765	APPPGYAMESDSFS
AP01766	FPPPGESAVDMSFFYALSNP
AP01769	KSLRPRCWIKIKFRCKSLKF
AP01778	DSMGAVKLAKLLIDKMKCEVTKAC
AP01783	FLPGVLRLVTKVGPAVVCAITRNC
AP01786	GKLQAFLAKMKEIAAQTL
AP01789	HSHACTSYWCGKFCGTASCTHYLCRVLHPGKMCACVHCSR
AP01793	GWINEEKIQKKIDEKIGNNILGGMAKAVVHKLAKGEFQCVANIDTMGNCETHCQKTSGEKGFCHGTK CKCGKPLSY
AP01794	FVDLKKIANIINSIFGK
AP01795	QIINNPITCMTNGAICWGPCPTAFRQIGNCGHFKVRCCKIR
AP01796	ASFPWSCPSLSGVCRKVCLPTELFFGPLGCGKGFLCGVSHFL
AP01797	GLWNSIKIAGKKLFVNVLDKIRCKVAGGCKTSPDVE
AP01798	SPRPDDKKNQGSASVDVQNERGEGTKVDARVRQELWRSDDGRTRAQAYGHWDRTYGGRNHGERSY GGGMRIEHTWGN
AP01799	KRGFGKKLRKRLKKFRNSIKKRLKNFNVVIPIPLPG
AP01800	KRGLWESLKRKATKLGDDIRNTLRNFKIKFPVPRQG

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01801	RTKRRIKLIKNGVKKVKDILKNNNIIILPGSNEK
AP01802	RPWAGNGSVHRYTVLSPRLKTQ
AP01804	GIRCPKSWKCKAFKQRVLKRLLAMLRQHAF
AP01815	DFGCARGMIFVCMRRCARMYPGSTGYCQGFRCMCDTMIPIRRPPFIMG
AP01824	FLPKLFAKITKKNMAHIR
AP01849	TSRCIFYRRKKCS
AP01852	SVMGTVKDLLIGAGKSAAQSVLKSLSCKISNDC
AP01860	ATNIPFKVHFRCKAAFC
AP01874	GLFSKFAGKGIVNFLIEGVE
AP01886	VVKCSYRLGSPDSQCN
AP01891	RFIYMKGFGKPRFGKR
AP01892	IPWKLPATFRPVERPFSKPFCRKD
AP01893	AAPRGGKGFFCKLFKDC
AP01899	FLKPLFNAALKLLP
AP01900	FLPVLAGVLSRA
AP00030	QRFSQPTFKLPQGRLTLSRKF
AP01919	FTLKKSQLLLFFLGTINFSLCEEERNAEEERRDYPEEKDVEVEKR
AP01922	GMWSKILGHLIR
AP01923	GKWMSLLKHILK
AP01939	CVISAGWNHKIRCKLTGNC
AP01940	FKTWKRPPFQTSCWGIIKE
AP01941	CVHWQTNTARTSCIGP
AP01943	SLWETIKNAGKGFIQNLDKIR
AP01952	FFPLLFGALSSHLPKLF
AP01953	FALGAVTKLLPSLLCMITRKC
AP01955	EYHLMNGANGYLTRVNGKTVYRVTKDPVSAVFGVISNCWGSAGAGFGPQH
AP01956	GFGMALKLLKKVL
AP01957	GTGLPMSERRKIMLMMR
AP01958	GLPRKILCAIAKKKGKCKGPLKLVCKC
AP01959	AILTTLANWARKFL
AP01963	ACQCPDAISGWTHTDYQCHGLENKMYRHVYAICMNGTQVYCRTEWGSSC
AP01964	IKLSPETKDNLKKVLKGAIKGAIAVAKMV
AP01965	LKIPGFVKDTLKKVAKGIFSAVAGAMTPS
AP01969	GPVGLLSSPGSLPPVGGAP
AP01971	VTSKSLCTPGCITGVLMCLTQNSCVSCNSCIRC
AP01972	STIVCVSLRICNWSLRFCPSFKVRCPM
AP01974	YGQSTHAVIYAQGYTYSSDWR
AP01975	KQIMTQFFNFARSPAVKD
AP01977	FLFSLIPSAISGLISAFKGRR
AP01978	FIGAIARLLSKIFGKR
AP01979	VAKCTEESGGKYFVFCCYKPTRICYMNEQKCESTCIGK
AP01981	GGKCTVDWGGQGGGRRLPSPLFCCYKPTRICYLNQETCETETCP
AP01982	ANKCIIDCMKVKTTCGDECKGAGFKTGGCALPPDIMKCCHNC
AP01993	TNWKKIGKCYAGTLGSAVLGFGAMGPVGYWAGAGVGYASFC
AP01995	ECELAKVDGGYTPKNCAMAVGGGMLSGAIRGGMSGTVFGVGTGNLAGAFAGAHIGLVAGGLACIGG YLGSH
AP01997	TPGGIDFISGGPHVAQDVLNAIKNFFK
AP02001	GMATKAGTALGKVAKAVIGAAL
AP02007	GLLGTLGNLLNGLGL
AP02011	GLFDVIKKVASVIGLASP

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP02012	VKVGINGFGRIGRLVTRAAFHGKKVEVVAIND
AP02013	FIGKLISAASGLLSHL
AP02014	GNANSNYEGGGSRSRNTGARNSLGRNAPTHIYSDPSTVKCANAVFSGMVGGAIKGGPVGMTRGTIGG AVIGQCLSGGGNGNGGGNRAGSSNCSGSNVGGTCSR
AP02021	FFPIVGKRLYGLL
AP02027	GFWGSLWEGVKSVV
AP02028	KRKCPKTPFDNTPGAWFAHLILGC
AP02029	DSIRDVSPTFNKIRRWFDGLFK
AP02030	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLV LRLR
AP02033	VSCGDVTSSIAPCLSYVMGRESSPSSSCCSGVRTLNGKASSSADRRTACSCLKNMASSFRNLNMGNAASI PSKCGVSVAFPISTSVDCSKIN
AP00012	GLFDIIKKIAESI
AP00013	GLFDIIKKIAESF
AP00014	GLLDIVKKVVGAFGSL
AP00015	GLFDIVKKVVGALGSL
AP00016	GLFDIVKKVVGAIGSL
AP00017	GLFDIVKKVVGTLAGL
AP00018	GLFDIVKKVVGAFGSL
AP00019	GLFDIAKKVIGVIGSL
AP00021	GLFDIVKKIAGHIASSI
AP00022	GLFDIVKKIAGHIVSSI
AP00058	GIGTKILGGVKTALKGALKELASTYAN
AP00060	GIGGKILSGLKTALKGAAKELASTYLH
AP00061	GIGGVLLSAGKAALKGLAKVLAEKYAN
AP00062	SIGAKILGGVKTFFKGALKELASTYLQ
AP00090	FLPLLAGLAANFLPTIICKISYKC
AP00128	KWKIFKKIEKVGRNIRNGIIKAGPAVAVLGEAKAL
AP00135	GWLKKIGKKIERVGQHTRDATIQTIGVAQQAANVAATLK
AP00139	KWKLFKKIEKVGQNIRDGIIKAGPAVAVVGQATQIAK
AP00165	ALWKNMLKGIGKLAGQAALGAVKTLVGAE
AP00176	ACYCRIPACIAGERRYGTCIYQGRLWAFCC
AP00178	DCYCRIPACIAGERRYGTCIYQGRLWAFCC
AP00200	LKLKSIVSWAKKVL
AP00214	KWCFRVCYRGICYRRCR
AP00242	GLLSVLGSVAQHVLPHVVPVIAEHL
AP00244	GLLSVLGSVVKHVIPHVVPVIAEHL
AP00245	GLFSVLGAVAKHVLPHVVPVIAEK
AP00246	GLFKVLGSVAKHLLPHVAPVIAEK
AP00247	GLFKVLGSVAKHLLPHVVPVIAEK
AP00248	GLFGVLGSIAKHVLPHVVPVIAEK
AP00260	GLFVGVLAKVAAHVVPAIAEHF
AP00345	GLLSVLGSVAKHVLPHVVPVIAEKL
AP00351	GLFDVIKKVASVIGGL
AP00352	GLFDIIKKVASVVGGL
AP00353	GLFDIIKKVASVIGGL
AP00366	GRFKRFRKKFKKLFKKLSPVIPLLHLG
AP00367	GGLRSLGRKILRAWKKYGPIIVPIIRIG
AP00396	RRRPRPPYLPRPRPPPFFPPRLPPRIPPGFPPRFPPRFP
AP00427	GLLGPLLKIAAKVGSNLL
AP00451	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00473	FFHHIFRGIVHVGKTIHRLVTG
AP00532	KTCENLADTFRGPCFATSNC
AP00541	IDWKKLLDAAKQIL
AP00565	FLIGMTQGLICLITRKC
AP00640	GLLGLLGSVVSHVVPAIVGHF
AP00770	GLLGLLGSVVSHVLPAITQHL
AP00809	GIKCRFCCGCCTPGICGVCCRF
AP00832	ILGPVISTIGGVLGGLLKNL
AP00878	FLPILASLAAKFGPKLFCLVTKKC
AP00964	GLWSKIKEAAKAAGKAALNAVTGLVNQGDQPS
AP01036	GIPCGESCVWIPCISSAIGCSCKSKVCYRN
AP01223	GFKDLLKGAAKALVKTVLF
AP01277	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK
AP01328	FIFHIIKGLFHAGKMIHGLVTRRRH
AP01412	FLPAIVGAAAKFLPKIFCAISKKC
AP01413	FLPIIAGVAAKVLPKIFCAISKKC
AP01414	FLPIIAGIAAKFLPKIFCTISKKC
AP01415	FLPVIAGVAANFLPKLFCAISKKC
AP01416	FLPIIAGAAAKVVQKIFCAISKKC
AP01417	FLPIIAGAAAKVVEKIFCAISKKC
AP01418	GLMDTIKGVAKTVAASWLDKLKCKITGC
AP01466	VNWKKILGKIIKVAK
AP01467	VNWKKILGKIIKVVK
AP01512	FFSLLPSLIGGLVSAIK
AP00494	KWKLFKKIPKFLHLAKKF
AP01774	GIPCGESCVFIPCITGAIGCSCKSKVCYRN
AP01777	GIPCGESCVFIPCITAAIGCSCKSKVCYRN
AP01989	IPCGESCVWIPCITAIAGCSCKNKVCYT
AP01990	AIPCGESCVWIPCISTVIGCSCSNKVCYR
AP01992	IPCGESCVWIPCISGMFGCSCKDKVCYS
AP00124	GLFLDTLKGAAKDVAGKLLEGLKCKIAGCKP
AP00151	RCVCTRGFCRCVCRRGVC
AP00160	ALWMTLLKKVLKAAAKALNAVLVGANA
AP00174	RRCICTTRTCRFPYRRLGTCIFQNRVYTFCC
AP00292	GIFSSRKCKTPSKTFKGYCTRDSNCDTSCRYEGYPAGD
AP00307	AGRGKQGGKVRAKAKTRSSRAGLQFPVGRVHRLLRKGNY
AP00336	AERVGAGAPVYL
AP00374	GKVWDWIKSTAKKLWNSEPVKELKNTALNAAKNLVAEKIGATPS
AP00385	FKLGSFLKKAWKSKLAKKLRAKGKEMLKDYAKGLLEGGSEEVPGQ
AP00550	KSCCRNTVARNCYNVCRIPGTPRPVCAATCDCKLITGTKCPPGYEK
AP00597	NFLGTLVNLAKKIL
AP00657	GIMDTVKNAAKDLAGQLDKLKCRITGC
AP00678	RLKELITTGGQKIGEKIRRIGQRIKDFFKNLQPREEKS
AP00755	ENFFKEIERAGQRIRDAIISAAPAVETLAQAQKIIKGGD
AP00781	FLGALIKGAIHGGRFIHGMIQNHH
AP00811	MTPFWRGVSLRPIGASCRDDSECITRLCRKRRCSLSVAQE
AP00731	SFGLCRLRRGFCARGRCRFPSIPIGRCSRFVQCCRRVW
AP00818	FLPLIGKILGTIL
AP00863	FLPIVGKLLSGLSGLL
AP00881	SLLGTVKDLLIGAGKSAAQSVLKGLSGKLSKDC

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00898	FLSGIVGMLGKLF
AP00900	FLSHIAGFLSNLF
AP01239	KFFRKLKKSVKKRAKEFFKKPRVIGVSIPF
AP01247	FMPIIGRLMSGSL
AP01248	INMKASAAVAKKLL
AP01259	ILPILGNLLNSLL
AP01286	RFIPPILRPPVRPPFRPPFRPPFRPPPIIRFFGG
AP01303	VIPFVASVAAEMMQHVYCAASKKC
AP01321	APGNKAECEREKGYCGFLKCSFPFVVSGKCSRFFFCCKNIW
AP01364	ATCDLLSAFGVGHAACAAHCIGHGYRGGYCNSKAVCTCRR
AP01439	FLPIIAGMAAKVICAITKKC
AP01468	GFGCPLNQGACHNHCRSIGRRGGYCAGIIKQTCTCYRK
AP01525	SWLSKTYKKLENSAKKRISEGIAIAIQGGPR
AP01544	IFGAIAGLLKNIF
AP01581	FLSLIPHIVSGVASIAKHF
AP01591	KCWNLRGSCREKCIKNEKLYIFCTSGKLCCLKPKFQPNMLQR
AP01596	CRQSCSFGPFTFVCDGNTK
AP01644	GAFGNFLKGVAKKAGLKILSIAQCKLSGTC
AP01721	FLPVLTGLTPSIVPKLVCLLTKKC
AP01750	RVRRFWPLVPVAINTVAAGINLYKAIRRK
AP01790	HPHVCTSYYCSKFCGTAGCTRYGCRNLHRGKLCFCLHCSR
AP01791	FLWGLIPGAISAVTSLIKK
AP01842	FIFPKKNIINSLFGR
AP01850	SFLSTFKELAINAAKNAGQSILHTLSCKLDKTC
AP01855	GLFSKFVGKGIKNFLIKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01877	GIFLKVLGVGKKVLCGVSGLC
AP01944	SLWETIKNAGKGFILNILDKIRCKVAGGCKT
AP00121	GLMDTVKNVAKNLAGHMLDKLKCKITGC
AP00142	GLKKLLGKLLKKLGKLLLK
AP00212	RRWCFRVCYKGFCYRKCR
AP00274	GIPCGESCVWIPCISAALGCSCKNKVCYRN
AP00277	VFQFLGRIIHHVGNFVHGFSHVF
AP00384	LLKELWTKIKGAGKAVLGKIKGLL
AP00445	GFCRCLCRRGVCRCICTR
AP00497	ILGPVLGLVSDTLDDVLGIL
AP00505	DSHAKRHHGYKRKFHEKHHSHRGY
AP00663	GFSSIFRGVAKFASKGLGKDLARLGVNLVACKISKQC
AP00706	GLLASLGKVFGGYLAEKLKPK
AP00724	RCLCRRGVCRCLCRRGVC
AP00725	RCICTRGFCRCICTRGFC
AP01269	GFLSILKKVLPKVMAHMK
AP01271	GIFPKIIGKGIVNGIKSLAKGVGMKVFKAGLNNIGNTGCNNRDEC
AP01273	GLFPKFNKKKVKTGIFDIIKTVGKEAGMDVLRTGIDVIGCKIKGEC
AP00074	FLPVLAGIAAKVVPALFCKITKKC
AP00094	FLPLIGRVLSGIL
AP00173	RCICTTRTCRFPYRRLGTCLFQNRVYTFCC
AP00180	ATCYCRTGRCATRESLSGVCEISGRLYRLCCR
AP00187	VVCACRRALCLPRERRAGFCRIRGRIHPLCCRR
AP00188	VVCACRRALCLPLERRAGFCRIRGRIHPLCCRR
AP00218	RGGRLCYCRRRFCICV

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00219	RGGGLCYCRRRFCVCVGR
AP00220	RGGRLCYCRGWICFCVGR
AP00221	RGGRLCYCRPRFCVCVGR
AP00222	VTCYCRRTRCGFRERLSGACGYRGRIYRLCCR
AP00223	VTCYCRSTRCGFRERLSGACGYRGRIYRLCCR
AP00224	CSCRTSSCRFGERLSGACRLNGRIYRLCC
AP00225	ACYCRIGACVSGERLTGACGLNGRIYRLCCR
AP00333	SCASRCKGHCRARRCGYYVSVLYRGRCYCKCLRC
AP00474	FIHHIFRGIVHAGRSIGRFLTG
AP00846	KYYGNGVSCNKKGCSVDWGKAIGIIGNNSAANLATGGAAGWKS
AP00023	AACARFIDDFCDTLTPNIYRPRDNGQRCYAVNGHRCDFTVFNTNNGGNPIRASTPNCKTVLRTAANR CPTGGRGKINPNAPFLFAIDPNDGDCSTNF
AP00025	HGVSGHGQHGVHG
AP00029	KWKLFKKIKFLHSAKKF
AP00506	KLAKLAKKLAKLAK
AP01026	GLPVCGETCVGGTCNTPGCSCSWPVCTRN
AP01031	GVPICGETCTLGTCYTAGCSCSWPVCTRN
AP01121	GIPCAESCVWIPCTVTALIGCGCSNKVCYN
AP01123	GTFPCGESCVFIPCLTSAIGCSCKSKVCYKN
AP01124	GLLPCAESCVYIPCLTTVIGCSCKSKVCYKN
AP01278	KSCCPNTTGRNIYNTCRFGGGSREVCARISGCKIISASTCPSDYPK
AP01279	KSCCPNTTGRNIYNTCRLTGSSRETCAKLSGCKIISASTCPSNYPK
AP01280	KSCCPNTTGRNIYNTCRFAGGSRERCAKLSGCKIISASTCPSDYPK
AP01281	KSCCPNTTGRNIYNTCRFGGGSRQVCASLSGCKIISASTCPSDYPK
AP01282	KSCCPNTTGRNIYNTCRLGGGSRERCASLSGCKIISASTCPSDYPK
AP01284	KSCCKNTTGRNIYNTCRFAGGSRERCAKLSGCKIISASTCPSDYPK
AP01775	GEFLKCGESCVQGECYTPGCSCDWPICKKN
AP01776	GLPTCGETCTLGTCYVPDCSCSWPICMKN
AP01784	GLPVCGETCAGGTCNTPGCSCSWPICTRN
AP01785	GLPVCGETCFGGTCNTPGCTCDPWPVCTRN
AP01806	GLPVCGETCVGGTCNTPGCACSWPVCTRN
AP01807	GLPVCGETCVGGTCNTPGCGCSWPVCTRN
AP01808	GSIPCGESCVFIPCISSVIGCACKSKVCYKN
AP01809	GIPCGESCVFIPCISSVIGCSCSSKVCYRN
AP01810	GSIPCGESCVFIPCISAVIGCSCSNKVCYKN
AP01811	GSIPCGESCVFIPCISAIIGCSCSSKVCYKN
AP01812	GSIPCEGSCVFIPCISAIIGCSCSNKVCYKN
AP01813	GIPCGESCVFIPCLTSAIDCSCKSKVCYRN
AP01983	GIACGESCVFLGCFIPGCSCKSKVCYFN
AP01984	GVIPCGESCVFIPCISSVLGCSCKNKVCYRD
AP01985	KLCGETCFKFKCYTPGCSCSYPFCK
AP01986	GDACGETCFTGICFTAGCSCNPWPTCTRN
AP01987	GIPCAESCVWIPPCTITALMGCSCKNNVCYNN
AP01988	GASCGETCFTGICFTAGCSCNPWPTCTRN
AP01991	GEYCGESCYLIPCFTPGCYCVSRQCVNKN
AP00003	DGVKLCDVPSGTWSGHCGSSSKCSQQCKDREHFAYGGACHYQFPSVKCFCKRQC
AP00031	DKLIGSCVWGAVNYTSDCNGECKRRGYKGGHCGSFANVNCWCET
AP00103	RECKAQGRHGTCFRDANCVQVCEKQAGWSHGDCRAQFKCKCIFEC
AP00148	ATYNGKCYKKDNICKYKAQSGKTAICKCYVKKCPRDGAKCEFDSYKGKCYC
AP00181	AFTCHCRRSCYSTEYSYGTCTVMGINHRFCCL

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP00190	HPLKQYWWRPSI
AP00286	QKLCERPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
AP00404	YGPGDGHGGGHGGGHGGGHGNGQGGGHGHGPGGGFGGGHGGGHGGGGRGGGGSGGGGSPGHGA GGGYPGGHGGGHHGGYQTHGY
AP00446	GADFQECMKEHSQKQHQHQG
AP00495	EQCGRQAGGKLCPNNLCCSQYGWCGSSDDYCSPSKNCQSNCKGGG
AP00646	FFPNVASVPGQVLLKKIFCAISKKC
AP00672	DCLSGRYKGPCAVWDNETCRRVCKEEGRSSGHCSPSLKCWCEGC
AP00716	RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC
AP00833	CLAGRLDKQCTCRRSQPSRRSGHEVGRPSPHCGPSRQCGCHMD
AP00889	APPGARPPPGPPPPGPPPPGP
AP00890	IDWKKVDWKKVSKKTCKVMLKACKFLG
AP00982	GTCKAECPTWEGICINKAPCVKCCKAQPEKFTDGHCSKILRRCLCTKPC
AP00983	QNNICKTTSKHFKGLCFADSKCRKVCIQEDKFEDGHCSKLQRKCLCTKNC
AP01049	GLPVCGETCFGGTCNTPGCSCTWPICTRD
AP01150	ELPKLPDDKVLIRSRSNCPKGKVWNGFDCKSPFAFS
AP01162	ETCASRCPRPCNAGLCCSIYGYCGSGAAYCGAGNCRCQCRG
AP01165	MINRTDCNENSYLEIHNNEGRDTLCFANAGTMPVAIYGVNWVESGNNVVTLQFQRNLSDPRLETITLQ KWGSWNPGHIHEILSIRIY
AP01166	AVRIGPCDQVCPRIVPERHECCRAHGRSGYAYCSGGGMYCN
AP01302	GFSPNLPGKGLRIS
AP01329	KQQLATEAESAGPIL
AP01371	GVTITVKPPFPGCVFYECIANCRSRGYKNGGYCTINGCQCLR
AP01478	RILSILRHQNLLKELQDLAL
AP01494	GHHPHGHHPHGHHPHGHHHPH
AP01528	RVCMKGSQHHSFPCISDRLCSNECVKEEGGWTAGYCHLRYCRCQKAC
AP01560	AKYTGKCTKSKNECKYKNDAGKDTFIKCPKFDNKKCTKDNNKCTVDTYNNAVDCD
AP01562	LSKFGGECSLKHNTCTYLKGGKNHVVNCGSAANKKCKSDRHHCEYDEHHKRVDCQTPV
AP01678	ITCQQVTSELGPCVPYLTGQGIP
AP01679	GRILSFIKGLAEHL
AP01680	ILGIITSLLKSLGKK
AP01681	KDLHTVVSAILQAL
AP01683	NEMGGPLVVEARTCESQSHKFKGTCLSDTNCANVCHSERFSGGKCRGFRRRCFCTTHC
AP01803	LMCTHPLDCSN
AP01970	EGPVGLADPDGPASAPLGAP
AP01976	VTCDVLSFEAKGIAVNHSACALHCIALRKKGGSCQNGVCVCRN
AP02017	GKVKVGVNGFGRIGRLVTRAAFNSGKVDIVA
AP02018	HTPTPTPICKSRSHEYKGRCIQDMDCNAACVKESESYTGGFCNGRPPFKQCFCTKPCKRERAAATLR WPGL
AP02024	RHRHCFSQSHKFVGACLRESNCENVCKTEGFPSGECKWHGIVSKCHCKRIC
AP02036	RQRDPQQQYEQCQKHCQRRETEPRHMQTCQQRCERRYEKEKRKQQKRYEEQQREDEEKYEERMK EEDN
AP02037	KRDPQQREYEDCRRRCEQQEPRQQHQCQLRCREQQRQHGRGGDMMNPQRGGSGRYEEGEEEQS
AP00028	CLGIGSCNDFAGCGYAVVCFW
AP00729	GLPVCGETCVGGTCNTPGCTCSWPVCTRN
AP00730	GSVLNCGETCLLGTCYTTGCTCNKYRVCTKD
AP01022	GVIPCGESCVFIPCISAAIGCSCKNKVCYRN
AP01023	GTACGESCYVLPCFTVGCTCTSSQCFKN
AP01024	GIPCGESCVFIPCLTTVAGCSCKNKVCYRN
AP01025	GFPCGESCVFIPCISAAIGCSCKNKVCYRN
AP01030	GLPICGETCVGGTCNTPGCSCSWPVCTRN

Table A.5: Xiao et al. (2013) Data Set Training AMP Sequences Continued...

Definition	Sequence
AP01034	GDPTFCGETCRVIPVCTYSAALGCTCDDRSDGLCKRN
AP01058	SISCGESCAMISFCFTEVIGCSCKNKVCYLN
AP01060	GIPCGESCVFIPCITSVAGCSCKSKVCYRN
AP01061	KIPCGESCVWIPCVTSIFNCKCKENKVCYHD
AP01062	KIPCGESCVWIPCLTSVFNCKCENKVCYHD
AP01063	KVCYRAIPCGESCVWIPCISAAIGCSCKN
AP01064	GIPCGESCVWIPCISAAIGCSCKSKVCYRN
AP01065	GSIPACGESCFKGKCYTPGCSCSKYPLCAKN
AP01066	GLPTCGETCFGGTCNTPGCTCDPWPVCTHN
AP01077	GGTIFDCGETCFLGTCYTPGCSCGNYGFCYGTN
AP01080	GVPCGESCVFIPCITGVIGCSCSSNVCYLN
AP01081	GIPCAESCVWIPCTVTALVGCSCSDKVCYN
AP01136	GGTIFDCGESCFLGTCYTKGCSCGEWKLCYGTN
AP01137	CLGVGSCNDFAGCGYAVVCFW
AP01138	CLGVGSCNDFAGCGYAIVCFW
AP01207	GICRCICGRGICRCICGR
AP01208	GICRCICGRRICRCICGR
AP01209	RICRCICGRRICRCICGR
AP01788	QEAQSVACTSYYCSKFCGSAGCSLYGCYLLHPGKICYCLHCSR

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences

Definition	Sequence
Q6YKB1	MFTNSIKNLIIYLMPLMVTLMLLSVSFVDAGKKPSGPNPGGNN
Q66125	MDVLTVVVSTADLHLANLQEVKRRRRRSHVRNRRARGYKSPSERARSIARLFQMLPFHGVDPVDWFP DVVRSPSVTSLVSYESFDDTDWFAGNEWAEGSF
Q8UYT3	MASIEIPLHEIIRKLERMNQKKQAQRKRHKLNRKERGHKSPSEQRRSELWHARQVELSAINSDNSSDEG TTLCRFDTFGSKSDAICDRSDWCLDQ
P80353	EPLCRRQFQQHQHLRACQRYLRRRAQRGEQRGPALRLCCNQLRQVNKPCVCPVLRQAAHQQLYQG QIEGPRQVRRLFRAARNLPNICKIPAVGRCQFTRW
O60516	MSTSTSCPIPGGRDQLPDCYSTTPGGTLYATTPGGTRIIYDRKFLLECKNSPIARTPPCCLPQIPGVTT PPTAPLSKLEELKEQETEEEIPDDAQFEMDI
P49172	MPIAQLYIIEGRTDEQKETLIRQVSEAMANSLDAPLERVRVLITEMPKNHFGIGGEPASKVRR
P14790	MLQSLIKKVWIPMKPYYTQAYQEIWVGTGLMAYIVYKIRSADKRSKALKASSAAPAHGHH
Q76ZQ4	MIGILLLIGICVAVTVAILYSMYNKIKNSQNPNPSPNLNSPPPEPKNTKFVNNLEKDHISSLYNLVKSSV
Q80HV6	MISNYEPLLLLVITCCVLLFNFTISSKTKIDIIFAVQTIVFIWFIFHFVHSAI
Q76ZQ3	MDMMLMIGNYFSGVLIAGIILLILSCIFAFIDFSKSTSPTRTWKVLSIMAFILGIIITVGMLIYSMWGKHCA PHRVSGVIHTNHSDISMN
P07608	MSWYEKYNIVLNPPKRCSFACADNLTTILAEDGNNIRAILYSQPKKLKILQDFLATSRNKMFLYKILDD EIRRVLT
P68596	MEDLNEANFSHLLINLSNNKDIDAQYASTLSVVHELLSAINFKIFNINKKSKKNSKSIEQHPVVHHAASAG REFNRR
P05623	MKTLALFLVLVCVLGLVQSWEWPWNRKPTKFPIPSPNPRDKWCRLNLGPAWGGRC
Q89VT6	MQAFNTDVRNRIIKLVKGILEQNALAADVTPQAKLVDVGLTSMDMVNLMLGVEAEFDFTIPQSEITPE NFQSVETLERMVMTQLQPATAA
A9CHM9	MNATIREILAKFGQLPTPVDTIADEADLYAAGLSSFASVQLMLGIEEAFDIEFPDNLLNRKSFASIKAIED TVKLILDGKEAA
P22702	MNGRPSVFTSQDYLSDHLWRALNA
P08521	MFSLSNSQYTCQDYISDHIWKTSSH
Q9P1F3	MNVDHEVNLLVEEIHRLGSKNADGKLSVKFGVLFRDDKCANLFEALVGTLKAAKRRKIVTYPGELLL QGVHDDVDIILLQD

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P08874	MFMKSTGIVRKVDELGRVVIPIELRRTLGIAEKDALEIYVDDEKIILKKYKPNMTCQVTGEVSDDNLKL AGGKLVLSKEGAEQIISEIQNQLQNLK
Q91J24	MGNLISTSCFNSKEKFRSQISDYSTWYPQPGQHISIRTFRELNPAPTSSPTSTRTETQLNGGNSRSTVEV LEEVNRQLTTHMPRR
P57752	MGLKEEFEEHAEKVNTLTELPSNEDLLILYGLYKQAKFGPVDTSRPGMFSMKERAKWDAWKAVEGK SSEEAMNDYITKVKQLLEVAASKAST
P45883	MSPQADFDKAAGDVKKLKTKPTDDELKELYGLYKQSTVGDINIECPGMLDLKGKAKWDAWNLKKG LSKEDAMSAYVSKAHELIEKYGL
A0R0B3	MAATQEEIIAGLAEIIEEVTGIEPSEVTPEKSFVDDLDIDSLSMVEIAVQTEDKYGVKIPDEDLAGLRTV GDVVAYIQKLEEENPEAAAALREKFAADQ
P80643	MADTLERVTKIIVDRLGVDEADVKLEASFKEDLGADSLDVVELVMELEDEFDMEISDEDAEKIATVGD AVNYIQNQQ
Q9ZCH9	MEFKIMSTTDKIEQKVIEMVAEKLNKDKAIITTDSRFIEDLKADSLDTVELMMAIEVEYGIDIPDDEATK IKTVSDVIKYIKERQS
P11830	MDRKEIFERIEQVLAEQLGIPAEQITEEADLREDLGMDSLDLVELVSALEDEVGMRVEQSQLEGIETVG HVMELTLDLVARLATASAADKPEAAS
O34163	MALIDEIKDVVANQLNISDKSKITDTASFVDDLNADSLDLVELIMELEKRYEIKIPQEDQEKIKNVADAA KYIEEHKK
Q3C258	MNQVMTIFLVLGVIVYSVESSSTPDGTWVKCRHDCFTKYKSCQMSDSCHDEQSCHQCHVKHTDCVNT GCP
Q3C256	MIYKAVFVCLVLVLLGDVFCSPRNSGGGTLNDNPFEKRTDCRFVGAKCTKANNPCVGKVCNGYQLY CPADDDHCIMKLTFIPG
P18281	MNPELQSAIGQGAALKHAETVDKSAPQIENVTVKKVDRSSFLEEVAKPHELKHAETVDKSGPAIPEDV HVKKVDRGAFLSEIEKAAKQ
P07032	MAGSEGLMSVDYEVSGRVQGVFFRKYTQSEAKRLGLVGWVRNTSHGTVQGQAQGPAARVRELQEW LRKIGSPQSRISRAEFTNEKEIAALEHTDFQIRK
O35031	MLQYRIIVDGRVQGVGFRYFVQMEADKRKLAGWVKNRDDGRVEILAEGPENALQSFVEAVKNGSPFS KVTDISVTESRSLEGHHRFSIVYS
Q83AB0	MTQKEKNETCIHVTVSGKVQGVFFRESVRKKAEELQLTGWVKNLSHGDVELVACGERDSIMILTEWL WEGPPQAAVSNVNWEEIVVEDYSDFRVR
P0AB65	MSKVCIIAWVYGRVQGVGFRYTTQYEAKRLGLTGYAKNLDDGSVEVVACGEEGQVEKLMQWLKSG GPRSARVERVLSEPHHPSGELTDFRIR
P15228	MKLFLLLVISASMLIDGLVNADGYIRGSNGCKVSCLLGNEGCNKECRAYGASYGYCWTWKLACWCQG LPDDKTWKSESNTCGGKK
P69943	GAPCLCANSGPNTRGNDLNGIVWVFGCPSGWHDCHGRAMVGYCCQED
P32781	MQLLRCFSIFSVIASVLAQELTTICEQIPSPTLESTPYSLSTTTILANGKAMQGVFEYYKSVTFVSNCGSH PSTTSKGSPINTQYVF
A7UNK4	ACWKANSCPGSAFESKDRLRSFALLYCRYNYKPPYGQGAFGYASAVSTHGWETEAQCINTFEQIITSC HGQSNGGTLELNSGRLSLAFGNCEEL
P17681	EEVFDDTDVGDELTNALESVLTDLKDKRDAEEPSAFMTRLRRQVAQMHIWRAVNHDRHHSTGSGRH SRFLTRNRYRYGGGHLSDA
P03086	MVLRQLSRKASVKVSKTWSGTKKRAQRILIFLLEFLLDFCTGEDSVDGKKRQRHSGLTEQTYSALPEP KAT
Q9LVC0	MEAMKMKLYVVVLVAVIAFSTVHQTVAAVDAPAPSPTSDASSFIPTFFASVAVMAFGFFF
Q9LYF6	MAISKASIVVLMMVIISVVASAQSEAPAPSPTSGSSAISASFVSAGVAAVAALVFGSALRI
O82337	MASRNSVTGFALFSFVFAVILSLAGAQSLAPAPAPTSDGTSIDQGIAYLLMVVALVLTYLIHPLDASSSYS FF
O55074	MGQLCCFPFAREEGKICEKDRKEPEDAELVRLSKRLVENAVLKAVQQYLEETQNKKQPGEGNSVKAE EGDRNGDGSDNNRK
P18829	MVQRCLVVALLVVVVAAALCSAQLNFTPNWGTGKRDAADFGDPYSFLYRLIQAEARKMSGCSN
P19872	MQVRAVLVLAVVALVAVATSRAQLNFTPWWGKRALGAPAAGDCVSASPQALLSILNAAQAEVQKLID CSRFTSEANS
P61855	MNPKSEVLIAAVLFMLLACVQCQLTFSPDWGKRSVGGAGPGTFFETQQGNCKTSNEMLLEIFRFVQS QAQLFLDCKHRE
P67788	MYKLTVFLMFIAFVIIAEAQLTFTSSWGGKRAMTNSISCRNDEAIAAIYKAIQNEAERFIMCQKN
P60848	MSSGTPTPSNVVLIGKKPVMNYVLAALTLLNQGVSEIVIKARGRAISKAVDTVEIVRNRFLPDKIEIKEI RVGSQVVTSQDGRQSRVSTIEIAIRKK
Q97ZF4	MTEKLNEIVVRKTKNVEDHVLDVIVLFNQGIDEVILKGTGREISKAVDVYNSLKDRLGDGVQLVNVQT GSEVRDRRRISYILLRLKRVY

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P81816	EYDDMYTEKRPKVYAFGL
P41839	DPLNEERRANRYGFGL
P81823	GYEDEDEDRPFYALGLGKRPRTYSFGL
P82153	AYSYVSEYKRLPVYNFGL
P85824	QVSLKYPEGKMYSFGL
O24172	DEAQFKECYDTCHKECSDKGNGFTFCEMKCDTDCSVKDVKEKLENYKPKN
P42559	QVRFRQCYFNPISCF
Q7M0J9	GFIGWGNNIFGHYSGDF
P09037	APGDRIYVHPF
P07457	MKSVILTGLLFVLLCVDHMTASQSVVATQLIPINTALTPAMMEGKVTNPIGIPFAEMSQIVGKQVNTPV AKGQTLMPNMVKTYVAGK
P80110	DLLEALSQDQKLLMAKFLPHIYAELANREGNWHEDAALRPLHDHDYPGWMDF
Q9UT86	MKVKILRYHAIANWTWDTPKDDVCGICRVPFDGCCPQCTSPGDNCPIVWGKCKHIFHAHCIQNWLA TSGSQGQCPMDRQTFVVADSTNEKSETQ
Q9BS18	MDSEVQRDGRILDLIDDAWREDKLPYEDVAIPLNELPEPEQDNGGTTESVKEQEMKWTDLALQYLHE NVPPIGN
Q9TUI9	MNLRRCVQALLLLWLCLSAVCGGPLLQTSDGKEMEEGTIRYLVQPRGPRSGPGPWQGGRRKFRRQR PRLSHKGPMPF
A9UGV5	MAGQLKSKIVAVAVAAVVVVASSLVGTASAADAPAPAPTSGATATAAAAPAFAAVSVAAAALGGYLF C
Q15847	MASKGLQDLKQQVEGTAQEAVSAAGAAAQQVVDQATEAGQKAMDQLAKTTQETIDKTANQASDTF SGIGKKFGLLK
P0CE37	MRLFLSLPVLVVVLSMVLEGPAPAQGAPEVSNPFDGLEELGKTLEDNTREFINRITQSELPAKMWDWF SETFRKVKEKLKIDS
P0DJG2	MKLLAATVLLLTICSLEGALVRRQAKEPCVESLVSQYFQTVTDYGKDLMEKVKSPELQAEAKSYFEKS KEQLTPLIKKAGTELVNFLSYFVELGTQPATQ
P02656	MQPRVLLVVALLALLASARASEAEDASLLSFMQGYMKHATKTAKDALSSVQESQVAQQARGWVTDG FSSLKDYWSTVKDKFSEFWDLDPEVRPTSAVAA
P02657	KSIFERDNRRDWLVIPDAVAAYVYETVNKMFPKVGQFLADAAQIPVIVGTRNFLIRETSKLSILAEQM MEKVKTLWNTKVLGYY
Q13794	MPGKKARKNAQPSPARAPAELEVECATQLRRFGDKLNFRQKLLNLISKLFCSGT
P36675	MSRYQHTKGQIKDNAIEALLHDPLFRQRVEKNKKGKGSYMRKGKHGNRGNWEASGKKVNHFFTTGL LLSGAC
P75993	MLEDTTIHNAITDKALASYFRSSGNLLEEESAVLGQAVTNLMLSGDNVNNKNIILSLIHSLETTSDILKAD VIRKTLEIVLRYTADDM
Q7M2N1	MSEPGDLSQTIVEEGGPEQETATPENGVIKSESLDEEEKLELQRRLVAQNQERRKSKSGAGKGKLTRS LAVCEESSARPGGESLQDQTL
Q589G4	MMLLVLSISAILQVSHSVSFCLLPIVPGPCTQYVIRYAFQPSISACRRFTFGGCEGNDNNFMTRRDCEHY CEELL
P32267	MNKNIDTVREIITVASILIKFSREDIVENRANFIAFLNEIGVTHEGRKLNQNSFRKIVSELTQEDKKTLIDE FNEGFEGVYRYLEMYTNK
P84702	YCGLFGDLCTLDGTLACCIALELECIPLNDFVGICL
Q9LVK3	MATESPNSVQKIVVHLRATGGAPILKQSKFKVSGSDKFANVIDFLRRQLHSDSLFVYVNSAFSPNPDES VIDLYNNFGFDGKLVVNYACSMAWG
Q5VTU8	MVAYWRQAGLSYIRYSQICAKVVRDALKTEFKANAKKTSGNSVKIVKVKKE
O00244	MPKHEFSVDMTCGGCAEAVSRVLNKLGGVKYDIDLPNKKVCIESEHSMDTLLATLKKTGKTVSYLGL E
P81450	MLKRFPTPILKVYWPFFVAGAAVYYGMSKAADLSSNTKEFINDPRNPRFAKGGKFVEVD
P81451	MGAAYHFMGKAIPPHQLAIGTLGLLGLLVVPNPFKSAKPKTVDIKTDNKDEEKFIENYLKKHSEKQDA
Q96253	MASNAAVPFWRAAGMTYISYSNICANIVRNCLKEPHKAEALTREKVHFSLSKWADGKPQKPVLRSDT PEV
P21306	MSAWRKAGISYAAYLNVAAQAIRSSLKTELQTASVLNRSQTDAFYTQYKNGTAASEPTPITK
Q00361	MVPPVQVSPLIKLGRYSALFLGMAYGAKRYNYLKPRAEEERRLAAEEKKKRDEQKRIERELAEAQED TILK
P13618	NKELDPVQKLFVDKIREYRTKRQTSGGPVDAGPEYQQDLDRELFKLKQMYGKADMNTFPNFTFEDP KFEAVEKPQS

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P03928	MPQLNTTVWPTMITPMLLTLFLITQLKMLNTNYHLPPSPKPMKMKNYNKPWEPKWTKICSLHSLPPQS
P61829	MQLVLAAKYIGAGISTIGLLGAGIGIAIVFAALINGVSRNPSIKDTVFPMAILGFALSEATGLFCLMVSFLLLFGV
P80286	MAELNVEIVSEERSIWSGAASAVSARTVNGEIGILPGHTPMLAVLGDGEVVVRTTDGGTVTAQAHGGFFSVDHDRVVIAATSARLGDAAAA
Q60189	MAECDDRPNLVERAVMKLGEPKNQARLLQVAWRISLLMMVIGFIIIIKTISPNFM
P81449	MSTVNVLRYSALGLGLFFGFRNDMILKCNAKKKEEQAQYEEKLKLVEEAKKEYAKLHPVVTPKDVPANASFNLEDPNIDFERVILNAVESLKEAST
P56134	MASVGECPAPVPVKDKKLLEVKLGELPSWILMRDFSPSGIFGAFQRGYYRYYNKYINVKKGSISGITMVLACYVLFSYSFSYKHLKHERLRKYH
P22483	MAFLGAAIAAGLAAVAGAIAVAIIVKATIEGTTRQPELRGTLQTLMFIGVPLAEAVPIIAIVISLLILF
P68699	MENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAIPMIAVGLGLYVMFAVA
Q8KRV3	MDMLFAKTVVLAASAVGAGTAMIAGIGPGVGQGYAAGKAVESVARQPEAKGDIISTMVLGQAVAESTGIYSLVIALILLYANPFVGLLG
Q2RFX4	MATIGFIGVGLAIGLAALGSGLGQGIASRGALEGMARQPEASGDIRTTLLLALAFMEALTLFSFVIAILMWTKL
O05331	MEGDIVQMGAYIGAGLACTGMGGAAVGVGHVVGNFISGALRNPSAAASQTATMFIGIAFAEALGIFSFLVALLLMFAV
P0A307	MNLTFLGLCIACMGVSVGEGLLMNGLFKSVARQPDMLSEFRSLMFLGVAFIEGTFFVTLVFSFIIK
O96910	MKVAIIILSLALVAAVFADQNCDIGNITSQCQMQHKNCEDANGCDTIIEECKTSMVERCQNQEFESAAGSTTLGPQ
Q7M460	NNKCDLEFASSECQMRYQDCGEASNCTALIEECKTSLQEECDQASSESSSTTIRPE
P38636	MAEIKHYQFNVVMTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVDVYTTLPYDFILEKIKKTGKEVRSGKQL
Q156A1	MQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
P22287	MKLSLLSVELALLIATTLPLCWAAALPVGLGVGLDYCNSSCTRAFDCLGQCGRCDFHKLQCVH
P0C1T6	MTLTMSTVVFFSLILLTLGLQPKDKDEGVMGRSRLGKRGLLMRSLDEELKSNDCPEYCPHGNECCEHHECRYDPWSRELKCLDSR
C1P605	MKLRKILKSMFNNYCKTFKDVPPGNMFR
Q6NTS2	MSSTSQKHRDFVAEPMGEKSVQCLAGIGEALGHRLEEKGFDKAYVVLGQFLVLKKDEELFKEWLKDICSANAKQSRDCYGCLKEWCDAFL
P11540	MKKAVINGEQIRSISDLHQTLKKELALPEYYGENLDALWDCLTGWVEYPLVLEWRQFEQSKQLTENGAESVLQVFREAKAEGCDITIILS
A8MTZ0	MLKAAAKRPELSGKNTISNNSDMAEVKSMFREVLPKQGPLFVEDIMTMVLCKPKLLPLKSLTLEKLEKMHQAAQNTIRQQEMAEKDQRQITH
P83346	RKCLIKYSQANESSKTCPSGQLLCLKKWEIGNPSGKEVKRGCVATCPKPWKNEIIQCCAKDKCNA
P11494	AAPCFCSGKPGRGDLWILRGTCPGGYGYTSNCYKWPNICCYPH
P0AE56	MYVCLCNGISDKKIRQAVRQFSPHSFQQLKKFIPVGNQCGKCVRAAREVMEDELMQLPEFKESA
Q9CA64	MSGRRSRSRQSSGTSRISEDQINDLIIKLQQLLPELRDSRRSDKVSAARVLQDTCNYIRNLHREVDDLSERLSELLANSDTAQAALIRSLLTQ
P0AB40	MKNVKTLIAAAILSSMSFASFAAVEVQSTPEGQQKVGTISANAGTNLGSLEEQLAQKADEMGAKSFRITSVTGPNTLHGTAVIYK
P84521	QQGEGGPYGGLSPLRFS
P85025	TPPAGPDVGPR
C0H419	MTVSIQMAGNLWKVHVKAGDQIEKGQEVAILESMKMEIPIVADRSGIVKEVKKKEGDFVNEGDVLLELSNSTQ
P62952	MYCLQWLLPVLLIPKPLNPALWFSHSMFMGFYLLSFLLERKPCTICALVFLAALFLICYSCWGNCFLYHCSDSPLPESAHDPGVVGT
Q8JFE6	MKCFAQIVVLLLVIAFSHGAVITGVCDRDAQCGSGTCCAASAFSRNIRFCVPLGNNGEECHPASHKVPYNGKRLSSLCPCNTGLTCSKSGEKFQCS
P0C739	MVHVLERALLEQQSSACGLPGSSTETRPSHPCPEDPDVSRLRLLLVVLCVLFGLLCLLLI
P24282	MEPIFIIGIILGLVILLFLSGSAAKPLKWIGITAVKFVAGALLLVCVNMFGGSLGIHVPINLVTTAISGILGIPGIAALVVIKQFII
P22499	MKKRYYTVKHGTLRALQEFADKHNVEVRREGGSKALRMYRPDGKWRTVVDFKTNSVPQGVRDRAFEEWEQIIIDNALLLNAD

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9H3K6	MELSAEYLREKLQRDLEAEHVEVEDTTLNRCSCSFRVLVVSAKFEGKPLLQRHRLVNACLAEELPHIH AFEQKTLTPDQWARERQK
P86994	MLLLSAVKTLLLAWLGIVLVFMSIIKSAMLDFLQEAGKLEGIETYKKEAQTSFMAPSWALGHLMGRK
P26814	MKKMLLATALALLITGCAQQTFTVQNKPAAVAPKETITHHFFVSGIGQKKTVDAAKICGGAENVVKT ETQQTFVNGLLGFITLGIYTPLEARVYCSQ
P85829	MVPVPVHHMADELLRNGPDTVI
P0C7J9	QQWPPGHHIPP
P0C7K1	QGPSPRHPIPP
P0C7K2	QGGAPWNPIPP
P0C7K3	QSAPGNEAIPP
P81754	ALFEESTVSAEPR
Q9TWD3	LRDYANRVINGGPVEAAGPPA
P0C7J7	QWPDPSSDIPP
P85038	QGRPPRPHIPP
P0C7S3	EEGGSPPPVVI
P85167	QARPPHPPIPP
P86284	QTLLQELPIPP
P85168	QGWAWPRPQIPP
P85169	QGGLPRPGPEIPP
P85163	QWAQWPRPTPQIPP
P0C7S8	QAPWPDTISPP
P84823	APVPGLSPFRVV
Q94JY4	MAKAGGITNAVNVGIAVQADWENREFISHISLNVRRLFEFLVQFESTTKSKLASLNEKLDLLERRLEML EVQVSTATANPSLFAT
Q8WUW1	MAGQEDPVQREIHQDWANREYIEIITSSIKKIADFLNSFDMSCRSRLATLNEKLTALERRIEYIEARVTK GETLT
Q0VTT9	MSFLKKSLFLVLFLGLVSSSICEEEKRETEEEENEDEIEEESEEKKREDPERPPGFTPFRVY
B0R5N8	MPITDLHCPRCGSDVKMGLPMGATVKSVTAASRQEPTSDTQKVRTVECRNDHEFFVRFEW
P59866	VTMGYIKDGDGKKIAKKKNKNGRKHVEIDLNKVG
C0H3U9	MHTCPRCDSKKGEVMSKSPVEGAWEVYQCQTCFFTWRSCEPESITNPEKYNPAFKIDPKETETAIEV PAVPERKA
P0A510	MAEDVRAEIVASVLEVVVNEGDQIDKGDVVVLLESMKMEIPVLAEAAGTVSKVAVSVGDVIQAGDLIA VIS
P80163	CSCADMTDKECLYFCHQDVIW
P81782	MECYRCGVSGCHLKITCSAEETFCYKWLNKISNERWLGCAKTCTEIDTWNVYNKCCTTNLCNT
P24649	MVYRRRRRSSTGATYGLTRRRRSSAGITRRRRSSGYRRRPGRPRTYRRSRSRSLTSRRSYRTRYY
P21808	MNRPVFLVLLLTGFLCIAAQEANVAHHYCGRHLANTLADLCWDTSVEKRSESSLASYSSRGWPWLPT PNFNKRAIKKRGVVDECCIQPCTLDVLATYC
P83660	SETGNTVTVKGFSPLR
Q8I6R2	MMSKLGVLVTICLLLFPLTALPLDGDQPADHPAKRTQDHNLASPISAWIDPSHYCCCGGGCTDDCVNC
Q9VNE7	MASTGEIWLQWFSCCFQQQRSPSRRPHQRLRIDRSMIGNPTNFVHTGHIGSADVELSANRLNAISTQM QSKGGYETNSIHSLHAC
Q9NRR8	MSEFWHKLGCCVVEKPQPKKKRRRIDRTMIGEPMNFVHLTHIGSGEMGAGDGLAMTGAVQEQMRS KGNRDRPWSNSRGL
P25938	AGDAAAGKTLYDASCASCHGMQAQGQGMFPKLAGLTSERIKTTLVAFKSGDTATLKKEGLGGPMSA IMAPNAAGLSEQDMDNLSAYIATLK
P00124	AVTKADVEQYDLANGKTVYDANCASCHAAGIMGAPKTGTARKWNSRLPQGLATMIEKSVAGYEGEY RGSKTFMPAKGGNPDLTDKQVGDAVAYMVNEVL
P86362	DGCPPHPVPGMHPCMCTNTC
P0C1W1	DGRCCHPACAKHFNC
P0CI05	GGCCSHPACQNNPDYC
P0C351	QSPGCCWNPACVKNRC
P85009	GCCSDPRCKHQC

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P60274	IRDECCSNPACRVNNPHVC
P85012	DYCCRRPPCTLIC
P50982	MFTVFLLVVLATTVGSFTLDRASDGRDAAANDKASDLIALTARRDPCCYHPTCNMSNPQICG
P0C8V4	GGCCSYPPCAVSNPQHC
P58782	GCCGPYPNAACHPCGCKVGRPPYCDRPSGG
P0C829	MGMRMMFTVFLSVVLATTVVSTPSDRASDGRNAAVHERQKELVPSVITTCCGYDPGTMCPPCRCTN SCPTKPKKPGRRND
P0C2C5	DCCGVKLEMCHPCLCDNSCKNYGK
P81783	MKTLLLTLVVVTIVCLDLGYTMKCKICNFDTCRAGELKVCASGEKYCFKESWREARGTRIERGCAAT CPKGSVYGLYVLCCTTDDCN
P82087	QQDYTGAHMDF
P82090	QQDYGTGWFDF
P60976	MKYFVVFCVLIIAVAAFTSAAEDGEVFEENPLEFPKTIQKRCISARYPCSNSKDCCSGNCGTFWTCYIR KDPCSKECLAP
P69165	CSNLSTCVLGKLSQDLHKLQTFPRTNTGAGVP
P82889	MDFNPSEVASQVTNYIQAIAAAGVGVLALAIGLSAAWKYAKRFLKG
P03619	MKKSVVAKIIAGSTLVIGSSAFAADDATSQAKAAFDSLTAQATEMSGYAWALVVLVVGATVGIKLFKK FVSRAS
P03620	MRVLSTVLAAKNKIALGAATMLVSAGSFAAEPNAATNYATEAMDSLKTQAIDLISQTWPVVTTVVVA GLVIRLFKKFSSKAV
P03621	MKAMKQRIAKFSPVASFRNLCIAGSVTAATSLPAFAGVIDTSAVESAITDGQGDMKAIGGYIVGALVIL AVAGLIYSMLRKA
P03623	MQSVITDVTGQLTAVQADITTIGGAIIVLAAVVLGIRWIKAQFF
P03622	SGVGDGVDVVSAIEGAAGPIAAIGGAVLTVMVGIKVYKWVRRAM
P57730	MADQLLRKKRRIFIHSVGAGTINALLDCLLEDEVISQEDMNKVRDENDTVMDKARVLIDLVTGKGPKS CCKFIKHLCEEDPQLASKMGLH
O28403	MLHLVVYDISDDGSRARLAKLLEKFGLQRVQYSAFRGELNPNDREVLARQVGKFVRDDRDCIFIIPLC QRCSSTAIVISNTGVELVKEKGVEFV
Q82W51	MLIIVTYDVSTETRAGRKRLRRVAKLCESIGQRVQKSVFECRINLMQYEELERRLLSEIDEQEDNLRLY RLTEPAELHVKEYGNFKAIDFEGPLTI
O27155	MVVTVYLLIVYDVGVERVNRVKSYLRTELHWVQNSVFEGEVTESQFRRIETNLERIIDRERDSVIIYSFR SERAMNRNVLGLEKSPLDVIL
Q9X2B6	MYVIMVYDVNEKRVAKILKIARKYLKWVQNSVLEGELSPGKYEKLKLEVSRLIDEKEDSVRFYVMDSQ KVFNLETLGVEKGEDGFIF
P84902	QTCVSCVNFGNGFCGDNCGNSWACSGC
P00948	MLFHVKMTVKLPVDMDPAKATQLKADEKELAQRLQREGTWRHLWRIAGHYANYSVFDVSSVEACN DTLMQLPLFPYMDIEVDGLCRHPSSIHSDDR
Q05595	MKKTLMLLAMVVALVILPFFINHGGEYGGSDGEAESQIQAIAPQYKPWFQPLYEPASGEIESLLFTLQG SLGAAVIFYILGYCKGKQRRDDRA
Q8T112	MFGYRSLLVLLVTLSLCLLLQSSHCSAVRTYGNDLDARARREIISLAARLIKLSMYGPEDDSFVKRNGG TADALYNLPDLEKIGKR
P58804	WATIDECEETCNVTFKTCCGPPGDWQCVEACPV
P0C8S5	TYGIYDAKPPFSCAGLRGGCVLPPNLRPKFKE
Q9Y2S6	MSGREGGKKKPLKQPKKQAKEMDEEDKAFKQKQKEEQKKLEELKAKAAGKGPLATGGIKKSGKK
P62552	MKQRITVTVDSDSYQLLKAYDVNISGLVSTTMQNEARRLRAERWKAENQEGMAEVARFIEMNGSFAD ENRDW
Q9TS44	AVQKVDGEPRAHLGALLARYIQQARKAPSGRMSVIKNLQNLDPSHRISDRDYMGWMDF
Q16627	MKISVAAIPFFLLITIALGTKTESSSRGPYHPSECCFTYTTYKIPRQRIMDYYETNSQCSKPGIVFITKRG HSVCTNPSDKWVQDYIKDMKEN
P10147	MQVSTAALAVLLCTMALCNQFSASLAADTPTACCFSYTSRQIPQNFIADYFETSSQCSKPGVIFLTKRS RQVCADPSEEWVQKYVSDLELSA
P0ABM5	MTPAFASWNEFFAMGGYAFFVWLAVVMTVIPLVVLVVHSVMQHRAILRGVAQQRAREARLRAAQQ QEAA
P72759	MQLAKVLGTVVSTSKTPNLTGVKLLLVQFLDTKGQPLERYEVAGDVVGAGLNEWVLVARGSAARKE RGNGDRPLDAMVVGIIDTVNVASGSLYNKRDDGR
Q99M08	MEVSQAASGTDGVRERRGSFEAGRRNQDEAPQSGMNGLPKHSYWLDLWLFILFDLALFVFVYLLP

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q5BLP8	MAPPPACRSPMSPPPPPLLLLLLSLALLGARARAEPAGSAVPAQSRPCVDCHAFEFMQRALQDLRKTA CSLDARTETLLLQAERRALCACWPAGH
P25063	MGRAMVARLGLGLLLLALLLPTQIYSSETTTGTSSNSSQSTSNTGLAPNPTNATTKAAGGALQSTASLF VVSLSLLHLYS
P31358	MKRFLFLLLTISLLVMVQIQTGLSGQNDTSQTSSPSASSNISGGIFLFFVANAIIHLFCFS
Q8NHZ8	MLRRKPTRLELKLDDIEEFENIRKDLETRKKQKEDVEVVGGSDGEGAIGLSSDPKSREQMINDRIGYKP QPKPNNRSSQFGSLEF
P0C1W6	MPKLAVVLLVLLILPLSYFDAAGGQAVQGDWRGNRLARDLQRGGRDDESECIINTRDSPWGRCCRTR MCGSMCCPRNGCTCVYHWRRGHGRSCPG
P62567	GLLDGLLGTLGL
Q6UWT4	MAVSVLRLTVVLGLLVLFLTCYADDKPDKPDDKPDDSGKDPKPDFPKFLSLLGTEIIENAVEFILRSMS RSTGFMEFDDNEGKHSSK
P0AE60	MKKPLRQQNRQIISYVPRTEPAPPEHAIKMDSFRDVWMLRGKYVAFVLMGESFLRSPAFTVPESAQR WANQIRQEGEVTE
Q5EE01	MALSTIVSQRKQIKRKAPRGFLKRVFKRKKPQLRLEKSGDLLVHLNCLLFVHRLAEESRTNACASKCR VINKEHVLAAAKVILKKSRG
A8MT69	MEGAGAGSGFRKELVSRLLHLHFKDDKTKVSGDALQLMVELLKVFVVEAAVRGVRQAQAEDALRVD VDQLEKVLPQLLLDF
P22790	SGQSWRPQGRF
Q06153	MQFSIATIALFLSSAMAAPYSGNSNSDSYDPCTGLLQKSPQCCNTDILGVANLDCHGPPSVPTSPSQFQ ASCVADGGRSARCCTLSLLGLALVCTDPVGI
Q96KF7	MSSAPEPPTFKKEPPKEKEFQSPGLRGVRTTTLFRAVNPELFIKPNKPVMAFGLVTLSLCVAYIGYLH AIQENKKDLYEAIDSEGHSYMRRKTSKWD
Q0VGL1	MTSALTQGLERIPDQLGYLVLSEGAVLASSGDLENDEQAASAISELVSTACGFRLHRGMNVPFKRLSVV FGEHTLLVTVSGQRVFVVKRQNRGREPIDV
P0C8W0	MSTLGILLPIALLLPLANPAENGDGQAMPRTRNLRSLSFGRTLRRLEKRGCDPTDGCQTTVCETDTGP CCCKPNFTCQISNSGTKSCSCSGQPSDCPV
P09931	RHKSSQESEESEELDQCCEQLNELNSQRCQCRALQQIYESQSEQCEGRQQEQQLEGELEKLPRICGFGP LRRCNINPDEE
P09930	FRSSEQSCKRQLQQVNLRHCENHIDQRIQQQQEEEED
P15472	ACSGRGSRCPPQCCMGLRCGRGNPQKCIGAHEDV
Q8N0T1	MAKNKLRGPKSRNVFHIASQKNFKAKNKAKPVTTNLKKINIMNEEKVNRVNKAFVNVQKELAHFAKS ISLEPLQKELIPQQRHESKPVNVDEATRLMALL
P15020	MAKVNIKPLEDKILVQANEAETTTASGLVIPDTAKEKPQEGTVVAVGPGRWDEDGEKRIPLDVAEGD TVIYSKYGGTEIKYNGEEYLILSARDVLAVVSK
Q1I5E1	MKLRPLHDRVVIRRSEEESKTAGGIVLPGSAAEKPNRGEVVAVGTGRILDNGEVRALAVKVGDKVVF GPYSGSNTVKVDGEDLLVMAENEILAVIEG
P0AE63	MPYKTKSDLPESVKHVLPSHAQDIYKEAFNSAWDQYKDKEDRRDDASREETAHKVAWAAVKHEYAK GDDDKWHKKS
Q9NYJ1	MSTSVPQGHTWTQRVKKDDEEEDPLDQLISRSGCAASHFAVQECMAQHQDWRQCQPQVQAFKDCM SEQQARRQEELQRRQEQAGAHH
P42718	ILGLLKGISALLS
P30814	RIFDTSCKGFYDRGLFAQLDRVCEDCYNLYRKPHVAAECRRDCYTTEVFESCLKDLMMHDFINEYKE MALMVS
P08365	MRITIKRWGNSAGMVIPNIVMKELNLQPGQSVEAQVSNNQLILTPISRRYSLDELLAQCDMNAAELSEQ DVWGKSTPAGDEIW
Q9BUW7	MSGPNGDLGMPVEAGAEGEEDGFGEAEYAAINSMLDQINSCLDHLEEKNDHLHARLQELLESNRQTR LEFQQQLGEAPSDASP
P0C607	MKLCATFLLVLVTLPLVTGEKSSERSLSGAILRGVRRTCSRRGHRCIRDSQCCGGMCCQGNRCFVAIR RCFHLPF
P84197	CQAYGESCSAVVRCCDPNAVCCQYPEDAVCVTRGYCRPPATVLT
Q7Z096	MKLCLTFLLVLMILASVTGEKSSKHTLSRAARVKNRGPSFCKADEKPCKYHADCCNCCLGGICKPSTS WIGCSTNVFLTR
Q9U3Z3	MMFRVTSVGCLLLVIVFLNLVVPTSACRAEGTYCENDSQCCLNECCWGGCGHPCRHPGKRSKLQEFF RQR
D2DGD3	MKLVLAIVVILMLLSLSTGAEMSDNHASRSATARRDRHLSPKAWPCGGVRASCSRHDDCCGSLCCFGT STGCRVAVRPCW
P69330	MKINQPAVAGTLESGDVMIRIAPLDTQDIDLQINSSVEKQFGDAIRTTILDVLARYNVRGVQLNVDDKG ALDCILRARLEALLARASGIPALPWEDCQ

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q0N4U8	MPSVRSVTCCCLLWMMFSVQLVTPGSPGTAQLSGHRTARFPRPRICNLACRAGIGHKYPFCHCRGKR DAVSSSMAV
P61165	MELEAMSRYTSPVNPAVFPHLTVVLLAIGMFFTAWFFVYEVTSTKYTRDIYKELLISLVASLFMGFGV LFLLLWVGIYV
Q6UW78	MDSLRKMLISVAMLGAGAGVGYALLVIVTPGERRKQEMLKEMPLQDPRSREEAARTQQLLLATLQEA ATTQENVAWRKNWMVGGEGGAGGRSP
Q9Z2N6	MSEILPYGEDKMGRFGADPEGSDLSFSCRLQDTNSFFAGNQAKRPPKLGQIGRAKRVVIEDDRIDDVL KGMGEKPPSGV
P07231	MHLYTYLYLLVPLVTFHLILGTGTLDDGGALTERRSADATALKAEPVLLQKSAARSTDDNGKDRLTQ MKRILKQRGNKARGEEELQENQELIREKSNGKR
P0C8E0	GEDEYAEGIREYQLIHGKI
P0C8E2	GEPEVAKWAEGLREKAASN
Q17868	MTTGNNDFYYSNKYEDDEFEYRHVHVTKDVSKLIPKNRLMSETEWRSLGIQQSPGWMHYMIHGPER HVLLFRRPLAATQKTGGNVRSGNAVGVR
P17684	GEEEYQKMLENLREAEVKKNA
Q3ECD6	MANLKFLLCLFLICVSLSRSSASRPMFPNADGIKRGRMMIEAEEVLKASMEKLMERGFNESMRLSPGG PDPRHH
O49519	MAKLSFTFCFLLFLLLSSIAAGSRPLEGARVGVKVRGLSPSIEATSPTVEDDQAAGSHGKSPERLSPGGP DPQHH
Q84W98	MATSNDQTNTKSSHSRTLLLLFIFLSLLLFSSLTIPMTRHQSTSMVAPFKRVLLESSVPASSTMDLRPKA STRRSRTSRRREFGNDAHEVPSGPNPISN
P0C6S2	GGVGRCIYNCMNSGGGLNFIQCKTMCY
D6C4I9	MNVTVMFLVLLLLTMPLTDGFNIRATNGGELFGPVQRDAGNVLDHGFQRRRECPPWCPTSHCNAGT C
P84707	WDAYDCIQFCMRPEMRHTYAQCLSICT
Q25088	MACKVKAELEAAFKKLDANGDGYVTALELQTFMVTLDAYKALSKDKVKEASAKLIKMADKNSDGKI SKEEFLNANAELLCQLK
Q9XF04	MDSKSFLLLLLLFCFLFLHDASDLTQAHAHVQGLSNRKMMMMKMESEWVGANGEAEKAKTKGLGLH EELRTVPSGPDPLHHHVNPPRQPRNNFQLP
P58623	CCKQSCTTCMPCCW
P58624	CCELPCHGCVPCCWP
Q9BPI0	MLKMGVVLFIFLVLFPLATLQLDADQPVERYAENKQLLNTDERREIILSALRTRVCCPFGGCHELCQC CEG
P0CH15	CCNWPCSFGCIPCCY
P58841	CCSQDCLVCIPCCPN
P01524	RDCCTPPRKCKDRRCKPMKCCA
P0C1N0	SKQCCHLAACRFGCTPCCW
P56529	MSKLGALLTICLLLFPITALLMDGDQPADRPAERMDYDISSEVHRLLERRHPPCCMYGRCRRYPGCSS ASCCQGG
P58926	CCKYGWTCLLGCSPCGC
P86263	GCCHPSTCHVRKGCSRCCS
P0C260	AFVKGSAQRVAHGY
G2TRJ8	MVDCQKEACNLQSCIQRNQYNQGNCEKFVNDLLLCCKRWYDKNSLTGNEAPHTCPELKPLLRQLSSR NLT
P85016	GCCPPQWCGPDCTSPCC
P85018	GCCPFPACTHTIICRCC
P85019	CCMALCSRYHCLPCC
P85020	GCCSPWNCIQLRACPCCPN
P84713	FHGGSWYRFPWGY
Q9NSA3	MNREGAPGKSPEEMYIQQKVRVLLMLRKMGSNLTASEEEFLRTYAGVVNSQLSQLPPHSIDQGAEDV VMAFSRSETEDRRQ
P04972	MNLEPPKAEIRSATRVMGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDITVI CPWEAFNHLELHELAQYGII
P56621	MADVEEWLTHARKVTQEASIGVDVTSIQECISAEPAQRVLVARRDAWRAICCAAFAALVAFAAINRVA TIMLEKPAPTWVATPSAASPFGLLIGK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P64467	MTVQDYLLKFRKISSLESLEKLYDHLNYTLTDDQELINMYRAADHRRAELVSGGRLFDLGQVPKSVW HYVQ
P0C8V5	MKLTCVVIVAVLFLTAWTFVMADDSRYGLKDLFPKARHEMKNPEASKLNKRDECFSPGTFCGIKPGL CCSAWCYSFFCLTLTF
P0C8V9	MKLTCVVIVAVLLLTACQLLTADDSRGTQKHRSLRSTTKVSKATDCIEAGNYCGPTVMKICCGFCSPF SKICMNYPQN
Q86RA3	MKLTCVLIIAVLFLTAYQLATAASHAKGKQKHRALRPADKHFRFTKRCNNRGGGCSQHPHCCSGTCN KTFGVCL
D2Y492	MKLTCVLIIAVLILTACQFIAADNTEYRKWRRSGTSTGMRLGSRDCGPWCWGQNKCCPDESCRSLHE SCT
Q8IS41	MNLACVLIVAVLFLTASQLATAASYARDKQEYPAVRSSDEMQDSEDLTLTKECTDDSQFCDPNDHDC CSGECIDEGGRGVCAIVPEHV
P0CH14	CCHSSWCKHLC
P56711	MEKLTILLLVAAVLMSTQAQNQEQRQQAKINFLSKRKPSAERWRRDCTSWFGRCTVNSECCSNSCDQ TYCELYAFPSFGA
P01025	SVQLMEKRMNKLGQYSKELRRCCEHGMRNNPMKFSCQRRAQFIHQGNACVKAFLNCCEYIAKLRQQ HSRNKPLGLAR
P12082	MLKKKIEEEAAKYRNAWVKKCCYDGAHRNDDETCEERAARIAIGPECIKAFKSCCAIASQFRADEHH KNMQLGR
P60513	DDCIKPYGFCSLPILKNGLCCSGACVGVCADL
P56708	ACRKKWEYCIVPIIGFIYCCPGLICGPFVCV
P0CH17	TCLARDELCGASFLSNFLCCDGLCLLICV
P0CH18	FGSFIPCAHKGEPCTICCRPLRCHEEKTPTCV
P24159	WCKQSGEMCNVLDQNCCDGYCIVFVCT
P05483	CKSPGTPCSRGMRDCCTSCLLYSNKCRRY
P58916	CKGKGAPCTRLMYDCCHGSCSSSKGRC
P0CH13	DCIPGGENCDVFRPYRCCSGYCILLLCA
G2TRM8	MFRTLKASQKRSLVNLMFGTTALFATATVIFPSLLPCPAMKNPYLDTQSDPGYEELPENSRVIVIDQQS PKSSLVASKQNNPPSKS
Q3E823	MRAVTRNKIVNNLYFSTFLIAFASVAIGSVLPCPAHSVDSDSPAVQQHKLQLAHEQELKRKDALSKKI
Q3E7B2	MVLNPSKYQDTRTWKMTPAMIRARKPFFKGNMLGLTLLLGVTGSVYYYTYHFLHKDNDFADVPIPPI DPQELEALKKEYEAKKKA
Q05809	MSETGETSEYYKQALEEYKEVQEDEDPDVWDTRISKTGCYVENLALQLCHAETGDWRQCFNEMALF RKCWEKNGNRERVSTVDVDGTTSKDSEKKK
Q86WW8	MPKYYEDKPQGGACAGLKEDLGACLLQSDCVVQEGKSPRQCLKEGYCNSLKYAFFECKRSVLDNRA RFRGRKGY
A6YR20	MKLTCVLVVLLLLLPYGDLITNNYIRGAARKVTPWRRNLKTRDVCDSLVGGHCIHNGCWCDQEAPHG NCCDTDGCTAAWWCPGTKWD
P58922	MEKLTILLLVAAVLMSTQALVERAGENHSKENINFLLKRKRAADRGMWGECKDGLTTCLAPSECCSE DCEGSCTMW
P01197	SYSMEHFRWGKPMGRKRRPIKVYPNSFEDESVENMGPEL
Q6UWE3	MAAALALVAGVLSGAVLPLWSALPQYKKKITDRCFHHSECYSGCCLMDLDSGGAFCAPRARITMICLP QTKGATNIICPCRMGLTCISKDLMCSRRCHMI
B2KPN7	MMSKLGVLLCIFLVLFPMATLQLDGDQTADRHADQRGQDLTEQQRNSKRVLKKRDWEYHAHPKPNS FWTLV
Q7Z4G1	MEASSEPPLDAKSDVTNQLVDFQWKLGMAVSSDTCRSLKYPYVAVMLKVADHSGQVKTKCFEMTIP QFQNFYRQFKEIAAVIETV
P25955	MNEKGFTLVEMLIVLFIISILLLITIPNVTKHNQTIQKKGCEGLQNMVKAQMTAFELDHEGQTPSLADL QSEGYVKKDAVCPNGKRIIITGGEVKVEH
P80355	MNRSGKHLISSIILYPRPSGECISSISLDKQTQATTSPLYFCWREK
P0CY50	MKQDMIDYLMKNPQVLTKLENGEASLIGIPDKLIPSIVDIFNKKMTLSKKCKGIFWEQ
P45453	MQDLINYFLNYPEALKKLKNKEACLIGFDVQETETIIKAYNDYYLADPITRQWGD
D0PX85	MIMRMTLTLFVLVVMTAASASGDALTEAKRIPYCGQTGAECYSWCIKQDLSKDWCCDFVKTIARLPP AHICSQ
Q9XYR5	MQTAYWVMVMMMVWIAAPLSEGGKLNDVIRGLVPDDITPQLILGSLISRRQSEEGGSNATKKPYILRA SDQVASGP

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P0C1X2	MEGRRFAAVLILTICMLAPGTGTLLPKDRPSLCDLPADSGSGTKAEKRIYYNSARKQCLRFDYTGQGG NENNFRRTYDCQRTCLYT
P25921	MVVDRKEEKKVAVTLRLTTEENEILNRIKEKYNISKSDATGILIKKYAKEEYGAF
P13920	MKKRLTITLSESVLENLEKMAREMGLSKSAMISVALENYKKGQEK
Q58AD3	MKQKLMVGAFIAAVSLSAAAVDMSNVVKTYDLQDGSKVHVFKDGKMGMENKFGKSMNMPEGKVM ETRDGTKIIMKGNEIFRLDEALRKGHSEGG
Q48271	MKATFQVPSITCNHCVDKIEKFVGEIEGVSFIDVSVEKKSVVVEFDAPATQDLIKEALLDAGQEVV
O32221	MEQKTLQVEGMSCQHCVKAVETSVGELDGVSAVHVNLEAGKVDVSFDADKVSVKDIADAIEDQGYD VAK
Q47840	MKQEFSVKGMSCNHCVARIEEAVGRISGVKKVKVQLKKEKAVVKFDEANVQATEICQAINELGYQAE VI
P11496	QTFQYSRGWTN
P58805	GPMGWVPVFYRF
P85871	GPMEDPLEIIRI
P83301	EDCIAVGQLCVFWNIGRPCCSGLCVFACTVKLP
Q9NDA7	MGKLTILVLVAVALLSTQVMVQGDGDQPADRDAVPRDDNPGGMSEKFLNALQRRGCPWQPYCG
P39103	MSKYAWYTRVTDTLHRLTVLTLVGGTLYMSGGLAYTLYMNGKKYEQQVTQQKALEEDNQQLQSPT APPTE
Q14061	MPGLVDSNPAPPESQEKKPLKPCCACPETKKARDACIIEKGEEHCGHLIEAHKECMRALGFKI
Q12287	MTETDKKQEQENHAECEDKPKPCCVCKPEKEERDTCILFNGQDSEKCKEFIEKYKECMKGYGFEVPS AN
Q49B96	MSTAMNFGTKSFQPRPPDKGSFPLDHLGECKSFKEKFMKCLHNNNFENALCRKESKEYLECRMERKL MLQEPLEKLGFGDLTSGKSEAKK
Q3E731	MSGNPGSSLSALRPTPPERGSFPLDHDGECTKYMQEYLKCMQLVQNENAMNCRLLAKDYLRCRMDH QLMDYDEWSHLGLPEDAPGNNGKTIKDATDNK
P19173	MAGGHVAHLVYKGPSVVKELVIGFSLGLVAGGFWKMHHWNSQRRTKEFYDMLEKGQISVVADEE
Q9CPQ1	MSSGALLPKPQMRGLLAKRLRVHIAGAFIVALGVAAAYKFGVAEPRKKAYAEFYRNYDSMKDFEEM RKAGIFQSAK
P26310	MSTGNESYNLRYPKGFKGYPYNMYKLEGYGTPKGYITLIGVVATLTVSGLFFAKTRSNKREYPTHNK EWRAKTLAYAKETNADPIYQLPKDKI
Q9VHS2	MMNLSRAVVRSFATTAGRRSAAVPKDQIEKGYFEIRKVQEHFQKKDGKPVFLKGSVVDNVLYRVTV ALALVGIGGMGKLFYELSVPKKE
P13183	MFPLAKNALSRLRVQSIQQAVARQIHQKRAPDFHDKYGNAVLASGATFCVAVWVYMATQIGIEWNPS PVGRVTPKEWREQ
P00430	MLGQSIRRFTTSVVRRSHYEEGPGKNIPFSVENKWRLLAMMTLFFGSGFAAPFFIVRHQLLKK
P10174	MANKVIQLQKIFQSSTKPLWWRHPRSALYLYPFYAIFAVAVVTPLLYIPNAIRGIKAKKA
P10176	MSVLTPLLLRGLTGSARRLPVPRAKIHSLPPEGKLGIMELAVGLTSCFVTFLLPAGWILSHLETYRRPE
P10175	MLRLAPTVRLLQAPLRGWAVPKAHITAKPAKTPTSPKEQAIGLSVTFLSFLLPAGWVLYHLDNYKKSS AA
P04039	MLCQQMIRTTAKRSSNIMTRPIIMKRSVHFKDGVYENIPFKVKGRKTPYALSHFGFFAIGFAVPFVAC YVQLKKSGAF
P07255	MTIAPITGTIKRRVIMDIVLGFSLGGVMASYWWWGFHMDKINKREKFYAELAERKKQEN
P82543	MEEKPKGALAVILVLTLTILVFWLGVYAVFFARG
P20609	MSHALPALIKMHITKDIGYGLLLGIIPGVWFKYQIGQSIQKREDFYAAYDKRN
Q9NRP2	MHPDLSPHLHTEECNVLINLLKECHKNHNILKFFGYCNDVDRELRKCLKNEYVENRTKSREHGIAMRK KLFNPPEESEK
G2TRU5	MNKSPIEEEFDLRKQHQKLLKKNCQKEIEDFVKCATGRTFSVTWKCRSENKTMKDCLTKAADEISEW QIRSEYNKAKQESFIGKQDEKK
P20610	MTHALPKVVKSQLVQDIGVALILGSIAGCFFKYGVDKKKQRERVAFYEKYDKEDL
Q96GX8	MGLKMSCLKGFQMCVSSSSSSHDEAPVLNDKHLDVPDIIITPPTPTGMMLPRDLGSTVWLDETGSCPD DGEIDPEA
Q9GU57	MHLSLARSAVLMLLLLFALGNFVVVQSGLITRDVDNGQLTDNRRNLQTEWNPLSLFMSRRSCNNSCQS HSDCASHCICTFRGCGAVNG
Q2I2P4	MTLTKSAVLILVLLLAFDNFADVQPGLITMGGGRLSNLLSKRVSIWFCASRTCSTPADCNPCTCESGVC VDWL
O22059	MFRSDKAEKMDKRRRRQSKAKASCSEEVSSIEWEAVKMSEEEEDLISRMYKLVGDRWELIAGRIPGRT PEEIERYWLMKHGVVFANRRRDFFRK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q95XP7	MHLWQLVLLVILFFGAAFGADLEGSGSGDVSTDAKEAILNAQTLLDAVSSDGSGADVEASGEDVQTFFF
P81033	RSAQGMGKMERLLASYRGALEPSTPLGDLSGSLGHPVE
A1L188	MSANGAVWGRVRSRLRAFPERLAACGAEAAAYGRCVQASTAPGGRLSKDFCAREFEALRSCFAAAA KKTLEGGC
P0C835	MSSGKKAVKVKTPAGKEAELVPEKVWALAPKGRKGVKIGLFKDPETGKYFRHKLPDDYPI
P81162	AVDFDKQCVPTADPGPCKGFMPMWWYNIFTSQCEEFIYGGCQGNDNRYRTKEECDKTCAEASATW DVNA
Q9BQ49	MIGDILLFGTLLMNAGAVLNFKLKKKDTQGFGEESREPSTGDNIREFLLSLRYFRIFIALWNIFMMFCM IVLFGS
P0C1W3	MMSKMGAMFVLLLLFTLASSQQEGDVQARKTHPKREFQRILLRSGRKCNFDKCKGTGVYNCGESCSC EGLHSCRCTYNIGSMKSGCACICTYY
P70964	MGNDSVKDKMKGGFNKAKGEVKDKVGDMADRTDMQAEGKKDKAKGEIQKDIGKAKDKFSDKD
P09928	MATRGWFSESSAQVAQIGDIMFQGHWQWVSNALQATAAAVDNINRNAYPGVSRSGSGEGAFSSSPSN GFRPKRIRSRFNR
P15528	MSGGGVFTDILAAAGRIFEVMVEGHWETVGMLFDSLGKGTMRINRNAYGNLGGGGGSLRGSSPEVSG FAVPTKAVESKFAK
Q46383	MSNGTNIDVAGAINTLAETFGKLFQMQIDVANTALKALADVAEPLGKTATDLIGSFTGAATQVLQSVS SAIAPKK
P45689	MADVTGIALGMIETRGLVPAIEAADAMTKAAEVRLVGRQFVGGGYVTVLVRGETGAVNAAVRAGAD ACERVGDGLVAAHIIARVHSEVENILPKAPQA
P60242	MKNTVKLEQFVALKEKDLQKIKGGEMRLSKFFRDFILQRKK
Q45096	MTVTGQVKWFNNEKGFGFIEVPGENDVFVHFSAIETDGFKSLEEGQKVSFEIEDGNRGPQAKNVIKL
P0A968	MEKGTVKWFNNAKGFGFICPEGGGEDIFAHYSTIQMDGYRTLKAGQSVQFDVHQGPKGNHASVIVP VEVEAAVA
P33911	MLVLSRKINEAIQIGADIEVKVIAVEGDQVKLGIDAPKHIDIHRKEIYLTIQEENNRAAALSSDVISALSSQ KK
P58844	KFLSGGFKEIVCHRYCAKGIAKEFCNCPD
P86258	KPCCSIHDSSCCGI
P58810	VGVCCGYKLCHPC
P86256	DPCCGYRMCVPC
P0C2F2	GNWCCSARVCC
P0C642	QCCPTMPECCRI
Q9BPG6	MRCFPVFIILLLLIASAPCFDARTKTDDDVPLSPLRDNLKRTIRTRLNIRECCEDGWCCTAAPLTGR
P58848	FCCPFIRYCCW
P86264	NIQIICCKHTPACCT
P58809	GICCGVSFCYPC
P80674	QADKYPAGLNPALCPNYPNCDNALIALYSNVAPAIPYAAAYNYPAGVSPAACPNYPFCGAIAPLGYHV REYPAGVHPAACPNYPYCV
P80676	QYYGYPYAAVLPQPVADTPEVASAKAAHFAAYNVAAAAAAAAPDYDAVGVVAPPYPGYSAYVGPL AGIPAIVNGVPADTPEVAAAKVAHFAAHAAANHY
P82171	ATVPAGTSVYQQPAVQYQQYQPYQTVVQSQPQQVVAGRVVPAVYPATVGSQVYPSGVVPATVYNS GIVRNVYPAGVVPAVAGAYQYFPTD
D2Y171	MRFYIGLMAALMLTSILRTDSASVDQTGAEGGLALIERVIRQRDAADVKPVARTNEGPGRDPAPCCQ HPIETCCRR
P80231	GYLGGYAAPALAYGAAPAVAYAAPAAYAPAALTSQSSNILRSYGNLGQVSTYTKTVDTPYSSVTKSD VRVSNDAIAHVAAPALAYAAPAAYAAPAYYH
P0DJB9	GCWICWGPNACCRGSVCHDYCPS
P83359	QAVRYANGYTYDIETGQVSSPYTGRVYETKGKAPFYGFGFEHPYHYYPGYYHGYPHAFY
P0DJB6	TFEPNAEECIVDGRCKHRSDWPCEMSSGTTGRCDVSLGACGCSN
P81588	AAPSKATVGESGIITPGGRLIQLPHGVSIILEGPSAALLSNGDFVTYESS
P81589	GDIIDVDNDLFEHEQDGVAGTSVHGEYEAYDAYGNEYEVKYIADHLGFRVL
P56561	QTGKDATIVELTNDNDGLGQYNFAYRTSDGIARQEQGALKNAGSENEALEVQGSYTYKGVDGKDYT VTFVANENGYQPRVQS
D2Y168	MKFLLFLSVALLLTSFIETEAGPVNEAGVERLFRALVGRGCPADCPNTCDSSNKCSPGFPG
P0CI41	NCPAGCRSQGCCM

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
D2Y3T1	MRCLSIFVLLVLLVSFAVAELDVEGEIVKQLLTRGTLKDADFWKRLEMQGCVCNANAKFCCGEGR
P0DJC3	DCQPCGHNVCC
P82122	AVVPASTVKTALAYTYPIHPYHSVYAHPHSVVIY
P81586	AVLLKGPSGVLFEDGQKRLLPPGVEIVLLTESGAVLSNGENVQF
P0DJB3	DNSCTPKPSCFF
P0DJB4	SLCDKPHHNCIDGQTCYHTCCQNGLKCVRYP
P0DJB5	RPKCCCVCGVVGRKCCSTWDKCHPVHLPCPSS
B3FIA5	MMPVILLLLLSLAIRCADGKAVQGDSDPSASLLTGDKNHDLPVKRDCTTCAGEECCGRCTCPWGDNC SCIEWGK
Q8T0W5	MCRYALIVLVVVVVATNLSEANIFDVPTCEPNRIYKTCGPACPPTCEDPDPDCNETPQCKAGCFCIPG LIENMKGGNCISPSLCP
Q8T0W4	MNAKIVALLIVVGFVGMFNVATAADPLCSLEPAVGLCKASIPRFASVGGKCQEFIYGGCGGNANNFQT QAECEAKCG
Q8T0W3	MRKPITLILVVALALVLLATSEVSAYRACGFPGRRCSPTEECCEGLVCQPRKNGPSMCYRPDP
P43497	MQFSTVASVAFVALANFVAAESAAAISQITDGQIQATTTATTEATTTAAPSSTVETVSPSSTETISQQTE NGAAKAAVGMGAGALAAAAMLL
P86500	FELLPSQDRSCCIQKTLECLENYPGQASQRAHYCQQDATTNCPDTYYFGCCPGYATCMSINAGNNVR SAFDKCINRLCFDPGH
P07471	MALPLKSLSRGLASAAKGDHGGTGARTWRFLTFGLALPSVALCTLNSWLHSGHRERPAFIPYHHLRIR TKPFSWGDGNHTFFHNPRVNPLPTGYEKP
P00429	MAEDIQAKIKNYQTAPFDSRFPNQNQTRNCWQNYLDFHRCEKAMTAKGGDVSVCEWYRRVYKSLC PISWVSTWDDRRAEGTFPGKI
P13184	MLRNLLALRQIAKRTISTSSRRQFENKVPEKQKLFQEDNGIPVHLKGGIADALLYRATLILTVGGTAYA MYELAVASFPKKQD
P01527	ISWPSYPGSEGIRSSNCQKKLNCGTKNIATKGVCKAFCLGRKRFWQKCGKNGSGSKGSKVCNAVLAH AVEKAGKGLIAVTDKAVAAIVKLAAGIA
P19004	RKCFNSPGRLVSKPCPEGNNLCYKMSNRMYPPGFNVRRGCAETCPRRNRLLEVVCCCDTDNCNK
P14541	LKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKIPIKRGCADNCPKNSALLKYVCCSTDKCN
P58990	HPTKPCMYCSFGQCVGPHICCGPTGCEMGTAEANMCSEEDEDPIPCQVFGSDCALNNPDNIHGHCVA DGICCVDDTCTTHLGCL
Q5I4E6	MSGHTSVSFLLLSIVALGMVATVICSCDSEFSSEFCERPEESCSCSTHTCCHWARRDQCMKPQRCISAQ KGNGRRRLIHMQK
B4YSU8	MQKATVLLLALLLLLPLSTAQDAEGSQEDAAQREVDIATRCGGTGDSCNEPAGELCCRRLKCVNSRC CPTTDGC
P80549	MAFTAMTVAPSALADLVLAQKSGCTVCHSVEAAIVGPAYKDVAAKYRGDAAAQDRLVAKVMAGGV GNWGQVPMPPNAHVPAADIKALVTWILGL
P38587	DGQSIYESGTSPTCASCHDRGTAGAPKINEPGDWDGIDLDAEALVDSTMDGKGAMPAYDGRADRDE VKEAVEYMLSTIE
P15452	MKKFLLVAVVGLAGITFANEQLAKQKGCMACHDLKAKKVGPAYADVAKKYAGRKDAVDYLAGKIK KGGSGVWGSVPMPPQNVTDAEAKQLAQWILSIK
P82903	AGDIEAGKAKAAVCAACHGQNGISQVPIYPNLAGQKEQYLVAALKAYKAGQRQGGQAPVMQGQATA LSDADIANLAAYYASLPADGQG
P82599	GGGNDTSNETDTGTSGGETAAVDAEAVVQQKCISCHGGDLTGASAPAIDKAGANYSEEEILDIILNGQ GGMPGGIAKGAEAEAVAAWLAEKK
P31330	SGDLGAERYAKMCKSCHGADGSNAAMSRALKGLPAEEVKAALIGYKEQTYGGKKKGMMERVVKSL TDEDIEVLATHIGTF
O25825	MKKVIMALGVLAFANALMATDVKALAKSCAACHGVKFEKKALGKSKIVNMMSEAEIEKDLMDFKSG ANKNPIMSAQAKKLSDEDIKALAKYIPTLK
P82291	MNETEATLPVFTLEQVAEHHSPDDCWMAIHGKVYDLTPYVPNHPGPAGMMLVWCGQESTEAWETK SYGEPHSSLAARLLQRYLIGTLEEIT
P00087	ADAPPPAFNQCKACHSIDAGKNGVGPSLSGAYGRKVGLAPNYKYSPAHLASGMTIDDAMLTKYLANP KETIPGNKMGAAFGGLKNPADVAAVIAYLKTVK
P00097	ATPAELATKAGCAVCHQPTAKGLGPSYQEIAKKYKGQAGAPALMAERVRKGSVGIFGKLPMTPTPP ARISDADLKLVIDWILKTP
P11732	GGGARSGDDVVAKYCNACHGTGLLNAPKVGDSAAWKTRADAKGGLDGLLAQSLSGLNAMPPKGTC ADCSDDELKAAIGKMSGL
P83391	VDAELLADGKKVFAGNCAACHLGGNNSVLADKTLKKDAIEKYLEGGLTLEAIKYQVNNGKGAMPAW ADRLDEDDIEAVSNYVYDQAVNSKW

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P20958	APEQSKSIPRGEILSLSCAGCHGTDGKSESIIPTIYGRSAEYIESALLDFKSGARPSTVMGRHAKGYSDEE IHQIAEYFGSLSTMNN
P0A646	MNVTVSIPTILRPHTGGQKSVSASGDTLGAVISDLEANYSGISERLMDPSSPGKLHRFVNIYVNDEDVRF SGGLATAIADGDSVTILPAVAGG
P80416	MIPGGLTEAKPATIEIQEIANMVKPQLEEKTNETYEEFTAIEYKSQVVAGINYYYIKIQTGDNRYIHIKVF KSLPQQSHSLILTGYQVDKTKDDELAGF
Q10992	SLEIDELARFAVDEHNKKQNALLEFGKVVNTKEQVVAGKMYYITLEATNGGVKKTYEAKVWVKPWE NFKELQEFKPVDAATS
P81063	AEVNPPEAFNQDVDTYLKIFRNGRYPLDKMAVICSQTGFKLDK
Q86609	MIIWILPCRMPMNLRKDERINISKTSSSKIKEINQLRHIIRKKNRQIQILIIMLNILRCCHRMKE
P20215	MSKEVLEKELFEMLDEDVRELLSLIHEIKIDRITGNMDKQKLGKAYFQVQKIEAELYQLIKVS
P87297	MSSENMDITENIQNEQNKDFDDIDFERRRRLLTLQISKSMNEVVNLMSALNKNLESINGVGKEFENVAS LWKEFQNSVLQKKDREMLDAP
Q12248	MMASTSNDEEKLISTTDKYFIEQRNIVLQEINETMNSILNGLNGLNISLESSIAVGREFQSVSDLWKTLYD GLESLSDEAPIDEQPTLSQSKTK
Q9UTG8	MLQARIEEKQKEYELICKLRDSSNDMVQQIETLAAKLETLTDGSEAVATVLNNWPSIFESIQIASQHSGA LVRIPPSTSNTNASATEQGDVEEV
P69850	MEHNLSPLQQEVLDKYKQLSLDLKALDETIKELNYSQHRQQHSQQETVSPDEILQEMRDIEVKIGLVGT LLKGSVYSLILQRKQEQESLGSNSK
Q50HP4	MNNPMEEQQSALLGRIISNVEKLNESITRLNHSLQLINMSNMNVELASQMWANYARNVKFHLEETHTL KDPI
P77527	MLARSGWLSLEALSEYGLSLAAVRAYVEIGFVEPLEVGGAWYFREEDLLRMAKAERIRKDLGANLIGA ALVVEILERT
P17615	MAYNKSDLVSKIAQKSNLTKAQAEAAVNAFQDVFVEAMKSGEGLKLTGLFSAERVKRAARTGRNPR TGEQIDIPASYGVRISAGSLLKKAVTE
Q46121	MTKADFISLVAQTAGLTKKDATTATDAVISTITDVLAKGDSISFIGFGTFSTQERAAREARVPSTGKTI KVPATRVAKFKVGKNLKEAVAKASGKKKK
P05514	MNKGELVDAVAEKASVTKKQADAVLTAALETIIEAVSSGDKVTLVGFGSFESRERKAREGRNPKTNE KMEIPATRVPAFSAGKLFREKVAPPKA
P02345	MVGISELSKEVAKKANTTQKVARTVIKSFLDEIVSQANGGQKINLAGFGIFERRTQGPRKARNPQTKK VIEVPSKKKFVFRASSKIKYQQ
P89113	MKVSQVFISAISVFGLATSVNAQNASNTTSNAAPALHAQNGQLLNAGVVGAAVGGALAFLI
Q99039	MEKKLEEVKQLLFRLELDIKETTDSLRNINKSIDQLDKYNYAMKIS
P68731	MDDKDLKLILHKTFIEIYSDLEELADIAKKGKPSMEKYVEEIEQRCKQNILAIEIQMKIK
P81160	MKTPLFLLLVVLASLLGLALSQDRNDTEWIQSQKDREKWCRLNLGPYLGGRCRK
P00273	MANEGDVYKCELCGQVVKVLEEGGGTLVCCGEDMVKQ
D2KX90	MAYLKIVLVALMLVLAVSAMRRPDQQDQDISVAKRVACKCDDDGPDVRSATFTGTVDLGSCNSGWE KCASYYTVIADCCRKPRG
P80178	PYDRISNSAFSDF
P07448	GVIAWELQHNEPGRKDSTAG
Q7M458	SINNTGGSGNRRLDKNGFAGQ
P14775	MKTTLIAAAIVALSGLAAPALAYDGTKCKAAGNCWEPKPGFPEKIAGSKYDPKHDPKELNKQADSIK QMEERNKKRVENFKKTGKFEYDVAKISAN
P0ABR1	MRIEVTIAKTSPLPAGAIDALAGELSRRIQYAFPDNEGHVSVRYAAANNLSVIGATKEDKQRISEILQET WESADDWFVSE
Q47150	MAANAFVRARIDEDLKNQAADVLAGMGLTISDLVRITLTKVAREKALPFDLREPNQLTIQSIKNSEAGI DVHKAKDADDLFDKLGI
P0C6B0	MNSANPCCDPITCKPRRGEHCVSGPCCRNCKFLNPGTICKRTMLDGLNDYCTGVTSDCPRNPWKSEE ED
P36235	SDDKCQGRPMYGCREDDDSVFGWTYDSNHGQCWKGSYCKHRRQPSNYFASQQECRNTCGA
P17497	SPPVCGNKILEQGEDCDCGSPANCQDRCCNAATCKLTPGSQCNYGECCDQCRFKKAGTVCRIARGD WNDDYCTGKSSDCPWNH
P82014	RMPSLSIDLPMSVLRQKLSLEKERKVQALRAAANRNFLNDI
P56618	SPTISITAPIDVLRKTWEQERARKQMVKNREFLNSLN
P82015	SFSVNPAVEILQHRYMEKVAQNNRNFLNRV
P56619	AGALGESGASLSIVNSLDVLRNRLLLEIARKKAKEGANRNRQILLSL
P23834	TGAQSLSIVAPLDVLRQRLMNELNRRRMRELQGSRIQQNRQLLTSI

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P85830	GLDLGLSRGFSGSQAAKHLMGLAAANYAGGP
P82172	MRLAYLLLLLVAVLFQAGGGSVEAFVQHRPRDCESINGVCRHKDTVNCREIFLADCYNDEQKCCRK
Q03AZ0	MADEAIKNGVLDILADLTGSDDVKTNLDLNLFETGLLDSMGTVQLLLELQSQFGVEAPVSEFDRSQWD TPNKIIAKVEQAQ
P19731	MSSLVYIAFQDNDNARYVVEAIIQDNPHAVVQHHPAMIRIEAEKRLEIRRETVEENLGRAWDVQEMLV DVITIGGNVDEDDDRFVLEWKN
P13123	MVKVKFKYKGEEKEVDTSKIKKVWRVGKMVSFTYDDNGKTGRGAVSEKDAPKELLDMLARAEREK K
Q96FX2	MAVFHDEVEIEDFQYDEDSETYFYPCPCGDNFSITKEDLENGEDVATCPSCSLIIKVIYDKDQFVCGET VPAPSANKELVKC
Q3E840	MSTYDEIEIEDMTFEPENQMFTYPCPCGDRFQIYLDDMFEGEKVAVCPSCSLMIDVVFDKEDLAEYYE EAGIHPPEPIAAAA
O94777	MATGTDQVVGLGLVAVSLIIFTYYTAWVILLPFIDSQHVIHKYFLPRAYAVAIPLAAGLLLLLFVGLFIS YVMLKTKRVTKKAQ
Q9P2X0	MTKLAQWLWGLAILGSTWVALTTGALGLELPLSCQEVLWPLPAYLLVSAGCYALGTVGYRVATFHD CEDAARELQSQIQEARADLARRGLRF
O94633	MQRIHKVILYYVSLTILYRVTYLFDLEEPWSTLRPYTPYLFILAFGSYLGITLLYNVATTNDKPEAYVDL VKDIKEAQDALRSKGMTIED
Q9C005	MEPEQMLEGQTQVAENPHSEYGLTDNVERIVENEKINAEKSSKQKVDLQSLPTRAYLDQTVVPILLQG LAVLAKERPPNPIEFLASYLLKNKAQFEDRN
Q9VQS6	MFAVMRIDNDDCRSDFRRKMRPKCEFICKYCQRRFTKPYNLMIHERTHKSPEITYSCEVCGKYFKQR DNLRQHRCSQCVWR
P24274	MSEHSFAFAYTVFFMVSVATMSELSSRRYELTQSEKRVIPTRHSTMKSCPWVVFHTRATLTRSFPCLC SNLNPSTTSCLAIFSEHSYVVVF
P13320	MAKKEMVEFDEAIHGEDLAKFIKEASDHKLKISGYNELIKDIRIRAKDELGVDGKMFNRLLALYHKDN RDVFEAETEEVVELYDTVFSK
Q95Y72	MSSTAAKKDVKSSAVPVTAVVEKKEFEEEEFEEFPVQEWAERAEGEEDDVNVWEDNWDDETHESEF SKQLKEELRKSGHQVA
O14140	MSRAALPSLENLEDDDEFEDFATENWPMKDTELDTGDDTLWENNWDDEDIGDDDFSVQLQAELKKK GVAAN
Q46582	MEEAKQKVVDFLNSKSGSKSKFYFNDFTDLFPDMKQREVKKILTALVNDEVLEYWSSGSTTMYGLKG AGKQAAAEHED
P82081	GLVPNLLNNLGL
P82085	GAVSGLLTNLGL
P23373	MNPKHWGRAVWTIIFIVLSQAGLDGNIEACKRKLYTIVSTLPCPACRRHATIAIEDNNVMSSDDLNYIY YFFIRLFNNLASDPKYAIDVTKVNPL
P15133	MIPRVLILLTLVALFCACSTLAAVAHIEVDCIPPFTVYLLYGFVTLILICSLVTVVIAFIQFIDWVCVRIAY LRHHPQYRDRTIADLLRIL
Q57231	MAVTYEKTFEIEIINELSASVYNRVLNYVLNHELNKNDSQLLEVNLLNQLKLAKRVNLFDYSLEELQAV HEYWRSMNRYSKQVLNKEKVA
P30569	MGCDDKCGCAVPCPGGTGCRCTSARSGAAAGEHTTCGCGEHCGCNPCACGREGTPSGRANRRANC SCGAACNCASCGSATA
P80936	ISDFDEYEPLNDADNNEVLDF
P80941	SAIDPNPDTPDSE
P25331	MANKLTAVIVVALAVAFMVNLDYANCSPAIASSYDAMEICIENCAQCKKMFGPWFEGSLCAESCIKAR GKDIPECESFASISPFLNKL
P58748	MSDFNLVGVIRVMPTDPDVNLDELEEKLKKVIPEKYGLAKVEREPIAFGLVALKFYVLGRDEEGYSFD EVAEKFEEVENVESAEVETVSRI
Q64214	MTDVLVVLKVFPDSDEVNLDNLYTDISSKLPKEYKIIRKETEPIAFGLNALILYVQMPEQTEGGTDNLE EVVNNIQGVSHAEVVGITRLGF
P82097	NEEEKVKWEPDVP
P11924	ISINQDLKAITDMLLTEQIQARQRCLAALRQRLLDLDSDVSLFNGDLLPNGRCS
P80594	GSGVSNGGTEMIQLSHIRERQRYWAQDNLRRRFLEK
Q03071	MSQDFVTLVSKDDKEYEISRSAAMISPTLKAMIEGPFRESKGRIELKQFDSHILEKAVEYLNYNLKYSG VSEDDDEIPEFEIPTEMSLELLLAADYLSI
P60003	MGRRKSKRKPPPKKKMTGTLETQFTCPFCNHEKSCDVKMDRARNTGVISCTVCLEEFQTPITYLSEP VDVYSDWIDACEAANQ
Q02973	MASQQEKKQLDERAKKGETVVPGGTGGKSFEAQQHLAEGRSRGGQTRKEQLGTEGYQQMGRKGGL STGDKPGGEHAEEEGVEIDESKFRTKT

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P96621	MSESSARTEMKISLPENLVAELDGVAMREKRSRNELISQAVRAYVSERTTRHNRDLMRRGYMEMAKI NLNISSEAHFAECEAETTVERLVSGG
Q8I817	MKTALPLLLLTCLVAAVQSTGSQGCPTYVSEKCTARLQECSNNQQQEPLQNCTAVHADCVVQATEDC QREQSQLNHDHLNNHTTTQQP
P62651	MRPRDQGFLVLGFTYSVLLISLATFYWLRNNDSFLHYWCVLLLCPATLWLWALIAWCDSEMFASSKD E
P68311	ISHGFRYDAIAKLFRPFFCGDGYGHRIGETVYYAGSLKYCARSFDVGAEIICKGFYYFGIYKRRVSEV
P95242	MTINYQFGDVDAHGAMIRAQAGLLEAEHQAIVRDVLAAGDFWGGAGSVACQEFITQLGRNFQVIYEQ ANAHGQKVQAAGNNMAQTDSAVGSSWA
P95243	MATRFMTDPHAMRDMAGRFEVHAQTVEDEARRMWASAQNISGAGWSGMAEATSLDTMAQMNQAF RNIVNMLHGVRDGLVRDANNYEQQEQASQQILSS
P0A564	MTEQQWNFAGIEAAASAIQGNVTSIHSLLDEGKQSLTKLAAAWGGSGSEAYQGVQQKWDATATELN NALQNLARTISEAGQAMASTEGNVTGMFA
Q99WU4	MAMIKMSPEEIRAKSQSYGQGSDQIRQILSDLTRAQGEIAANWEGQAFSRFEEQFQQLSPKVEKFAQLL EEIKQQLNSTADAVQEQDQQLSNNFGLQ
P0A567	MAEMKTDAATLAQEAGNFERISGDLKTQIDQVESTAGSLQGQWRGAAGTAAQAAVVRFQEAANKQ KQELDEISTNIRQAGVQYSRADEEQQQALSSQMGF
P0A568	MSQIMYNYPAMLGHAGDMAGYAGTLQSLGAEIAVEQAALQSAWQGDTGITYQAWQAQWNQAMEDL VRAYHAMSSTHEANTMAMMARDTAEAAKWGG
P0AEJ8	MKLAVVTGQIVCTVRHHGLAHDKLLMVEMIDPQGNPDGQCAVAIDNIGAGTGEWVLLVSGSSARQA HKSETSPVDLCVIGIVDEVVSGGQVIFHK
P0C8E8	MRALLARLLLCVLVVSDSKGLVSTIESRTSGDGADNFDVVSCNKNCTSGQNECPEGCFCGLLGQNKKG HCYKIIGNLSGEPPVVRR
Q7W7Q2	MASSKQADPQTDARPLPQDFETALAELESLVSAMENGTLPLEQSLSAYRRGVELARVCQDRLAQAEQ QVKVLEGDLLRPLDPAALDDE
P0A8G9	MPKKNEAPASFEKALSELEQIVTRLESGDLPLEEALNEFERGVQLARQGQAKLQQAEQRVQILLSDNE DASLTPFTPDNE
P0DJ94	HSDATFTAEYSKLLAKLALQKYLESILGSSTSPRPPSS
P20394	MKIILWLCVFGLFLATLFPVSWQMPVESGLSSEDSASSESFASKIKRHSDGTFTSDLSKQMEEEAVRLFI EWLKNGGPSSGAPPPSG
P21937	MLYKKLRSQGNFRKNDSAYFKLENKRELKGDNLPVEEKVRQTIEKFKDDVSEIRRLADDSDFGCNGK ETGGAMHIVCFFQKNYDWMKGQWQN
P03639	MVRWTLWDTLAFLLLLSLLLPSLLIMFIPSTFKRPVSSWKALNLRKTLLMASSVRLKPLNCSRLPCVYA QETLTFLLTQKKTCVKNYVRKE
Q91DM1	MGLVWSLISNSIQTIIADFAISVIDAALFFLMLLALAVVTVFLFWLIVAIGRSLVARCSRGARYRPV
P0C6Y6	MGSLWSKISQLFVDAFTEFLVSVVDIAIFLAILFGFTVAGWLLVFLLRVVCSALLRSRSAIHSPELSKVL
P85029	ITFSKIYRSCKSDSDCGNQKCARGRCV
Q99J19	MNWKVLEHVPLLLYILAAKTLILCLAFAGVKMYQRRSLEGKLQAEKRKQSEKKAS
Q0P4B9	MTSSSTPRMHTYKRTSSPRSPTNTGELFTPAHEENVRFIHDTWLCVLRDIKCPQNHERNDRGPQEYVE KNPNPNLHSFIPVDLSDLKKRNTQDSKKS
P20222	MKSTPFFYPEAIVLAYLYDNEGIATYDLYKKVNAEFPMSTATFYDAKKFLIQEGFVKERQERGEKRLY LTEKGKLFAISLKTAIETYKQIKKR
P85468	SLNTKNDFMRF
P85462	TPPQPADNFIRF
P83321	DGRTPALRLRF
P43173	SDIGISEPNFLRF
P41869	APNQPSDNMIRF
P30273	MIPAVVLLLLLLVEQAAALGEPQLCYILDAILFLYGIVLTLLYCRLKIQVRKAAITSYEKSDGVYTGLST RNQETYETLKHEKPPQ
D5ARY6	MAMKIDPELCTSCGDCEPVCPTNAIAPKKGVYVINADTCTECEGEHDLPQCVNACMTDNCINPAA
P30438	MKGACVLVLLWAALLLISGGNCEICPAVKRDVDLFLTGTPDEYVEQVAQYKALPVVLENARILKNCV DAKMTEEDKENALSVLDKIYTSPLC
P0AEL3	MQYTPDTAWKITGFSREISPAYRQKLLSLGMLPGSSFNVVRVAPLGDPIHIETRRVSLVLRKKDLALLE VEAVSC
A6TF33	MASLMEVRDMLALQGRMEAKQLSARLQTPQPLIDAMLERMEAMGKVVRISETSEGCLSGSCKSCPEG KAACRQEWWALR
O67065	MKVIINGKEFDIPKGVRFGELSHEIEKAGIEFGCTDGQCGVCVARVIKGMECLNEPSEEEEETLWRVG AVDEDQRLTCQLVIEKEDCDEIVIESED

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P00210	MARKFYVDQDECIACESCVEIAPGAFAMDPEIEKAYVKDVEGASQEEVEEAMDTCPVQCIHWEDE
P08813	MGWTVTVDTDKCTGDGECVDVCPVEVYELQDGKAVPVNEEECLGCESCVEVCEAGAITVEEN
P18324	MTMRVSADRTVCVGAGLCALTAPGVFDQDDDGIVTVLTAEPAADDDRRTAREAGHLCPSGAVRVVE DTE
P00251	MATYKVTLINEEEGINAILEVADDQTILDAGEEAGLDLPSSCRAGSCSTCAGKLVSGAAPNQDDQAFLD DDQLAAGWVMTCVAYPTGDCTIMTHQESEVL
P00237	AYKVTLKTPDGDITFDVEPGERLIDIGSEKADLPLSCQAGACSTCLGKIVSGTVDQSEGSFLDDEQIEQ GYVLTCIAIPESDVVIETHKEDEL
P18325	MRIHVDQDKCCGAGSCVLAAPDVFDQREEDGIVVLLDTAPPAALHDAVREAATICPAAAITVTD
P59799	MPKVIVANINAEFEGIENETIMQILYRNGIEIDSACGGHGQCTSCKVLIISGSENLYPAEFEEKDTLEENG MDPETERLSCQAKLNGKGDVVIYLP
P50727	MAKYTIVDKDTCIACGACGAAAPDIYDYDDEGIAFVTLDENKGVVEVPEVLEEDMIDAFEGCPTDSIK VADEPFEGDPLKFE
P00195	MAYKIADSCVSCGACASECPVNAISQGDSIFVIDADTCIDCGNCANVCPVGAPVQE
P00209	PIEVNDDCMACEACVEICPDVFEMNEEGDKAVVINPDSDLDCVEEAIDSCPAEAIVRS
P00202	PATVNADECSGCGTCVDECPNDAITLDEEKGIAVVDNDECVECGACEEACPNQAIKVEE
P00203	MKVTVDQDLCIACGTCIDLCPSVFDWDDEGLSHVIVDEVPEGAEDSCARESVNECPTEAIKEV
P29603	MAWKVSVDQDTCIGDAICASLCPDVFEMNDEGKAQPKVEVIEDEELYNCAKEAMEACPVSAITIEEA
P03942	MPHVICEPCIGVKDQSCVEVCPVECIYDGGDQFYIHPEECIDCGACVPACPVNAIYPEEDVPEQWKSYI EKNRKLAGLE
P21149	MLSQVCRFGTITAVKGGVKKQLKFEDDQTLFTVLTEAGLMSADDTCQGNKACGKCICKHVSGKVAA EDDEKEFLEDQPANARLACAITLSGENDGAVFEL
P0A8P3	MSRTIFCTFLQREAEGQDFQLYPGELGKRIYNEISKEAWAQWQHKQTMLINEKKLNMMNAEHRKLLE QEMVNFLFEGKEVHIEGYTPEDKK
P0A6R3	MFEQRVNSDVLTVSTVNSQDQVTQKPLRDSVKQALKNYFAQLNGQDVNDLYELVLAEVEQPLLDMV MQYTRGNQTRAALMMGINRGTLRKKLKKYGMN
Q5F881	MASVVIRNLSEATHNAIKFRARAAGRSTEAEIRLILDNIAKAQQTVRLGSMLASIGQEIGGVELEDVRGR NTDNEVSL
P82064	AGPVSKLVSGIGL
P0AEM4	MSIDRTSPLKPVSTVQPRETTDAPVTNSRAAKTTASTSTSVTLSDAQAKLMQPGSSDINLERVEALKLA IRNGELKMDTGKIADALINEAQQDLQSN
P74931	MMTQGAVLGLIREGVFQVVLLVAPVLCTALVVGLIVAIFQAVTSIQEQTLTFVPKMLTILGMIALLGG WMLTMLQNYTVRLFDIIPQLVRSGPV
Q7YX32	MNFSGFEFSSIVAFFLLILQLSTAAVLPADYAYGVADEMSALPDSGSLFAEQRPSKRAQTFVRFGKRAQ TFVRFGKRGQTFVRFGRSAPFEQ
Q9XVX1	MSFQLTLFSMLFLLIAVVVGQPIQSQNGDLKMQAVQDNSPLNMEAFNDDSALYDYLEQSDPSLKSMEK RWANQVRFGKRASWASSVRFG
O17058	MLSSRTSSIILILAILVAIMAVAQCRNIQYDVEEMTPEAAFRYAQWGEIPHKRVPSAGDMMVRFGKRSI
Q8MPY9	MKVMFMLALLFSSLVATSAFRLPFQFFGANEDFNSGLTKRNYYESKPYKREFNADDLTLRFGKRGGA GEPLAFSPDMLSLRFGK
Q18184	MFSLTQILTFLLVAITLMTFSSAQPIDEERPIFMERREASAFGDIIGELKGKGLGGRMRFGKRSSSPDISL AEMRAIYGGDQSNIFNFK
P19369	MPNFFRNGCIALVGSVAAMGAAHAEGGIAEAAGKALDSAQSDVTITAPKVMMVVATVVGVGILINMM RKA
Q12497	MLRTTFLRTPRQLMRKSPRASFSIVTRAAFPHLKNNQDEAEKKEQGLFDSNKKRLDTLEHGKNPDYK QPGMEDLKKKGDDARIEQNRPDDGVY
Q10731	MASKNLFVLFFIFALFAANIAALQCPKNSEVRNSPCPRTCNDPYGQNSCITVIRETCHCKGELVFDSDSI CVPISQC
P0AAP3	MPSTPEEKKKVLTRVRRIRGQIDALERSLEGDAECRAILQQIAAVRGAANGLMAEVLESHIRETFDRND CYSREVSQSVDDTIELVRAYLK
Q867W1	MCVQTRMLVAVAVVLVVLAVLSDPVSAGYRKPPFNGSIFGKRAGGDSLYEPGKALASACQVAVEACA AWFPGPEKK
P14903	MAAKNSEMKFAIFFVVLLTTTLVDMSGISKMQVMALRDIPPQETLLKMKLLPTNILGLCNEPCSSNSDC IGITLCQFCKEKTDQYGLTYRTCNLLP
Q9UTP3	MEDKYILLSAVETFKSRLEELLMQSAKVQKQTMLRKELASSMNDMASTVQEALNKKKSS
P80680	EVASDDVAAEEAAAAPKIGRRVRVTAPLRVYHVLKAPDLDIQGMEGVVKQYVCVWKGKRVTANFPF KVEFELAVEGQPKPVRFFAHLREDEFEFVDG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q55781	MNVGDRVRVTSSVVVYHHPEHKKTAFDLQGMEGEVAAVLTEWQGRPISANLPVLVKFEQRFKAHFR PDEVTLIED
Q9KUJ3	MRVFALTLSLLLVWLLYTLMWGKNGVMDFRAVQAEIEVQQQVNANLHLRNQEMFAEIDDLRQGLDA IEERARNELGMVKDGETFYRIIGEESRQ
Q63113	MEGITCAFLLVLAGLPVLEANGPVDKGSPFYYDWESLQLGGMIFGGLLCIAGIAMALSGKCKCRRNHT PSSLPEKVTPLITPGSAST
Q9D164	METVLVLCSLLAPVVLASAAEKEKEKDPFYYDYQTLRIGGLVFAVVLFSVGILLILSRRCKCSFNQKPR APGDEEAQVENLITTNAAEPQKAEN
P59648	MATPTQSPTNVPEETDPFFYDYATVQTVGMTLATIMFVLGIIIILSKKVKCRKADSRSESPTCKSCKSE LPSSAPGGGGV
O80294	MSELGNLETTVTGKIKRFNNGGGYYYTTVVSPAADAYSFPPVIRIKSKKSLGRVGDEIADIHCRITGYE RSFPYTDKQTGEQSRGFNVDMLLELLE
P03670	MLTVEIHDSQVSVKERSGVSQKSGKPYTIREQEAYIDLGGVYPALFNFNLEDGQQPYPAGKYRLHPAS FKINNFGQVAVGRVLLESVK
P03672	MNIQITFTDSVRQGTSAKGNPYTFQEGFLHLEDKPFPLQCQFFVESVIPAGSYQVPYRINVNNGRPELA FDFKAMKRA
P68676	MKVQIMSSAVAVRSFPAREGKPATHFREQTAAVLREGDFPLPFTIGLDEDQPPYGEGFYIIDPKSLQN NKFGGLEFGRRIRLIPDLTAKLQQQPAKVG
P69535	MEQVADFDTIYQAMIQISVVLCFALGIIAGGQR
P69538	MSVLVYSFASFVLGWCLRSGITYFTRLMETSS
P47215	GWTLNSAGYLLGPHAIDNHRSFNEKHGIA
P04564	QLGLQDPPHMVADLSKKQGPWVEEEEAAYGWMDF
P68807	MTKVTREEVEHIANLARLQISPEETEEMANTLESILDFAKQNDSADTEGVEPTYHVLDLQNVLREDKAI KGIPQELALKNAKETEDGQFKVPTIMNEEDA
Q5XAC6	MKISEEEVRHVAKLSKLSFSESETTTFATTLSKIVDMVELLNEVDTEGVAITTTMADKKNVMRQDVAE EGTDRALLFKNVPEKENHFIKVPAILDDGGDA
Q9WY94	MIKVTKDLVLHLENLARLELSEDQRESLMKDFQEILDYVELLNEVDVEGVEPMYTPVEDSAKLRKGDP RFFEMRDLIKKNFPEEKDGHIKVPGIHR
Q9LCX4	MELSPELLRKLETLAKIRLSPEEEALLLQDLKRILDFVDALPRVEEGGAEEALGRLREDEPRPSLPQAE ALALAPEAEDGFFRVPPVLE
P02698	MPVINIEDLTEKDKLKMEVDQLKKEVTLERMLVSKCCEEFRDYVEERSGEDPLVKGIPEDKNPFKELK GGCVIS
P63214	MKGETPVNSTMSIGQARKMVEQLKIEASLCRIKVSKAAADLMTYCDAHACEDPLITPVPTSENPFREK KFFCALL
Q9NFZ3	MDPSALQNMDRDALKKQIENMKYQASMERWPLSKSIAEMRSFIEENEKNDPLINAPDKKNNPWAEKG KCVIM
Q95ZG8	MSESQLKKVLKENETLKAQLEKSTTILKVSEACESLQDYCTKTSDPFIPGWSGENEWTKPLKGNGCSV L
Q01821	MNKMQGKKKKKEEEEEEERIIPPELWKLIIEQYKRQLAKTDVMKVSETVKLHEEKIKEKVPTDHIIHA QKPNAWVEETKKSGGCLLV
P22800	ENFSGGCVAGYMRTPDGRCKPTFYQ
P11470	MKEKEFQSKPLLTKREREVFELLVQDKTTKEIASELFISEKTVRNHISNAMQKLGVKGRSQAVVELLR MGELEL
P62165	MPAMVGHIRIVNIGSSGIFHIGDVFAIRPISYSRAFAGAGSFNVGDNVSVYNYQSATTVNDSDVVDQAII GST
O06720	MNFYINQTIQINYLRLESISNSSILQIGSAGSIKSLSNLYNTGSYVEPAPEVSGSGQPLQLQEPDTGSLVPL QPPGR
P0A3T9	MNLNVVNRELKVGQIKMNGVSSSALFLIGDANLLILSSILDTPFETVTEGPFVPLVTDVPPTPG
P62183	MPSVVGNLVVQNSNGSFNLGDFYNVSPKENTKAYNGSGASNVGFVVNTFNGVSATNTFDSDVADQD QIGTA
O06716	MPAIVGPIAINSISGGVVNFGDSFYLSPKSSSKSALGSGAGNTGDFLLLNNAVNATNYIDPDVNDQDMV GNG
P30047	MPYLLISTQIRMEVGPTMVGDEQSDPELMQHLGASKRRALGNNFYEYYVDDPPRIVLDKLERRGFRV LSMTGVGQTLVWCLHKE
Q4G0N0	MNVKGKVILSMLVVSTVIIVFWEFINSTEGSFLWIYHSKNPEVDDSSAQKGWWFLSWFNNGIHNYQQG EEDIDKEKGREETKGRKMTQQSFGYGTGLIQT
P84848	EEAIDAQALVDQNCTGCHGSEVYTRDERRVESLDALHGQVRMCEQNLELTWFDDQVDAVTTLLNRE YYNFEP

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q76CA1	MASFRTLFACVVILCCVLWSSMARYGEDMEVETEMNKRDVGVACTGQYASSFCLNGGTCRYIPELGE YYCICPGDYTGHRCEQMSV
P09680	YAEGTFISDYSIAMDKIRQQDFVNWLLAQKGKKSDWIHNITQ
P26649	MDHSLNSLNNFDFLARSFARMHAEGRPVDILAVTGNMDEEHRTWFCARYAWYCQQMMQARELELE H
P80051	FPDDNFLTPGCPECRLKENLRFSNMGIGRIYQCSGCCYSRAYPTPMRSKKTMLVPKNITSEAKCCVAK TQYRVTVMDNVKIENHTACHCSTCLYHKS
P06028	MYGKIIFVLLLSEIVSISALSTTEVAMHTSTSSSVTKSYISSQTNGETGQLVHRFTVPAPVVIILIILCVMA GIIGTILLISYSIRRLIKA
P68688	MQTVIFGRSGCPYCVRAKDLAEKLSNERDDFQYQYVDIRAEGITKEDLQQKAGKPVETVPQIFVDQQ HIGGYTDFAAWVKENLDA
P00276	MFKVYGYDSNIHKCVYCDNAKRLLTVKKQPFEFINIMPEKGVFDDEKIAELLTKLGRDTQIGLTMPQV FAPDGSHIGGFDQLREYFK
P81027	HAGTYTSDVSSYLQDQAAKEFVSWLKTGRGRRD
P23063	HADGIYTSDVASLTDYLKSKRFVESLSNYNKRQNDRRM
Q77CE4	MPRSPLIVAVVAAALFAIVRGRDPLLDAMRREGAMDFWSAGCYARGVPLSEPPQALVVFYVALTAVM VAVALYAYGLCFRLMGASGPNKKESRGRG
P28980	MLSTRFVTLAILACLLVVLGLARGAGGDPGVKQRIDVAREEERRDFWHAACSGHGFPITTPSTAAILF YVSLLAVGVAVACQAYRAVLRIVTLEMLQHLH
Q87088	MVSSAGLSLTLVAALCALVAPALSSIVSTEGPLPLLREESRINFWNAACAARGVPVDQPTAAAVTFYIC LLAVLVVALGYATRTCTRMLHASPAGRRV
P0C764	MGSITASFILITMQILFFCEDSSGEPNFAERNFWHASCSARGVYIDGSMITTLFFYASLLGVCVALISLAY HACFRLFTRSVLRSTW
P37042	MEKSRKILVGVLLFTASVAICLAQHWSYGLQPGGKRNAENLVESFQEIANEMESLGEGQKAECPGSYQ HPRLSDLKETMASLIEGEARRKEI
P33439	MGIKRALWWMVVCVVVLQVSAQHWSHGLNPGGKRAVMQESAEEIPRSSGYLCDYVAVSPGNKPFRL KDLLTPVAGREIEE
P51919	MAPQTSNLWILLLLVVVMMMSQGCCQHWSYGLSPGGKRDLDSLSDTLGNIIERFPHVDSPCSVLGCVE EPHVPRMYRMKGFIGSERDIGHRMYKK
P45656	MKAFPTFALLFLVLLFSAHVSDAQHWSYGLRPGGKRDTESLQDMYHETPNEVALFPELERLECSVPQS RLNVLRGALMNWLEGENRKKI
P51924	MVHICRLFVVMGMLLCLSAQFASSQHWSHGWYPGGKREIDVYDSSEVSGEIKLCEAGKCSYLRPQGR NILKTILLDAIIRDSQKRK
P51917	MEGKGRVLVQLLMLACVLEVSLCQHWSYGWLPGGKRSVGEVEATFRMMDSGDAVLSIPMDSPMERL SPIHIVSEVDAEGLPLKEQRFPNRRGRD
P51922	MRPYNVIVVMVVLLALVLHAVLSQHWSYGWLPGGKRSVGELEATIRMMGTGGVVSLPEETSAQTQE RLRPYNIINDGGYFNRKKRFFHE
P0CI74	MLADKVKLSAKEILEKEFKTGVRGYKQEDVDKFLDMIIKDYETFHQEIEELQQENLQLKKQLEEASKK QPVQSNTTNFDILKRLSNLEKHVFGSKLYD
P40754	FGFLPIYRRPAS
P08072	MVPRDLVATLLCAMCIVQATMPSLDNYLYIIKRIKLCNDDYKNYCLNNGTCFTVALNNVSLNPFCACH INYVGSRCQFINLITIK
P31886	APAPSGGGSAPLAKIYPRGSHWAVGHLM
P86961	MRRSVLVVFLVLAVTNVAVEAISRRGSFLAGGLLGLGLGAAASRGFGFPGYYGGYYGGGYYPMGGY YPMGGYYPMGGFYPSYHTFGGYYG
P08958	MAQPDSSGLAEVLDRVLDKGVVVDVWARVSLVGIEILTVEARVVAASVDTFLHYAEEIAKIEQAELTA GAEAAPEA
D2Y2C5	MTTVGVSLFRRSPEKITMKIATFLGLSFLLIASYVLICEAQHPGFQELLILEENMRDPENSKERSCAKPR ENCNRMNILCCRGECVCPTFGDCFCYGD
P11910	MKKSLFAAALLSLALAACGGEKAAEAPAAEASSTEAPAAEAPAAEAPAAEAAAAEAPAAEAPAAEAP AAEAAATEAPAAEAPAAEAAK
D2Y236	MKSIVFVALFGLALLAVVCSASEDAHKELLKEVVRAMVVDKTDAVQAEERECRWYLGGCSQDGDCC KHLQCHSNYEWCIWDGTFSK
O43504	MEATLEQHLEDTMKNPSIVGVLCTDSQGLNLGCRGTLSDEHAGVISVLAQQAAKLTSDPTDIPVVCLE SDNGNIMIQKHDGITVAVHKMAS
O13916	MLRRNPTAIQITAEDVLAYDEEKLRQTLDSESTTEEALQKNEESTRLSPEKKKIIRERRIGITQIFDSSMH PSQGGAAQS
Q9UBK5	MIHLGHILFLLLLPVAAAQTTPGERSSLPAFYPGTSGSCSGCGSLSLPLLAGLVAADAVASLLIVGAVFL CARPRRSPAQEDGKVYINMPGRG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9KV11	MAKGQSLQDPFLNALRRERIPVSIYLVNGIKLQGQIESFDQFVILLKNTVNQMVYKHAISTVVPARPVS HHSGDRPASDRPAEKSEE
P0ACE3	MSEKPLTKTDYLMRLRRCQTIDTLERVIEKNKYELSDNELAVFYSAADHRLAELTMNKLYDKIPSSVW KFIR
P76106	MKQSEFRRWLESQGVDVANGSNHLKLRFHGRRSVMPRHPCDEIKEPLRKAILKQLGLS
Q9P298	MSANRRWWVPPDDEDCVSEKLLRKTRESPLVPIGLGGCLVVAAYRIYRLRSRGSTKMSIHLIHTRVAA QACAVGAIMLGAVYTMYSDYVKRMAQDAGEK
Q9KMG5	MALEFKDKWLEQFYEDDKRHRLIPSSIENALFRKLEILDAAQAESDLRIPPGNRFEHLEGNLKGWCSIR VNKQYRLIFQWVDGVALNTYLDPHKY
P00266	GTNASMRKAFNYQEVSKTAGKNCANCAQFIPGASASAAGACKVIPGDSQIQPTGYCDAYIVKK
P83342	AELERLSEDDATAQALSYTHDASGVTHDSYQEGSRCSNCLLYSNPDAKDWGPCSVFPKHLVAEGGWC TAWVGRG
P04169	GLPDGVEDLPKAEDDHAHDYVNDAADTDHARFQEGQLCENCQFWVDYVNGWGYCQHPDFTDVLV RGEGWCSVYAPA
P23873	MMSFQKIYSPTQLANAMKLVRQQNGWTQSELAKKIGIKQATISNFENNPDNTTLTTFFKILQSLELSM TLCDAKNASPESTEQQNLEW
B3EBZ3	SAPANAVSADDATAIALKYNQDATKSERVSAARPGLPPEEQHCANCQFMQADAAGATDEWKGCQLF PGKLINVNGWCASWTLKAG
P00264	QDLPPLDPSAEQAQALNYVKDTAEAADHPAHQEGEQCDNCMFFQADSQGCQLFPQNSVEPAGWCQS WTAQN
P80882	AAPLVAETDANAKSLGYVADTTKADKTKYPKHTKDQSCSTCALYQGKTAPQGACPLFAGKEVVAKG WCSAWAKKA
Q07558	MFSLKLFVVFLAVCICVSQAVSYTDCTESGQNYCLCVGGNLCGGGKHCEMDGSGNKCVDGEGTPKP KSQTEGDFEEIPDEDILN
P0A5B1	MQQSLAVKTFEDLFAELGDRARTRPADSTTVAALDGGVHALGKKLLEEAGEVWLAAEHESNDALAE EISQLLYWTQVLMISRGLSLDDVYRKL
Q9JLS0	MKFMLNLYVLGIMLTLLSIFVRVMESLGGLLESPLPGSSWITRGQLANTQPPKGLPDHPSRGVQ
P85219	MSGIVEAISNAVKSGLDHDWVNMGTSIADVVAKGADFIAGFFS
P06116	SNTRNFVLRDEEGNEHGVFTGKQPRQAALKAANRGDGTKSNPDVIRLRERGTKKVHVFKAWKEMV EAPKNRPDWMPEKISKPFVKKEKIEKIE
P48781	MGELPIAPIGRIIKNAGAERVSDDARIALAKVLEEMGEEIASEAVKLAKHAGRKTIKAEDIELARKMFK
Q23794	MSDSPVKKGRGRPAKAKPEETASPKAAKKEEKKVEEVPKKIEESTKPENGAAPKKGRGRPSKGDKA APKRPASGKGRGRPAKNAKKVDDADTEEVNSSD
P11873	AKSKDDSKPAPPKRPLSAFFLFKQHNYEQVKKENPNAKITELTSMIAEKWKAVGEKEKKKYETLQSE AKAKYEKDMQAYEKKYGKPEKQKKIKKNKKGSK
O97980	MEEQPECREEKRGSLHVWKSELVEVEDDVYLRHSSSLTYRL
P0C7T4	MEIFIEVFSHFLLQLTELTLNMCLELPTGSLEKSLMISSQVLQIPVANSTKQR
Q59041	MLPKATVKRIMKQHTDFNISAEAVDELCNMLEEIIKITTEVAEQNARKEGRKTIKARDIKQCDDERLK RKIMELSERTDKMPILIKEMLNVITSEL
P0CH77	FECSVSCEIEKEGNKDCKKKKCKGGWKCKFNMCVKDI
P37305	MPQKYRLLSLIVICFTLLFFTWMIRDSLCELHIKQESYELAAFLACKLKE
P0ACG4	MKQHKAMIVALIVICITAVVAALVTRKDLCEVHIRTGQTEVAVFTAYESE
Q9BPY8	MSAETASGPTEDQVEILEYNFNKVDKHPDSTTLCLIAAEAGLSEEETQKWFKQRLAKWRRSEGLPSEC RSVTD
O94483	MRRSTIVPTSRTSSSSPSPSQMKSFQMNRLVDQLSKLQSNMSHLENHLHVTAIQAEAIRRLGALQASLL MASGRVMSEARIQKSSDAVEEDVPM
Q84F15	MRKHSDCMNFCAVDATKGICRLSKQMINLDDAACPEIKVMPKCKNCKNFVEANDEGIGKCVGLEKED WVYSTLNAITCEGHVFNE
P0AFX0	MQLNITGNNVEITEALREFVTAKFAKLEQYFDRINQVYVVLKVEKVTHTSDATLHVNGGEIHASAEGQ DMYAAIDGLIDKLARQLTKHKDKLKQH
P0A0V6	MAHHEEQHGGHHHHHHHTHHHHYHGGEHHHHHHSSHHEEGCCSTSDSHHQEEGCCHGHHE
P24337	ALITRPSCPDLSICLNILGGSLGTVDDCCALIGGLGDIEAIVCLCIQLRALGILNLNRNLQLILNSCGRSYPS NATCPRT
O00198	MCPCPLHRGRGPPAVCACSAGRLGLRSSAAQLTAARLKALGDELHQRTMWRRRARSRRAPAPGALP TYWPWLCAAAQVAALAAWLLGRRNL
P69852	MNANKQRQYNQLAHELRELQTNLQETTKQLDIMSKQCNENLVGQLGKVHGSWLIGSYIYYMEQMLG KTQ

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P01559	MKKLMLAIFISVLSFPSFSQSTESLDSSKEKITLETKKCDVVKNNSEKKSENMNNTFYCCELCCNPACAGCY
O50319	MKKIVFVLTLMLFSFGTLGQETASGQVGDVSSSTIATEVSEAECGTQSATTQGENDWDWCCELCCNPACFGC
P22542	MKKNIAFLLASMFVFSIATNAYASTQSNKKDLCEHYRQIAKESCKKGFLGVRDGTAGACFGAQIMVAAKGC
P0A4M3	MRNLFIALMLLFSSIAFSQTVENNKKTVQQPQQIESKVNIKKLSENEECPFIKQVDENGNLIDCCEICCNPACFGCLN
Q17128	MNHLVKVLIVVVAIALVLCEAQVNFSPGWGTGKRSAVQDSPCKGSAESLMYIYKLVQNEAQKILECEKFSSN
Q9URQ5	MSQNNTISSMNPERAYNNVTLKNLTAFQLLSQRENICELLNLVESTERHNSIINPERQRMSLEEMKKMLDALKNERKK
Q6Q5K6	MTMDQGLNPKQFFLDDVVLQDTLCSMSNRVNKSVKTGYLFPKDHVPSANIIAVERRGGLSDIGKNTSN
P18527	EVKLVESGGGLVKPGGSLKLSCAASGFTFSSYAMSWVRQTPEKRLEWVASISSGGSTYYPDSVKGRFTISRDNARNILYLQMSSLRSEDTAMYYCAR
P01155	ARYGKSPYLYPLGY
P52754	MKFFAIAALFAAAAVAQPLEDRSNGNGNVCPPGLFSNPQCCATQVLGLIGLDCKVPSQNVYDGTDFRNVCAKTGAQPLCCVAPVAGQALLCQTAVGA
P84954	LRPAVIRPKGK
P0AAM3	MCIGVPGQIRTIDGNQAKVDVCGIQRDVDLTLVGSCDENGQPRVGQWVLVHVGFAMSVINEAEARDTLDALQNMFDVEPDVGALLYGEEK
P84189	AEIDFSGIPEDIIKQIKETNAKPPA
P12924	MVDAITVLTAIGITVLMLLMVISGAALIVKELNPNDIFTMQSLKFNRAVTIFKYIGLFIYIPGTIILYATYVKSLLMKS
P01093	DAGNRIAAPACVHFTADWRYTFVTNDCSIDYSVTVAYGDGTDVPCRSANPGDILTFPGYGTRGNEVLGAVLCATDGSA
P20596	SDASEPAPACVVMYESWRYTTAANNCADTVSVSVAYQDGATGPCATLPPGAVTTVGEGYLGEHGHPDHLALCPSS
P80403	CIPKWNRCGPKMDGVPCCEPYTCTSDYYGNCS
P01000	EIYFEPDFGFPPDCKVYTEACTREYNPICDSAAKTYSNECTFCNEKMNNDADIHFNHFGECEY
P26461	MAFFSSRVRALFILVLVLPLCSETGFARSKKTRKEPDCDVYRSHLFFCTREMDPICGTNGKSYANPCIFCSEKLGRNEKFDFGHWGHCREYTSAARS
P01066	EASSSSDDNVCCNGCLCDRRAPPYFECVCVDTFDHCPASCNSCVCTRSNPPQCRCTDKTQGRCPVTECRS
P07679	GDEKRPWECCDIAMCTRSIPPICRCVDKVDRCSDACKDCEETEDNRHVCFDTYIGDPGPTCHDD
P19860	SGERPWKCCDLQTCTKSIPAFCRCRDLLEQCSDACKECGKVRDSDPPRYICQDVYRGIPAPMCHEHQ
P19399	DQVRKCLSDTDCTNGEKCVQKNKICSTIVEIQRCEKEHFTIPCKSNNDCQVWAHEKICNKLPWGL
P83039	ELDEGNGLRRPVCGEMAELLACPLVNLPVCGTDGNTYANECLLCVQKMKTRQDIRILNNGRCRDT
P16064	QEQGTNPSQEQNVPLPRNYKQALETNTPTKTSWPELVGVTAEQAETKIKEEMVDVQIQVSPHDSFVTADYNPKRVRLYVDESNKVTRTPSIG
P01053	MSSVEKKPEGVNTGAGDRHNLKTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAQVPRVG
P16062	MSSMEGSVLKYPEPTEGSIGASSAKTSWPEVVGMSAEKAKEIILRDKPNAQVEVIPVDAMVHLNFDPNRVFVLVAVARTPTVG
P01052	KEFECDGKLQWPELIGVPTKLAKEIIEKQNSLISNVHILLNGSPVTMDFRCNRVRLFDDILGSVVQIPRVA
P01051	TEFGSELKSFPEVVGKTVDQAREYFTLHYPQYDVYFLPEGSPVTLDLRYNRVRVFYNPGTNVVNHVPHVG
P82381	SRRCPGKNAWPELVGKSGNMAAATVERENRNVHAIVLKEGSAMTKDFRCDRVWVIVNDHGVVTSVPHIT
P83472	MKLLFAIVALLALAFLCADISAVKTSWPELVGETLEEAKAQILEDRPDAVIKVQPEHSPVTYDYRPSRVIIFVNKDGNVAETPAAG
P03170	MSWALEMADTFLDTMRVGPRTYADVRDEINKRGREDREAARTAVHDPERPLLRSPGLLPEIAPNASLGVAHRRTGGTVTDSPRNPVTR
Q4WZ11	MKFSLACLLALAGLQAALADPATCEKEAQFVKQELIGQPYTDAVANALQSNPIRVLHPGDMITMEYIASRLNIQVNENNEIISAHCA
P10296	EEERICPLIWMECKRDSDCLAQCICVDGHCG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q50298	MQPKFNNQAKQDKLVLTGKILEIIHGDKFRVLLENNVEVDAHLAGKMRMRRLRILPGDLVEVEFSPYD LKLGRIIGRK
P0C7I2	MKKRSVSGCNITILAVVFSHLSAGNSPCGNQANVLCISRLEFVQYQS
Q51472	MGALTKAEIAERLYEELGLNKREAKELVELFFEEIRQALEHNEQVKLSGFGNFDLRDKRQRPGRNPKT GEEIPITARRVVTFRPGQKLKARVEAYAGTKS
Q06607	MIRSELIAKIAEENPHLFQRDVEKIVNTIFEEIIEAMARGDRVELRGFGAFSVKKRDARTGRNPRTGTS VAVDEKHVPFFKTGKLLRDRLNGGEE
P83579	VIEPDCKKYEGKKCPPDIALVCGTNGREYYNECALCVFIRDSTLKADKAIKIKKWGKC
P83037	NHHDDCRDLGISIDDQQNRLLVVKNGDPLVVQFAKANRGGDD
P0ADG1	MMQHQVNVSARFNPETLERVLRVVRHRGFHVCSMNMAAASDAQNINIELTVASPRSVDLLFSQLNKL VDVAHVAICQSTTTSQQIRA
P0ADF8	MQNTTHDNVILELTVRNHPGVMTHVCGLFARRAFNVEGILCLPIQDSDKSHIWLLVNDDQRLEQMISQ IDKLEDVVKVQRNQSDPTMFNKIAVFFQ
C1P619	MNNSTKFCFSRFRTGN
P82706	MKFFSVVTVFVLGLLAVANAVPLSPDPGNVIINGDCRVCNVHGGK
P82705	MKFFQAAALLLAMFAALANAEPVPQPGTVLIQTDNTQYIRTG
P83869	MNCLKICGFFFALIAALATAEAGTQVIHAGGHTLIQTDRSQYIRKN
P82701	MKLIALCCLLLLGLLGFLAAPGVASPSRHTGPGNGSGSGAGSGNPFRSPSSQQRPLYYDAPIGKPSKTM YA
Q9ZHG0	MTLLSFGFSPVFFSVMAFCIISRSKFYPQRTRNKVIVLILLTFFICFLYPLTKVYLVGSYGIFDKFYLFCFI STLIAIAINVVILTINGAKNERN
P11424	FKNPECGEPHSLDGSPNGISCRGYFPSWSYNPDAQQCVSFVYGGCGGNNNRFGSQNECEERCI
Q06578	MKSKISEYTEKEFLEFVEDIYTNNKKKFPTEESHIQAVLEFKKLTEHPSGSDLLYYPNENREDSPAGVV KEVKEWRASKGLPGFKAG
P02984	MGLKLDLTWFDKSTEDFKGEEYSKDFGDDGSVMESLGVPFKDNVNNGCFDVIAEWVPLLQPYFNHQI DISDNEYFVSFDYRDGDW
P13476	MKLSPKAAIEVCNEAAKKGLWILGIDGGHWLNPGFRIDSSASWTYDMPEEYKSKIPENNRLAIENIKDD IENGYTAFIITLKM
Q03708	MELKNSISDYTEAEFVQLLKEIEKENVAATDDVLDVLLEHFVKITEHPDGTDLIYYPSDNRDDSPEGIV KEIKEWRAANGKPGFKQG
P11899	MNKMAMIDLAKLFLASKITAIEFSERICVERRRLYGVKDLSPNILNCGEELFMAAERFEPDADRANYEI DDNGLKVEVRSILEKFKL
P13190	VPTQRLCGSHLVDALYFVCGERGFFYSPKQIRDVGPLSAFRDLEPPLDTEMEDRFPYRQQLAGSKMK RGIVEQCCHNTCSLVNLEGYCN
P68988	SALTGAGGTHLCGSHLVEALYVVCGDRGFFYTPSKTGIVEQCCHRKCSIYDMENYCN
P58609	GADEDCLPRGSKCLGENKQCCEKTTCMFYANRCVGI
P56615	RICTNCCAGRKGCNYYSADGTFICEGESDPNNPKACPRYCDTRIAYSKCPRSEGN
P01094	MNTDQQKVSEIFQSSKEKLQGDAKVVSDAFKKMASQDKDGKTTDADESEKHNYQEQYNKLKGAGHK KE
P01095	MTKNFIVTLKKNTPDVEAKKFLDSVHHAGGSIVHEFDIIKGYTIKVPDVLHLNKLKEKHNDVIENVEED KEVHTN
P03718	MKTFKEFTSTTTPVSTITEATLTSEVIKANKGREGKPMISLVDGEEIKGTVYLGDGWSAKKDGATIVIS PAEETALFKAKHISAAHLKIIAKNLL
P03719	MKTYQEFIAEARVGAGKLEAAVNKKAHSFHDLPDKDRKKLVSLYIDRERILALPGANEGKQAKPLNA VEKKIDNFASKFGMSMDDLQQAAIEAAKAIKDK
Q3SX13	MTDVETTYADFIASGRTGRRNAIHDILVSSASGNSNELALKLAGLDINKTEGEEDAQRNSTEQSGEAQG EAAKSES
P27775	MTDVESVISSFASSARAGRRNALPDIQSSLATGGSPDLALKLEALAVKEDAKMKNEEKDQGQPKKPLD EDK
Q9Y2B9	MMEVESSYSDFISCDRTGRRNAVPDIQGDSEAVSVRKLAGDMGELALEGAEGQVEGSAPDKEAGNQP QSSDGTTSS
P0AAN9	MKNLIAELLFKLAQKEEESKELCAQVEALEIIVTAMLRNMAQNDQQRLIDQVEGALYEVKPDASIPDD DTELLRDYVKKLLKHPRQ
P50500	MSEKDKMITRRDALRNIAVVVGSVATTTMMGVGVADAGSMPKAAVQYQDTPKGKDHCSVCAQFIAP HSCKVVAGNISPNGWCVAFVPKSA
P10832	DKPTTKPICEQAFGNSGPCFAYIKLYSYNQKTKKCEEFIYGGCKGNDNRFDTLAECEQKCIK
P0C0L9	MGLKWTDSREIGEALYDAYPDLDPKTVRFTDMHQWICDLEDFDDDPQASNEKILEAILLVWLDEAE

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
B2C6F2	MDKITNTYSALLSSITNFKNRNIRAYFKRKAFDEYLECVNGVRPVEKYLKDNEELKGVMDRQSVIYNFYED
Q6Q560	MPGFTAPTRRQVLSLYKEFIKNANQFNNYNFREYFLSKTRTTFRKNMNQQDPKVLMNLFKEAKNDLGVLKRQSVISQMYTFDRLVVEPLQGRKH
Q9PSM2	EPDSAADTGTEAACSNYDLKKGCAKIFDPVCGTDNILYSNECLLCFQNLQRKTNVRIKRRGTCQEPSPR
P20155	MALSVLRLALLLLAVTFAASLIPQFGLFSKYRTPNCSQYRLPGCPRHFNPVCGSDMSTYANECTLCMKIREGGHNIKIIRNGPC
P09036	MKVAVIFLLSALALLSLAGNTFSAKVTGKEASCHDAVAGCPRIYDPVCGTDGITYANECVLCFENRKRIEPVLIRKGGPC
P37109	MAVRLWVVALALAALFIVDREVPVSAEKQVFSRMPICEHMTESPDCSRIYDPVCGTDGVTYESECKLCLARIENKQDIQIVKDGEC
P01001	MKTSGVFLLLSLALFCFFSGVFGQGAQVDCAEFKDPKVYCTRESNPHCGSDGQTYGNKCAFCKAVMKSGGKINLKHRGKC
Q5DT21	MRATAIVLLLALTLATMFSIECAKQTKQMVDCSHYKKLPPGQQRFCHHMYDPICGSDGKTYKNDCFFCSKVKKTDGTLKFVHFGKC
Q9TWH3	VEVATVKNCGKKLLATPR
P84778	EDEECAKTDQICPPNAPNYCCSGSCVPHPRLRIFVCA
P34950	MASVAESSGVVEVIELISDGGNDLPRKIMSGRHGGICPRILMPCKTDDDCMLDCRCLSNGYCG
P16589	AHMDCTEFNPLCRCNKMLGDLICAVIGDAKEEHRNMCALCCEHPGGFEYSNGPCE
P01069	LLMPVKPNDDRVIGCWCISRGYLCGCMPCKLNDDSLCGRKG
B1P1E6	MKTSVLLVILGIAAITVQCTASESVEQDSLRTFVDAVLGWNAEMASEARCGGWMAKCADSDDCCETFHCTRFNVCGK
B1P1B0	MNATIFALLLLLNLAMYNAAEQSSETDMDDTLLIPEINRGRCIEEGKWCPKKAPCCGRLECKGPSPKQKKCTRP
B1P1H3	MRFHTLLFLSFLLLVSCALICTAQHPGLEKSGMFHENVGKGQHIEEKRSCIERMQTCEVEAGLPCCSGAPCICPYIGDCICIQ
P0CH53	TCEPSGKPCRPLMRIPCCGSCVRGKCA
B1P1F0	MKAAILLAFAGLALLSVICHASENVEQDSFEEVFSAIFAMEDDLKPKERVCRGYGLPCTPEKNDCCQRLYCSQHRLCSVKAGK
B1P1F1	MRALLIIAGLALFLVVCNASQVNEQRKLNEMLSVMFAVEEPQERDDCLGMFSSCNPDNDKCCEGRKCDRRDQWCKWNPW
B1P1D7	MKVSVLITLAVLGVMFVWASAAELEQSGSDQKDSDSPAWLKSMERIFQSEERDCRKMFGGCSKHEDCCAHLACKRTFNYCAWDGSFSK
P0C5X7	MKGSAFAIILGLVVLCACSFAEDEQDQFASPNELLRSMFLESRHELIPEVEGRYCQKWMWTCDSERKCCEGYVCELWCKYNLG
B1P1C7	MKTLVLFIIFGLAALFLLSSANELEETERGCGLLMDACDGKSTFCCSGYNCSPTWKWCVLDCPNLFLLPPTKTLC
P62520	MKNTSILFILGLALLLVLAFEAQVGESDGECGGFWWKCGRGKPPCCKGYACSKTWGWCAVEAP
P0CH44	ECTKLLGGCTKDSECCPHLGCRKKWPYHCGWDGTF
P69592	MSKGKKRSGARPGRPQPLRGTKGKRKGARLWYVGGQQF
Q9BKB4	MKIFFAILLILAVCSMAIWTVNGTPFAIKCATNADCSRKCPGNPPCRNGFCACT
P84777	MKAFYGMLVIFILCSTCYISVDSQIDTNVKCSGSSKCVKICIDRYNTRGAKCINGRCTCYP
Q9NBG9	MKIFSILLVALIICSISICTEAFGLIDVKCFASSECWTACKKVTGSGQGKCQNNQCRCY
Q95NJ8	MKFIIVLILISVLIATIVPVNEAQTQCQSVRDCQQYCLTPDRCSYGTCYCKTTGK
P0C1X6	MHFSGVAFILISMVLIGSIFETTVEAGEGPKSDCKPDLCEAACKDLGKPMDFCKDGTCKCKD
P0C183	GCTPEYCSMWCKVKVSQNYCVKNCKCPGR
Q9U8D2	MSRLYAIILIALVFNVVMTITPDMKVEAATCEDCPEHCATQNARAKCDNDKCVCEPK
P60209	VGCAECPMHCKGKMAKPTCENEVCKCNIGKKD
P0C8X9	MKNIAMKTTVVLTILLLSVLTAINADTMKKRSDYCSNDFCFFSCRRDRCARGDCENGKCVCKNCHLN
Q86QT3	MKVLILIMIIASLMIMGVEMDRDSCVDKSRCAKYGYYQECQDCCKNAGHNGGTCMFFKCKCA
Q9BKB7	MKISFVLLLTLFICSIGWSEARPTDIKCSESYQCFPVCKSRFGKTNGRCVNGFCDCF
Q0GY43	MVATNRCCVFALLVALLLIHSLAEAGKGKEVLGKIKNKLVEVKEKIKAGWDKLTSKSEYACPVIDKFCEDHCAAKNAIGKCDDFKCQCLNS
P38393	MIAHHFGTDEIPRQCVTPGDYVLHEGRTYIASANNIKKRKLYIRNLTTKTFITDRMIKVFLGRDGLPVKAESW

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P82851	GHACYRNCWREGNDEETCKERCG
P86110	MESSRKSYVLMLFLAFVIMNVCSVSGEPKDGEIAGFEMEEARYDACVNACLEHHPNVRECEEACKNP VPP
P0DJ34	MKTSGTVYVFLLLLAFGIFTDISSACSEQMDDEDSYEVEKRGNACIEVCLQHTGNPAECDKACDK
P0DJ36	MKSTLMTASVLILVLLSIVDYASVYAEFIDSEISLERQWINACFNVCMKISSDKKYCKYLCGKN
P0DJ41	MKLLPLLFVILIVCAILPDEASCDQSELERKEENFKDESREIVKRSCKKECSGSRRTKKCMQKCNREHG HGR
Q25297	MATTYEEFSAKLDRLGEEFNRKMQEQNAKFFADKPDESTLSPEMKEHYEKFERMIKEHTEKFNKKM HEHSEHFKQKFAELLEQQKAAQYPSK
P83056	MRLWFCLSFLIILCVEHFPGTLAVERNVPESEEKTEQFLRDLFEISRLQRRPAGFTPFRGKFHSQSLRGL SETKRIYNAIWPCKHCNKCKPGLLCKK
P02447	MACCAPRCCSVRTGPATTICSSDQFCRCGVCLPSTCPHDISLLQPTFCDNSPVPYHVPDTYVPTCFLLN SSHPTPGLSGINLTTFIQPGCENACEPRC
P02451	ACNDLCGPCGPTPLANSCNEPCVRQCEASRVVIQPSTVVVTLPGPILSSFPQSTAVGGSASSSVGNELL ASQGVPYFGGGFGLGGLGCFSGRRGCYPC
P62584	MSEYMKNEILEFLNRHNGGKTAEIAEALAVTDYQARYYLLLLEKAGMVQRSPLRRGMATYWFLKGE KQAGQSCSSTT
P60985	MKIPVLPAVVLLSLLVLHSAQGATLGGPEEESTIENYASRPEAFNTPFLNIDKLRSAFKADEFLNWHAL FESIKRKLPFLNWDAFPKLKGLRSATPDAQ
Q9UNZ5	MAQGQRKFQAHKPAKSKTAAAASEKNRGPRKGGRVIAPKKARVVQQQKLKKNLEVGIRKKIEHDVV MKASSSLPKKLALLKAPAKKKGAAAATSSKTPS
P14269	MALTCRLRFPVPGFRGRMHRRRGMAGHGLTGGMRRAHHRRRRASHRRMRGGILPLLIPLIAAAIGAV PGIASVALQAQRH
P0C5F3	FGESCIAGRFIVPLGQQVTDQRDCALYKCVNYNKKFALETKRCATVNLKSGCKTVPGGAGAAFPSCC PMVTCK
O88038	MNLFDLQSMETPKEEAMGDVETGSRASLLLCGDSSLSITTCN
P80060	MKFALALCAAVLLVVLVQAEEKCTPGQVKQQDCNTCTCTPTGVWGCTRKGCQPAKREISCEPGKTF KDKCNTCRCGADGKSAACTLKACPNQ
P82384	MKFVIVLACLLAVVFANEEADVVKSDSEVNLLDFNYAYELSNHIRAVQTGALKEHDNWVVSGEYEYV APNGKTVKVVYTADETGYHPKVVEA
P84870	QKACTMDYFPVCCQIVLGGGAYTAGNPCSCQGFVASKGTCDNPYPCPCTTEAQFPVCCTTQWGLVS ATGNCACGCMGGVPVSDGPCPEVY
P85888	FGPGCGPSTFSCTSPQKILPGSSVSFPSGYSSIYLTTESGSASVYLDRPDGFWVGGADSRGCSNFGGFN GNGDSKVGNWGDVPVAAWACN
P12308	VTSYTLNEVVPLKDVVPEWVRIGFSATTGAEFAAHEVLSWSFHSELGGTSASKQS
P18675	NGPNGKSQSIIVGPWGDRVTN
O43261	MRPCIWIHVHLKPPCRLVELLPFSSALQGLSHLSLGTTLPVILPERNEEQNLQELSHNADKYQMGDCCK EEIDDSIFY
P80588	SAPAQWKLWLVMDPRTVMIGTAAWLGVLALLIHFLLLGTERFNWIDTGLKEQKATAAAQAAITPAP VTAAAK
P51756	MWRIWRLFDPMRAMVAQAVFLLGLAVLIHLMLLGTNKYNWLDGAKKAPVATAVAPVPAEVTSLAQ AK
P80103	MWKLWKFVDFRMTAVGFHIFFALIAFAVHFACISSERFNWLEGAPAAEYYMDENPGIWKRTSYDG
P95655	MNNAKMWLVVKPTVGIPLFLVACAIASFLVHLMLVLTTGWMGDYYSGSFEAASLVSNATTLLS
P77799	MNQGKVWRVVKPTVGVPVYLGAVAVTALILHGGLLAKTDWFGAYWNGGKKAAAAAAAVAPAPVA APQAPAQ
P35104	MNQGRIWTVVKPTVGLPLLLGSVAIMVFLVHFAVLTHTTWVAKFMNGKAAAIESSIKAV
P07503	MQPRSPVRTNIVIFTILGFVVALLIHFIVLSSPEYNWLSNAEGGALLLSAARALFGI
P97253	MSNPKDDYKIWLVINPSTWLPVIWIVATVVAIAVHAAVLAAPGFNWIALGAAKSAAK
P80586	MTNGKIWLVVKPTVGLPIGMLFAALLAVLIHGLLFVDGRLKSWWSEFPVAKPAVVSVQAAPAPVAAE VK
P04123	MATEYRTASWKLWLILDPRRVLTALFVYLTVIALLIHFGLLSTDRLNWWEFQRGLPKAASLVVVPPAV G
P11696	TDIRTGLTDEECQEIHEMNMLGMHAYWSIGLIANALAYAWRPFHQGRAGNRLEDHAPDYVRSALT
P09927	MRDDDDLVPPKWRPLFNNQDWLLHDIVVKSFYGFGVIAAIAHLLVYLWKPWLP
Q2RQ23	MAEVKQESLSGITEGEAKEFHKIFTSSILVFFGVAAFAHLLVWIWRPWVPGPNGYSALETLTQTLTYLS
P80587	ADANKVWPTGLTVAEAEELHTYVTNGFRVFVGIAVVAHVLVFAAHPWGRGGALVA

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P04124	MADLKPSLTGLTEEEAKEFHGIFVTSTVLYLATAVIVHYLVWTARPWIAPIPKGWVNLEGVQSALSYL V
P04126	YFAADGSVVPSISDWNLWVPLGILGIPTIWIALTYR
P08948	MSAVPFTRVLLISGFLAHLLLSTFVTLTVCKEVTEESDDLSKRNVLQRQLWAVGSLMGKKSLENTNRR SDEDMEISALFRGSPLKVKRSD
Q38162	MKKLFLAMAVVLLSACSTFGPKDIKCEAYYMQDHVKYKANVFDRKGDMFLVSPIMAYGSFWAPVSY FTEGNTCEGVF
P38497	AHRFAAEDFGALDTA
P22395	GAVPAAQFSPRL
P41490	RLHQNGMPFSPRL
P08497	MKKAERGASPKQIKPHRYYLLAVH
A8DYP9	MIRIISRANSVTSSNEVNRLVTGQIPHD
P20404	QDSGDGWPQQPFVPRL
Q9HKT2	MVLNYTMHMMYSKNWKAKKGLIRVTLDLDGNRIKDIHISGDFFMFPEDSINRLEDMLRGSSIEKINDII RDFYNQGVITPGVEPEDFIQALRVI
P0CI69	MDPEPTPLPRWRIFLFR
P17323	MKRLFLSFVALALLAGSIAACGQKGPLYLPDDEKAKKEHSKDRYGF
P69776	MKATKLVLGAVILGSTLLAGCSSNAKIDQLSSDVQTLNAKVDQLSNDVNAMRSDVQAAKDDAARANQ RLDNMATKYRK
P0AD92	MKAIFVLKGWWRTS
Q9Y333	MLFYSFFKSLVGKDVVVELKNDLSICGTLHSVDQYLNIKLTDISVTDPEKYPHMLSVKNCFIRGSVVRY VQLPADEVDTQLLQDAARKEALQQKQ
Q9Y7M4	MESAQAVAEPLDLVRLSLDEIVYVKLRGDRELNGRLHAYDEHLNMVLGDAEEIVTIFDDEETDKDKAL KTIRKHYEMLFVRGDSVILIAPPRN
P57743	METPLDLLKLNLDERVYIKLRGARTLVGTLQAFDSHCNIVLSDAVETIYQLNNEELSESERRCEMVFIR GDTVTLISTPSEDDDGAVEI
P40089	MSLPEILPLEVIDKTINQKVLIVLQSNREFEGTLVGFDDFVNVILEDAVEWLIDPEDESRNEKVMQHHG RMLLSGNNIAILVPGGKKTPTEAL
Q06406	MSGKASTEGSVTTEFLSDIIGKTVNVKLASGLLYSGRLESIDGFMNVALSSATEHYESNNNKLLNKFNS DVFLRGTQVMYISEQKI
O00453	MLSRNDDICIYGGLGLGGLLLLAVVLLSACLCWLHRRVKRLERSWAQGSSEQELHYASLQRLPVPSSEG PDLRGRDKRGTKEDPRADYACIAENKPT
P84810	MMKLVLFGIIVILFSLIGSIHGISGNYPLNPYGGYYYCTILGENEYCKKICRIHGVRYGYCYDSACWCET LKDEDVSVWNAVKKHCKNPYL
P84809	MISVQVIFIAFISIIAFSMVCGGNVFPNRELGILYGCKGYGNAFCDKVCKMHLARGGGRCGEPNPVMW ACECIDIDEDNGYFLNALEKQCPLLKG
Q9NU23	MAASRLPPATLTLKQFVRRQQVLLLYRRILQTIRQVPNDSDRKYLKDWAREEFRRNKSATEEDTIRM MITQGNMQLKELEKTLALAKS
Q9HD34	MAASSRAQVLSLYRAMLRESKRFSAYNYRTYAVRRIRDAFRENKNVKDPVEIQTLVNKAKRDLGVIRR QVHIGQLYSTDKLIIENRDMPRT
P09181	MCGKILLILFFIMTLSACQVNHIRDVKGGTVAPSSSSRLTGLKLSKRSKDPL
Q980W8	MVNLKCPICGGEITVEDDALPGELVEHECGAQLEVVKQNGKLSLRLAEQIGEDWGE
P05780	MSLLTEVETPIRNEWGCRCNDSSDPLVIAANIIGILHLILWILDRLFFKCIYRRFKYGLKRGPSTEGVPES MREEYRKEQQNAVDVDDGHFVNIELE
Q9NQG1	MASDLDFSPPEVPEPTFLENLLRYGLFLGAIFQLICVLAIIVPIPKSHEAEAEPSEPRSAEVTRKPKAAVP SVNKRPKKETKKKR
P86916	ICNGQWTSVGSAGLYYTIKADSMCVDIHYTDGFIQPSCQGLQVIGPCNRYQNGPRDFVACQTSGGSGH PICIQSTNGNIELCANCYCPQ
P29399	MKTTPFFANLLASQTRELTENELEMTAGGTASQQSPVQEVPEQPFATMRYPSDSDEDGFNFPV
P42716	INWKKMAATALKMI
P30659	MKQILLISLVVVLAVFAFNVAEGCDATCQFRKAIDDCQKQAHHSNVLQTSVQTTATFTSMDTSQLPGN SVFKECMKQKKKEFSSGK
P0A5E9	MPFLVALSGIISGVRDHSMTVRLDQQTRQRLQDIVKGGYRSANAAIVDAINKRWEALHDEQLDAAYAA AIHDNPAYPYESEAERSAARARRNARQQRSAQ
P0CL59	MKTAISLPDETFDRVSRRASELGMSRSEFFTKAAQRYLHELDAQLLTGQIDRALESIHGTDEAEALAVA NAYRVLETMDDEW
Q8VJR2	MSTSTTIRVSTQTRDRLAAQARERGISMSALLTELAAQAERQAIFRAEREASHAETTTQAVRDEDREW EGTVGDGLG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P0CL61	MKLSVSLSDDDVAILDAYVKRAGLPSRSAGLQHAIRVLRYPTLEDDYANAWQEWSAAGDTDAWEQT VGDGVGDAPR
P0AE72	MIHSSVKRWGNSPAVRIPATLMQALNLNIDDEVKIDLVDGKLIIEPVRKEPVFTLAELVNDITPENLHEN IDWGEPKDKEVW
P0C7B4	MLSFSQNRSHSLEQSLKEGYSQMADLNLSLANEAFPIECEACDCNETYLSSNSTNE
E3YBA4	MTVKIAQKKVLPVIGRAAALCGSCYPCSCM
P63452	MTSSPSTVSTTLLSILRDDLNIDLTRVTPDARLVDDVGLDSVAFAVGMVAIEERLGVALSEEELLTCDT VGELEAAIAAKYRDE
P01499	MISMLRCTFFFLSVILITSYFVTPTMSIKCNCKRHVIKPHICRKICGKNG
P04567	MCICKNGKPLPGFIGKICRKICMMQQTH
O86200	MSYKKLYQLTAIFSLPLTILLVSLSSLRIVGEGNSYVDVFLSFIIFLGFIELIHGIRKILVWSGWKNGS
P81511	MFLLVFLCCLHLVISSHTPDESFLCYQPDQVCCFICRGAAPLPSEGECNPHPTAPWCREGAVEWVPYS TGQCRTTCIPYVE
P86912	FHVPDDRPCINPGRCPLVPDATCTFVCKAADNDFGYECQHVWTFEGQRVGCYA
P01076	MAQKFTILFTILLVVIAAQDVMAQDATLTKLFQQYDPVCHKPCSTQDDCSGGTFCQACWRFAGTCGP YVGRAMAIGV
O32712	MEQITLSFPASRALSGRALAGVVGSGDMEVLYTAAQSATLNVQITTSVDNSQARWQALFDRLNLINGL PAGQLIIHDFGATPGVARIRIEQVFEEAAHA
O60200	MGNIMSASFAPECTDLKTKYDSCFNEWYSEKFLKGKSVENECSKQWYAYTTCVNAALVKQGIKPALD EAREEAPFENGGKLKEVDK
Q61845	MATSDVKPKSISRAKKWSEEIENLYRFQQAGYRDEIEYKQVKQVAMVDRWPETGYVKKLQRRDNTF FYYNKERECEDKEVHKVKVYVY
P83441	NPEDWFTPDTCAYGDSNTAWTTCTTPGQTCYTCCSSCFDVVGEQACQMSAQC
P83235	DIEDFYTSETCPYKNDSQLAWDTCSGGTGNCGTVCCGQCFSFPVSQSCAGMADSNDCPNA
P12350	MNKLAILAIIAMVLFSANAFRFQSRIRSNVEAKTETRDLCEQSALQCNEQGCHNFCSPEDKPGCLGMV WNPELCP
P26887	DECANAAAQCSITLCNLYCGPLIEICELTVMQNCEPPFS
P58548	DICDIAIAQCSLTLCQDCENTPICELAVKGSCPPPWS
P13113	MKKLFASLAIAAVVAPVWAATQTVTLSVPGMTCSACPITVKKAISKVEGVSKVNVTFETREAVVTFD DAKTSVQKLTKATEDAGYPSSVKK
P34166	MQPITTASTQATQKDKSSEKKDNYIIKGLFWDPACVIA
P34069	MDSIATNTHSSSIVNAYNNNPTDVVKTQNIKNYTPKVPYMCVIA
P64512	MKKFRWVVLVVVVLACLLLWAQVFNMMCDQDVQFFSGICAINQFIPW
P55846	RVINDDCPNLIGNRDLYKKVEWICEDCSNIFRNTGMATLCRKNCFFNEDFLWCVYATERTEEMSQLR QWVGILGAGRE
P0A734	MALLDFFLSRKKNTANIAKERLQIIVAERRRSDAEPHYLPQLRKDILEVICKYVQIDPEMVTVQLEQKD GDISILELNVTLPEAEELK
O25099	MSLFDFFKNKGSAATATDRLKLILAKERTLNLPYMEEMRKEIIAVIQKYTKSSDIHFKTLDSNQSVETIE VEIILPR
P81077	DEVDPDGKVLNSLIDTLMHLQKEFANLKYAFLTVHKARSFGSGSERLYVTNKEIKNFEPLGDI
P64602	MSESLSWMQTGDTLALSGELDQDVLLPLWEMREEAVKGITCIDLSRVSRVDTGGLALLLHLIDLAKKQ GNNVTLQGVNDKVYTLAKLYNLPADVLPR
P01207	DGDDYKFGHFRWSVPL
Q54GL7	MSDEKTQLIEAFYNFDGDYDGFVSVEEFRGIIRDGLPMTEAEITEFFEAADPNNTGFIDYKAFAAMLYS VDES
P41989	YVMGHFRWDKF
P60694	MNHNVIIVIALIIVVISMLAMLIRVVLGPSLADRVVALDAIGLQLMAVIALFSILLNIKYMIVVIMMIGILAF LGTAVFSKFMDKGKVIEHDQNHTD
P30748	MIKVLFFAQVRELVGTDATEVAADFPTVEALRQHMAAQSDRWALALEDGKLLAAVNQTLVSFDHPL TDGDEVAFFPPVTGG
Q7A441	MKVLYFAEIKDILQKAQEDIVLEQALTVQQFEDLLFERYPQINNKKFQVAVNEEFVQKSDFIQPNDTV ALIPPVSGG
Q9Y8C2	MSTFQIHYFASASTYTGRNTESLPAPLPLSSLFDTLEAKYPGIKEKVLSSCSISLGDEYVDLVSDGEKSG NEGLLIQGGDEVAIIPPVSSG
O96033	MVPLCQVEVLYFAKSAEITGVRSETISVPQEIKALQLWKEIETRHPGLADVRNQIIFAVRQEYVELGDQ LLVLQPGDEIAVIPPISGG
P02881	FREIKGYEYQLYVYASDKLFRADISEDYKTRGRKLLRFNGPVPPP

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P02882	GEWEIIDIGPFTQNLGKFAVDEENKIGQYGRLTFNKVIRPCMKKTIYEEN
P08854	MSISARNQLKGKVVGLKKGVVTAEVVLEIAGGNKITSIISLDSVEELGVKEGAELTAVVKSTDVMILA
Q502X0	MNGFGRLEHFSGAVYEGQFKDNMFHGLGTYTFPNGAKYTGNFNENRVEGEGEYTDIQGLEWSGNFH FTAAPDLKLKLHM
P69207	QPPGGSKVILF
Q5TGZ0	MSESELGRKWDRCLADAVVKIGTGFGLGIVFSLTFFKRRMWPLAFGSGMGLGMAYSNCQHDFQAPY LLHGKYVKEQEQ
Q96VH5	MSEQAQTQQPAKSTPSKDSNKNGSSVSTILDTKWDIVLSNMLVKTAMGFGVGVFTSVLFFKRRAFPV WLGIGFGVGRGYAEGDAIFRSSAGLRSSKV
Q9NRJ1	MGECPHLVDVRLGHRSLATGPEQSDICHTGSEARWTTTWYGSLSFSRHKYKMLADLTPGVEMSCRH WARWLTPVIPALWKAEAGGLPELRSSRPAWTTW
P19863	FVPIFTHSELQKIREKERNKGQ
P0C777	MDIEPEVPVVEKQMLAGNRGKQKTRRSVAKDAIRKPASDSTNGGNWVNVADKIEVHIHFNF
Q89682	MDSQRTVELTNPRGRSKERGDSGGKQKNSMGRKIANDAISESKQGVMGASTYIADKIKVTINFNF
P17461	MDPERIPYNSLSDSDATGKRKKGGEKSAKKRLVASHAASSVLNKKRNEGSASHGGTWVIVADKVEVSI NFNF
Q89846	MACCRCDSSPGDYSGALLILFISFVFFYITSLSPQGNTYVHHFDSSSVKTQYVGISTNGDG
Q7TD19	MKVLLVTGVLGLLLLIKWKSQSTSTSNQTCQCPTSPWVIYAFYNSLSLVLLLCHLIPEIKPIHTSYNTHD SSKQQHISINTGNGK
P89035	MATGKCYCPEDPRVGPLLVLCLLLLLILFSRSWNVAPVVVPSYHTVYHHEKYQNIEIQK
P43174	MNNEKNVSFEFIGSTDEVDEIKLLPCAWAGNVCGEKRAYCCSDPGRYCPWQVVCYESSEICSQKCGK MRMNVTKNTI
P14947	AAPVEFTVEKGSDEKNLALSIKYNKEGDSMAEVELKEHGSNEWLALKKNGDGVWEIKSDKPLKGPFN FRFVSEKGMRNVFDDVVPADFKVGTTYKPE
P10414	MKNIFMLTLFILIITSTIKAIGSTNEVDEIKQEDDGLCYEGTNCGKVGKYCCSPIGKYCVCYDSKAICNK NCT
P34394	MTFPKENINNYASSSFNYKYLIEEFLQWSSSCNRISYPTKYFDFLLVIIMTFVFLCLVYLLIWIEVLLNSSR PDPKHRVYYPRTNFWQRTD
Q46865	MEKRTPHTRLSQVKKLVNAGQVRTTRSALLNADELGLDFDGMCNVIIGLSESDFYKSMTTYSDHTIW QDVYRPRLVTGQVYLKITVIHDVLIVSFKEK
O75012	MSGKPPVYRLPPLPRLKVKKPIIRQEANKCLVLMSNLLQCWSSYGHMSPKCAGLVTELKSCTSESALG KRNNVQKSNINYHAARLYDRINGKPHD
Q9RGZ0	MFQSILMIVLVVMSISLFVCFIRTLIGPTMSDRIVALDTFGINLIGFIGVIMMLQETLAYSEVVLVISILAFI GSIALSKFIERGVVFDRG
A7VN15	MKVFFILIIFSFTLATCQGECYGSVPLPIDGEDVPLRTCVDTHDGQKHLIVSTWKTANSFSCECTQIGLQ CCQKYVAVA
D2X5V5	SCSRKGPKFVEYKYMAGSAQYCEHKNMKFMIGSNFIDFDDCTRCTCYNHGLQCCGIGANAGVFGVP GCEAVNDHCELVFLKKNTDQLCFIN
P83242	YCFQKINRPGESDEGCILDGKLYPFGEISRTENCYRCSCSRDAMRCCTLFHTPVGYNKEKCKVVFNKE SCNYDVVQKDDPSKECFVYSRV
P56277	MPQKDPCQKQACEIQKCLQANSYMESKCQAVIQELRKCCAQYPKGRSVVCSGFEKEEEENLTRKSAS K
Q59584	MEMLPLVKIAPEYNLTLDPSTGMIGAALGREVIILSMDEINEQIAALEATADDLINSLDPTTIPEGSYPG REGVYLTAGKLTNIVYGFILGLIILFALLL
P80654	MAEEHEKGVPMVLAPQMGAIDATVESIRYRAQLIARNQKLDSGVAATGIIGFAAGFLFSLLMVIVLPVA VGL
Q50773	MIILSNKPNIRGIKNVVEDIKYRNQLIGRDGRLFAGLIATRISGIAIGFLLAVLLVGVPAMMSILGVI
Q50774	MSEEEKTTIPRVLVSADEFNKANEKLDEIEEKVEFTVGEYSQRIGQQIGRDIGILYGIVIGLIILAVTNILF AGLLKGLLKSLFGL
P30331	MTSTTLVKCACEPCLCNVDPSKAIDRNGLYYCSEACADGHTGGSKGCGHTGCNCHG
Q09902	MGKLVKHCHTSLHSELVILFYAKNRFCISIDHLRPLKSHRTHGHYCLLNFSLRENKNYLIIVYLPIEGFSA NHMCISHGTSFNIK
Q96DR8	MKFLAVLVLLGVSIFLVSAQNPTTAAPADTYPATGPADDEAPDAETTAAATTATTAAPTTATTAAST TARKDIPVLPKWVGDLPNGRVCP
Q08542	MDCVLRSYLLLAFGFLICLFLFCLVVFIWFVYKQILFRTTAQSNEARHNHSTVV
P46979	GFKDGAADRISHGF
P0CI28	MRVIRMTNYEAGTLLTCSHEGCGCRVRIEVPCHCAGAGDAYRCTCGDELAPVK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9M0N8	MDEEASRTARESLELVFRMSNILDTGLDRHTLSVLIALCDLGVNPEALATVVKELRRESIPDSVTTTPSIH
Q08AG7	MASSSGAGAAAAAAAANLNAVRETMDVLLEISRILNTGLDMETLSICVRLCEQGINPEALSSVIKELRKATEALKAAENMTS
O95777	MTSALENYINRTVAVITSDGRMIVGTLKGFDQTINLILDESHERVFSSSQGVEQVVLGLYIVRGDNVAVIGEIDEETDSALDLGNIRAEPLNSVAH
P23059	MDILKLSDFIGNTLIVSLTEDRILVGSLVAVDAQMNLLLDHVEERMGSSSRMMGLVSVPRRSVKTIMIDKPVLQELTANKVELMANIV
P0A9I5	MHTNWQVCSLVVQAKSERISDISTQLNAFPGCEVAVSDAPSGQLIVVVEAEDSETLIQTIESVRNVEGVLAVSLVYHQQEEQGEETP
Q3E7Y6	MIPALTPEERQKLRSAILHRMQLELETTEKLIENIKEETLKKLNLLQQPDATSAPQSKELIREVLEQEGRRIE
P35087	MLPPLPDFSLSVEQQFDLQKYRQQVRDISREDLEDLFIEVVRQKMAHENIFKGMIRQGS
P27372	MEDLLGLLLSETGLLAIIYLGLSLAYLLVFPALLYWYLQKRWYVASSVERLVMYFLVFLFFPGLLVLSPVLNLRPRRQAA
P74771	MAAKMKKGSLVRVIRAQLENSLEAQASDRRLPDYLFHSKGEVLDLNEEYALVRFYVPTPNVWLRLDQIEALA
Q02377	MWFEVLPGIAVMGVCLFIPGMATARIHRFSNGGKEKRVAHYPYQWYLMERDRRVSGVNRSYVSKGLENID
Q02370	MAAAAAIRGVRGKLGLREIRIHLCQRSPGSQGVRDFIEKRYVELKKANPDLPILIRECSDVQPKLWARYAFGQEKNVSLNNFSADQVTRALENVLSSKA
Q02371	MAERVAAFLKNVWAKEPVLVASFAIAGLAVILPTLSPYTKYSLMINRATPYNYPVPLRDDGNMPDVPSHPQDPQGPSLEWLKRL
P42117	MSGTIPHFWAQPFRYIRWSAREKPAYFYSCVIAGLGPVFLTVVPPVRKYFGDVNPAPIPVTYPIPTGPRKQLTGYDDDTEEA
Q01321	MLRQIIGQAKRHPSLIPLFIFIGAGGTGAALYVTRLALFNPDVSWDRKNNPEPWNKLGPNDQYKFYSVNVDYSKLKKEGPDF
O75438	MVNLLQIVRDHWVHVLVPMGFVIGCYLDRKSDERLTAFRNKSMLFKRELQPSEEVTWK
Q02365	MAHGHGHEHGPSKMELPDYKQWKIEGTPLETVQEKLAARGLRDPWGRNEAWRYMGGFANNVSFVGALLKGFKWGFAAFVVAIGAEYYLESQKKDKKHH
Q02376	MAPSALLRPFWKLLAPARFPSVSSSRSKFYIQEPPHGSPNWLKVGLTLGTSVFLWIYLIKQHNEDVLEYKRRNGLE
P61849	MSFAQFFVACCLAIVLLAVSNTRAAVQGPPLCQSGIVEEMPPHIRKVCQALENSDQLTSALKSYINNEASALVANSDDLLKNYNKRTDVDHVFLRFGKRR
P58522	NPLFGIAGEDGPTGPSGIVGQ
P06020	MCSNEKARDWHRADVIAGLKKRKLSLSALSRQFGYAPTTLANALERHWPKGEQIIANALETKPEVIWPSRYQAGE
P84883	AQENETNESGSID
P35722	MDCCTESACSKPDDDILDIPLDDPGANAAAAKIQASFRGHMARKKIKSGERGRKGPGPGGPGGAGGARGGAGGGPSGD
P81796	QAIVSKARRPYIL
G2TRS1	MQKKEKIIENLEKAGNTFHLIIKCNTIARILKNLSTYAILLNSNILEEHVDLLLIYLSTQRMWKLNNLKKTLQYKGSKKK
P11633	MAATKEAKQPKEPKKRTTRRKKDPNAPKRGLSAYMFFANENRDIVRSENPDVTFGQVGRILGERWKALTAEEKQPYESKAQADKKRYESEKELYNATRA
Q03561	MSVDLKQQLELADYLGALAVWCIFFGVLFILSVIFNFVCIKKDDDVTALERWGYKKNIDMKLGPHRRSMVARQIPQTVVADH
P82900	MAGMMKKQVVTALMLALVVLAAAPGGARAACQASQLAVCASAILSGAKPSGECCGNLRAQQGCFCQYAKDPTYGQYIRSPHARDTLTSCGLAVPHC
P01297	APLSWDLPEPRSRAGKIRVHPRGNLWATGHFM
P34963	YKVDEDLQGAGGIQSRGYFFFRPRN
P04685	MADQLTLEIISAINKLVKAENGERTSVALGEITTDTELTSLGIDSLGLADVLWDLEQLYGIKIEMNTADAWSNLNNIGDVVEAVRGLLTKEV
Q9V5P6	MYLMYTINENGDRVYTLKKRTEDGRPTLSAHPARFSPEDKYSRQRLTIKKRFGLLLTQKPEPIY
P81303	MVEMRMKKCPKCGLYTLKEICPKCGEKTVIPKPPKFSLEDRWGKYRRMLKRALKNKNKAE
P61580	MNPSEMQRKGPPRRWCLQVYPTAPKRQRPSRTGHDDDGGFVEKKRGKCGEKQERSNCYCVCVERSRHRRLHFVMC
Q8MP00	MTFSTSSSFSRRALVALLVCTLLIDLSSFTDARPQDDPTSVAEAIRLLQELETKHAQHARPRFGKRSYLNPAGYGQDEQEDDWQDSTFTR

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P41334	KVVHLRPRSSFSSEDEYQIYLRNVSKYIQLYGRPRF
P86442	MSQSRPLALLVVAALVAAAVLVAAAEAQQADGNKLEGLADALKYLQELDRYYSQVARPRFGKRAEL RPDVVDDVIPEEMSADKFWRRFARRR
P41967	PDKDFIVNPSDLVLDNKAALRDYLRQINEYFAIIGRPRF
Q9VU58	MAKLAICILVFALFALALSARVPREESNPAQEFLTKAQGDFNEFIEKLKALDAKKVEGLFKDGLNTVQ EGLQKLNETFLQAPAAST
Q9VV28	MFKLCVFVALLSLAAAAPAPAPAPAPAPGLIGPGIVAPGIWGPTVVGSPLLAPQVVSVVPGAISHAAIT QVHPSPLLIKSVHGLGPVVIG
Q9VQ66	MFKLLVVVFAALFAAALAVPAPVARANPAPIPIASPEPAPQYYYGASPYAYSGGYYDSPYSYYG
P15506	AGEGLSSPFWSLAAPQRF
P0C0P6	MISSVKLNLILVLSLSTMHVFWCYPVPSSKVSGKSDYFLILLNSCPTRLDRSKELAFLKPILEKMFVKRSF RNGVGTGMKKTSFQRAKS
Q27441	MQRVILVVLLLSCMAVLSVRADNSEMLAPPPRPEEFTSAQQLRQYLAALNEYYSIMGRPRFGKRGDSF RKREFFRTNGERYPEDAAAWTEFQ
P0AC65	MRITIYTRNDCVQCHATKRAMENRGFDFEMINVDRVPEAAEALRAQGFRQLPVVIAGDLSWSGFRPD MINRLHPAPHAASA
Q48708	MVTVYSKNNCMQCKMVKKWLSEHEIAFNEINIDEQPEFVEKVIEMGFRAAPVITKDDFAFSGFRPSEL AKLA
Q16612	MVYYPELFVWVSQEPFPNKDMEGRLPKGRLPVPKEVNRKKNDETNAASLTPLGSSELRSPRISYLHFF
P19740	MQGKCWFLENKALKPFVCFYGGDQFLYIGDRIVSYFSTNDLYVALRGRIDKDLSLSRKVELYNGECVY LFCEHPAVGIVNTDFKLEIH
Q9UH64	MRQRGQEHLPTSVKSEPRACNNPTVAENRRVPSGLAAVIRNLTALWNPSLGVSERRGGDWEPSRIPR LWARVGWIQLPG
P0C6Z2	MNHLQQRQLFLENLLVGVNNMFHQMQKRPVNTCCRSLQKILDHLILLQTIHSPAFRLDQMQLRQMQT LACLWIHQYNHDHQVTLGAIKWISPLIKELK
P03902	MSMVYMNIMMAFTVSLVGLLMYRSHLMSSLLCLEGMMLSLFVMAALTILNSHFTLASMMPIILLVFAA CEAALGLSLLVMVSNTYGTDYVQNLNLLQC
Q9B6D4	MFIGTIILVLSFLGFVFNRRNIILAFICLETMLLGINLILLRNSVLFDDISGSLFAIVIIILAGVESAIGLSLLVS YYRLRGVINSYGI
C5W716	MIPLQHGLILAAILFVLGLTGLVIRRNLLFMLIGLEIMINASALAFVVAGSYWGQTDGQVMYILAISLAA AEASIGLALLLQLHRRRQNLNIDSVSEMRG
O60356	MATFPPATSAPQQPPGPEDEDSSLDESDLYSLAHSYLGGGGRKGRTKREAAANTNRPSPGGHERKLV TKLQNSERKKRGARR
P83490	QAIGPPYGLCFQCNQKTSSDCTEARRCSPFHEKCYTLYQPDENWMKSSGLSHFGCGKQCPTAGPEGR VTCCLTPRCN
P01526	ASSTWGGSYPACENNCRKQYDDCIKCQGKWAGKRGKCAAHCAVQTTSCNDKCKKH
B2BRQ5	MKTLLLTLVVVTIVCLDFGGGLICYMGPKTPRTCPPGQNLCYTKTWCDGFCGSRGKVVVLGCAATC PTVKPGVDITCCATDKCNPFPKTKAPWERP
P15815	RTCLISPSSTSQTCPKGQDICFTKAFCDRWCSSRGPVIEQGCAATCPEFTSRYKSLLCCTTDNCNH
P86096	LTCHTCPYNTCANSETCPAGKNICYQKKWEEHQGERIERRCVANCPKLGSNDKSLLCCRRDDCN
P60237	RICHSQMSSQPPTTTFCRVNSCYRRTLRDPHDPRGTIIVRGCGCPRMKPGTKLECCTSDKCNV
P83302	LICHRVHGLQTCEPDQKFCFRKTTMFFPNHPVLLMGCTSSCPTEKYSVCCSTDKCNK
P86094	LKCYGIFRKIMTCPQGQNICEKFAYSPMHNGWMYSWGCTSNCHKGPLDKCCSTDLCNY
P10808	RICLNQQQSTPEDQPTNGQCYIKTDCQNKTWNTHRGSRTDRGCGCPKVKPGINLRCCKTDKCNE
P14534	LKCHKAQFPNIETQCKWQTLCFQRDVKPHPSSMIVLRGCTSSCGKGAMCCATDLCNGPSTPST
P86421	LICYVSKDGETATCPPGQKCEKYAVSASHTGHWFYMYDCTSTCHIGPYNVCCSTDLCNR
P86422	LTCFNDFSPTAHTVEDCQRGITTCYMKTWRVHRETVIERGCGCPKVKPGIRLKCCTGNTCNY
P0CK21	MLVVIMFFIAFAFCSWLSYSYLRPYISTKELNKSR
P13155	MTDNAVLLGEGFTLMCLGMGFVLVFLLLLIFAIRGMSLAVNRLFPEPPAAPKPAPAAVAPADDFARLK PAIVAAIHHHRRLHP
Q8TAD7	MGCGNSTATSAGAGQGPAGAAKDVTEESVTEDDKRRNYGGVYVGLPSEAVNMVSSQTKTVRKN
Q9BZK8	MPVAPSNHCDNQCPHIFSKALVVSVAPSPPRDKPAPYTFTDVSSLCGLQKKCEGGKAMLFTLKRDRFS FLLFVSHC
P83518	NKLKPSQWISL
P19407	MESVAKPATTKEGSAKQAAIVVGVLALGWFAIQVAFIPLFNKVRGGGSDKKDDDVNAFTPDT
F0T376	MKKAVLLATVFCGVVGLTSCCRIVDCCFEDPCAPKPCNPCGNKKDKGCSPCGVYTPSCSKPCGSECN PGVQGPQAKGCTSLDGRCKQ

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
A8ABZ0	MEAREVEEMRRSRLLTLGGIGYTAVIALAALVLVMGALGLVLKVAAAAGALPSEVAKVANALPGLKA SVDANPAAGSLSSVSVST
P80045	YYEAPPDGRHLLLQPAPAAPAVAPAAPASWPHQQRRQALDEFAAAAAAAADAQFQDEEEDGGRRV
B6DT16	MSFFYSSAYAADKRNFDEIDRSGFNSFIKKKKNFDEIDRSGFDGFVKRNFDEIDRVGFGSFIKKSEPHH
P56717	QPLPDCCRQKTCSCRLYELLHGAGNHAAGILTL
P59636	MDPNQTNVVPPALHLVDPQIQLTITRMEDAMGQGQNSADPKVYPIILRLGSQLSLSMARRNLDSLEAR AFQSTPIVVQMTKLATTEELPDEFVVVTAK
P25511	IYVRPTNDELNYCGDFRELGQPDKKCRCDGKPCTVGRCKFARGDNDDKCISA
P0C6T2	MITDVQLAIFANMLGVSLFLLVVLYHYVAVNNPKKQE
Q99380	MISDEQLNSLAITFGIVMMTLIVIYHAVDSTMSPKN
Q92316	MTYEQLYKEFHSSKSFQPFIHLDTQPKFAICGLIVTLAVLSSALFAVGSKSSYIKKLFFYTILSVIGSLFAG LTTVFASNSFGVYV
Q800Y1	MKTLAILVLCSLAAICLTSSASAGAQPAGDSPVQGGLFMEKDQASAVVRQTRAAKELTLAQTESLREV CETNMACDEMADAQGIVAAYQAFYGPIPF
P02820	MRTPMLLALLALATLCLAGRADAKPGDAESGKGAAFVSKQEGSEVVKRLRRYLDHWLGAPAPYPDP LEPKREVCELNPDCDELADHIGFQEAYRRFYGPV
Q89769	MPTKAGTKSTANKKTTKGSSKSGSSRGHTGKTHASSSMHSGMLYKDMVNIARSRGIPIYQNGSRLTKS ELEKKIKRSK
P0C215	MLFRLLSPLSPLALTALLLFLLPPSDVSGLLLRPPPAPCLLLFLPFQILSGLLFLLFLPLFFSLPLLLSPSLPI TMRFPARWRFLPWKAPSQPAAAFLF
P32510	MALDGSSGGGSNVETLLIVAIIVVIMAIMLYYFWWMPRQQKKCSKAEECTCNNGSCSLKTS
P11130	MVPLKISTLESQLQPLVKLVATETPGALVAYARGLSSADRSRLYRLLRSLEQAIPKLSSAVVSATTLAAR GL
P84699	MSDDNDEDAPAVELGEGARVAGAPIARVASRLTWALQKSEVVRKEGDTVVRTPDGPRDLDAVLEDV STPYFETRREFERDVRAAMGAGPVPTE
Q37958	MQKPSGKGLKYFAYGVAISAAGAILAEYVRDWMRKPKAKS
Q9XJR8	MINKTTIKTVLITLGVLAAVNKVSALRSVKRLIS
Q9XJR5	MLGALMGVAGGAPMGGASPMGGMPSIASSSSAETGQQTQSGNFTGGGINFGSNNNNQLLIVGAVVIG LFLVIKRK
P0C790	MDTKTLIDKYNIENFTNYINFIIRNHQAGKGNLRFLVNLLKTTGGSNLKELDINPVEIENFNIDIYLDFLE FCLDSKFIF
O27252	MYIIFRCDCGRALYSREGAKTRKCVCGRTVNVKDRRIFGRADDFEEASELVRKLQEEKYGSCHFTNPS KRE
P76078	MSNVYWPLYEVFVRGKQGLSHRHVGSLHAADERMALENARDAYTRRSEGCSIWVVKASEIVASQPEE RGEFFDPAESKVYRHPTFYTIPDGIEHM
O31178	MTVFVDHKIEYMSLEDDAELLKTMAHPMRLKIVNELYKHKALNVTQIIQILKLPQSTVSQHLCKMRGK VLKRNRQGLEIYYSINNPKVEGIIKLLNPIQ
P68248	MPPRWASLLLLACSLLLLAVPPGTAGPSQPTYPGDDAPVEDLIRFYNDLQQYLNVVTRHRYGRRSSSR VLCEEPMGAAGC
P01298	MAAARLCLSLLLLSTCVALLLQPLLGAQGAPLEPVYPGDNATPEQMAQYAADLRRYINMLTRPRYGK RHKEDTLAFSEWGSPHAAVPRELSPLDL
P0C8W6	KQLLKEALAPEPAPKPAPEPAPEPAPEPAPEAAPEPAAAAPEAAPE
P58091	MASKNTSVVLGDHFQAFIDSQVADGRYGSASEVIRAGLRLLEENEAKLAALRAALIEGEESGFIEDFDF DAFIEERSRASAPQGFHEE
P0CW73	MVVNRALLASVDALSRDEQIELVEHINGNLAEGMHISEANQALIEARANDTDDAHWSTIDDFDKRIRAR LG
P0CW74	MNKPAKPAADDVDDLFGRPLTPAEEDTWFEHNREAIGQLVDEAWAEFERGEYDERSFAEIIAQGVAE HNAKR
Q9A458	MPSNGIVRSWRRAMATMNVSLPDAMREWVEGQTQSGRYHNASEYVRDLIRRDQERADKIAHLQRLID EGLDSGVGERSLHEIRAEARRRAGVDHEL
P22995	MSRLTIDMTDQQHQSLKALAALQGKTIKQYALERLFPGDADADQAWQELKTMLGNRINDGLAGKVS TKSVGEILDEELSGDRA
P95255	MSSRYLLSPAAQAHLEEIWDCTYDRWGVDQAEQYLRELQHAIDRAAANPRIGRACDEIRPGYRKLSA GSHTLFYRVTGEGTIDVVRVLHQRMDVDRNL
Q9A4S4	MGRVIRTRPVSGDLDRVFRDVCENNGVKVASAQLNRIESVFHRLSAFPRLGRDRSDLRPGLRTFSVKP WQVLYRLNGEDVVILRILDGRMNLAAQLGKKT
Q9A459	MWIMSYRLSRKAEQDLIDIYVAGVGLFGVAQAERYQDTLEAAFGAIAAFPHIGRERPELRPPVRVHPC KSHIILYVLDERGALIVRVRHAGEDWVGEAGG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q09098	MNSVTKISTLLIVILSFLCFVEGLICNSCEKSRDSRCTMSQSRCVAKPGESCSTVSHFVGTKHVYSKQMC SPQCKEKQLNTGKKLIYIMCCEKNLCNSF
O52748	MKAIMLVNFCDERGSGR
P0C574	MGQEQDTPWILSTGHISTQKREDGQQTPRLEHHNSTRLMDHCQKTMNQVVMPKQIVYWKQWLSLRS PTPVSLKTRVLKRWRLFSKHEWTS
P43511	LADDMPATMADQEVYRPEPEQIDSRNKYFSPRL
Q3E833	MTSKREKSLDHTLELKIPFETERQATIATKVLSPDPILKPQDFQVDYSSEKNVMLVQFRSIDDRVLRVG VSSIIDSIKTIVEAMDVLS
P63055	MSERQSAGATNGKDKTSGDNDGQKKVQEEFDIDMDAPETERAAVAIQSQFRKFQKKKAGSQS
P83586	NSELINAILGSPTLFGEV
D2IT41	MRFIILGVLFIAVASMILSNGVMAQSRDFSISEREIVASLAKQLLRVARMGYVPEGDLPRKRNAELINSLL GVPRVMSDAGRR
P09929	MTAMAVSGKLLTALVLSTYILGLALTIQATQYEEDKYQENEVKYGRELASWLAQLAHKNEPAICAHK RNSEIINSLLGLPKLLNDAGRK
O80446	MAGTGVVAVYGEGAMTETKQKSPFSVKVGLAQMLRGGVIMDVVNAEQARIAEEAGACAVMALERV PADIRAQGGVARFC
P86488	DAQEKRQPWLPFG
Q79F93	MEKMSHDPIAADIGTQVSDNALHGVTAGSTALTSVTGLVPAGADEVSAQAATAFTSEGIQLLASNASA QDQLHRAGEAVQDVARTYSQIDDGAAGVFAE
P80577	AGEDVSHELEEKEKALANHSE
P80578	EELRPEVLPDVSE
P13975	MHTTRLKRVGGSVMLTVPPALLNALSLGTDNEVGMVIDNGRLIVEPYRRPQYSLAELLAQCDPNAEIS AEEREWLDAPATGQEEI
Q9LV87	MEKSDRRSEESHLWIPLQCLDQTLRAILKCLGLFHQDSPTTSSPGTSKQPKEEKEDVTMEKEEVVVTS RATKVKAKQRGKEKVSSGRPGQHN
Q08362	MVEPLLCGIVLGLVPVTIAGLFVTAYLQYLRGDLATY
P50369	MIFDFNYIHIFMLTITSYVGLLIGALVFTLGIYLGLLKVVKLI
P83795	MILGAVFYIVFIALFFGIAVGIIFAIKSIKLI
Q42496	MAMSIAARSACCGVAAPRSSTVRVAAARPAVRPSLRTAGQKAAPSRGVATKAVNELAMIAGEAEFIA GTALTMVGMTLVGLAIGFVLLRVESLVEEGKI
P83796	MTEEMLYAALLSFGLIFVGWGLGVLLLKIQGAEKE
P0C1D4	MLAEGEPAIVQIGWAATCVMFSFSLSLVVWGRSGL
P82694	SDLTWTYQSPGDPTNSKN
P82696	AFLTLTPGSHVDSYVEA
P82697	TDPLWQLPGAHLEQYLS
P23816	KNGDLRAPYVEIFDARGCDAKNSQYTGPKSGDMNDDQCVKVSMAVPKVSEATAEKKRQEFLGFKET AINVPQIAGKTKKY
Q06253	MQSINFRTARGNLSEVLNNVEAGEEVEITRRGREPAVIVSKATFEAYKKAALDAEFASLFDTLDSTNKE LVNR
P11860	AKAKRSPRKKKAAVKKSSKSKAKKPKSPKKKKAAKKPAKKAAKKK
P25271	KLSYDDKVFENVEFTPRL
P65722	MARNRLTESEMNEALRALDGWQKVDGREAITRSFKFKDFSTAFGFMAQAALYAEKLDHHPEWFNAY NRVDVTLATHSENGVTELDIKMARKMNAIAG
A0R2K7	MAVLSNDQVDAALPNLPGWERAAGALRRSVKFPTFLDGIDAVRRVAEFAEEKDHHPDIDIRWRTVTF ALVTHAAGGITEKDVQMAEEINRILSD
P86282	FLISIPYSASIGGTATLTGTA
Q7M4T6	SAGKFIVIFKNDVSEDKIRETKDEVIAEGGTITNEYNMPGMKGFAGELTPQSLTKFQGLQGDLIDSIEE DGIVTTQ
Q3MUY2	MFLSLPTLTVLIPLVSLAGLFYSASVEENFPQGCTSTASLCFYSLLLPITIPVYVFFHLWTWMGIKLFRH N
P35476	IEVLLGSDDGGLAFVPGNFSVSAGEKITFKNNAGFPHNVVFDEDEIPAGVDVSKISMSEEEYLNGPGET YSVTLSEKGTYTFYCAPHQGAGMVGKVTVN
Q02325	MEHKEVVLLLLLFLKSGQGEPLDDYVNTQGPSLFSVTKKQLGAGSREECAAKCEEDKEFTCRAFQYH SKEQQCVIMAENRKSSIIIRMRDAVLFEK
P85036	ECNIGDICVVHGDCSECKCSDGRAAKCDREHGEPPHCSCSHR

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P56513	MAPLHHILVLCVGFLTTATAEAPQEHDPFTYDYQSLRIGGLIIAGILFILGILIVLSRRCRCKFNQQQRTG EPDEEEGTFRSSIRRLSTRRR
P82595	LSCASCENAACPAIGLPCKPSEYVYTPCGCCPQCPLELGQPCGSFTQRCQFDLWCLRRKGNKIEAYKY VPWHLDFKGVCARVDV
P84388	SADIVKKLWDNPAL
P84389	SVDMVMKGIKLWPL
P40975	MLMSTLPGGVILVFILVGLACIAIISTIIYRKWQARQRGLQRF
P87284	MDSAKIINIILSLFLPPVAVFLARGWGTDCIVDIILTILAWFPGMLYALYIVLQD
P69392	RIKIGLFDQLSRL
P55791	FGGFTGARKSARKLANQ
P81765	AVTDNEIVPQCLANGSKCYSHDVCCTKRCHNYAKKCVT
Q8VWY7	MAEADDPQDIADRERIFKRFDANGDGQISATELGETLQTLGSVTPEEVKYMMDEIDTNKDGFISFQEFI EFARANRGLIRDVAKIF
P01154	RGEVKDASGELEPPPGPFIQRGRASCWLGRTGSCQNCWLCSQGNCAGV
P80524	MKNEGEKEGLNISRCRVCKPGSTLINKTGGWRNFRPVYIYEKCTKCGICHIVCPDMSVKPRENGFFEY DYDYCKGCGICANECPADAIEMILEEK
Q56316	MSLKSWKEIPIGGVIDKPGTAREYKTGAWRVMRPILHKEKCIDCMFCWLYCPDQAIIQEGGIMKGFN YDYCKGCGLCANVCPKQAIEMRPETEFLSEEG
P86295	ESIVRPPPVEAKVEETPE
P41736	FLPLLILGSLLMTPPVIQAIHDAQR
P0A9L5	MAKTAAALHILVKEEKLALDLLEQIKNGADFGKLAKKHSICPSGKRGGDLGEFRQGQMVPAFDKVVF SCPVLEPTGPLHTQFGYHIIKVLYRN
P82684	DEGGTQYTPRL
P84411	DHLPHDVYSPRL
P85741	SASGGAGESSGMWFGPRL
P85864	DGAETPGAAASLWFGPRV
A4IFH6	MDKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLISICLLLICIIVMLL
P01372	DVPSANANANNQRTAAAKPQANAEASS
P31992	MPHIDIKCFPRELDEQQKAALAADITDVIIRHLNSKDSSISIALQQIQPESWQAIWDAEIAPQMEALIKKP GYSMNA
P27532	MQWTKPAFTDLRIGFEVTMYFEAR
Q8P6M8	MSTISRDSCPALRAGVRLQHDRARDQWVLLAPERVVELDDIALVVAQRYDGTQSLAQIAQTLAAEFD ADASEIETDVIELTTTLHQKRLLRL
Q00LT9	MCTTLFLLSTLAMLWRRRFANRVQPEPSGADGAVVGSRSERDLQSSGRKEEPLK
P85798	MSSLRIAMFLLVVLMVLIDCSTAIPAADKERLLNEVDLVDDDGSIETALINYLFTKQIVKRLRNQLDIGD LQRKRSYWKQCAFNAVSCFGK
P81264	MKAVGAWLLCLLLLGLALQGAASRAHQHSMEIRTPDINPAWYAGRGIRPVGRFGRRRAAPGDGPRP GPRRVPACFRLEGGAEPSRALPGRLTAQLVQE
P83266	ARRRHSMKKKRKSVRRRKTRKNQRKRKNSLGRSFKQHGFLKQPPRFRP
P30259	GCKKRKARKRPKCKKARKRPKCKRRKVAKKKC
P08433	MGSCKPKKKQAPCFLRRRHLRRLNVCKRDTSKTYRRRRHVRRLPKKRRRRC
P0A411	MSHTVKIYDTCIGCTQCVRACPTDVLEMVPWDGCKAAQVASSPRTEDCVGCKRCETACPTDFLSIRV YLGAETTRSMGLAY
P12352	MQALSSRVNIAAKPQRAQRLVVRAEEVKAAPKKEVGPKRGSLVKILRPESYWFNQVGKVVSVDQSGV RYPVVVRFENQNYAGVTTNNYALDEVVAAK
P12975	MALNRGDKVRIKRTESYWYGDVGTVASVEKSGILYPVIVRFDRVNYNGFSGSASGVNTNNFAENELEL VQAAAK
P23079	MATAFLPSILADASFLSSIFVPVIGWVVPIATFSFLFLYIEGEDVA
P17227	MINLPSLFVPLVGLLFPAVAMASLFLHVEKRLLFSTKKIN
P23080	MADKADQSSYLIKFISTAPVAATIWLIITAGILIEFNRFFPDLLFHPLP
P23317	MLTSTLLAAATTPLEWSPTIGIIMVIANVIAITFGRQTIKYPSAEPALPSAKFFGGFGAPALLATTAFGHI LGVGIILGLHNLGRF
P23318	MVLATTLPDTTWTPSVGLVVILSNLFAIALGRYAIQSRGKGPGLPIALPALFEGFGLPELLATTSFGHLL AAGVVSVGLQYAGAL
Q7NHY3	MAATVVSGAQVAIAFVVALIAGIAALLLSTALGK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P0A403	MALTDTQVYVALVIALLPAVLAFRLSTELYK
P23319	MAKAKTPAVANTGAKPPYTFRTAWALLLLGVNFLVAAYYFHIIQ
P0C029	MQSYNVFPALVIITTLVVPFMAAAALLFIIERDPS
Q8DIP0	MAGTTGERPFSDIITSVRYWVIHSITIPALFIAGWLFVSTGLAYDVFGTPRPDSYYAQEQRSIPLVTDRF EAKQQVETFLEQLK
Q8DIN9	MTSNTPNQEPVSYPIFTVRWVAVHTLAVPTIFFLGAIAAMQFIQR
P22666	MATGTSKAKPSKVNSDFQEPGLVTPLGTLLRPLNSEAGKVLPGWGTTVLMAVFILLFAAFLLIILEIYN SSLILDDVSMSWETLAKVS
Q54697	MLTLKIAVYIVVGLFISLFIFGFLSSDPTRNPGRKDFE
P59087	MMSEGGRIPLWIVATVAGMGVIVIVGLFFYGAYAGLGSSL
P12163	MLNIFSLICLNSALYSSSFFFGKLPEAYAFLSPIVDFMPVIPLFFFLLAFVWQAAVSFR
Q55354	MDRNSNPNRQPVELNRTSLYLGLLLVAVLGILFSSYFFN
Q8DHA7	MEVNQLGLIATALFVLVPSVFLIILYVQTESQQKSS
P61840	MEALVYTFLLVSTLGIIFFAIFFREPPKISTKK
Q9F1R6	MTITPSLKGFFIGLLSGAVVLGLTFAVLIAISQIDKVQRSL
Q8DKM3	MDWRVLVVLLPVLLAAGWAVRNILPYAVKQVQKLLQKAKAA
P92276	MTSILQVALLALIFVSFALVVGVPVVFATPNGWTDNKGAVFSGLSLWLLLVFVVGILNSFVV
Q9I317	MMTALETRLSVADGTHAAALRQRLQAALAECRRELARGACPEHFQFLQQQARALEGGLGILSQLTED
A2YFB4	MVNPGRTARALCLLCLALLLLGQDTHSRKLLLQEKHSHGVGNGTTTTQEPSRENGGSTGSNNNGQLQ FDSAKWEEFHTDYIYTQDVKNP
Q9FRF9	MSPKVIAICLVALLLPISISHGGRIGPIEPSKASSKVVERGNYDGRVEGCEEDDCLVERLLVAHLDYIYTQ GKHN
Q9AR88	MAARTVAVAAALAVLLIFAASSATVAMAGRPTPTTSLDEEAAQAAAQSEIGGGCKEGEGEEECLARR TLTAHTDYIYTQQHHN
Q9FS10	MSSKAITLLLIALLFSLSLAQAARPLQPADSTKSVHVIPEKVHDEACEGVGEEECLMRRTLTAHVDYIY TQDHNP
P0C7Y0	MGIIAGIIKVIKSLIEQFTGK
P0C804	MEFVAKLFKFFKDLLGKFLGNN
P0C817	MAIVGTIIKIIKAIIDIFAK
P15785	LRIPCCPVNLKRLLVVVVVVVLVVVVIVGALLMGL
P32696	MLELLFVIGFFVMLMVTGVSLLGIIAALVVATAIMFLGGMLALMIKLLPWLLLAIAVVWVIKAIKAPKV PKYQRYDRWRY
Q941C7	MTFVVRLLVCLLLTLTITSSLARNPVSVSGGFENSGFQRSLLMVNVEDYGDPSANPKHDPGVPPSATG QRVVGRG
P0AA04	MFQQEVTITAPNGLHTRPAAQFVKEAKGFTSEITVTSNGKSASAKSLFKLQTLGLTQGTVVTISAEGE DEQKAVEHLVKLMAELE
P07515	MEKKEFHIVAETGIHARPATLLVQTASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGSDVTITVDGA DEAEGMAAIVETLQKEGLAE
P45611	MAKFSAIITDKVGLHARPASVLAKEASKFSSNITIIANEKQGNLKSIMNVMAMAIKTGTEITIQADGNDA DQAIQAIKQTMIDTALIQG
O50515	MAERRVNVGWAEGLHARPASIFVRAATATGVPVTIAKADGSPVNAASMLAVLGLGAQGGEEIVLASD AEGAEAALERLAKLVAEGLEELPETV
P0A435	MKRKIIVACGGAVATSTMAAEEIKELCQSHNIPVELIQCRVNEIETYMDGVHLICTTARVDRSFGDIPL VHGMPFVSGVGIEALQNKILTILQG
Q7VSX7	MIHAHSNARLLRWAILAIAPVTLGACAPNGPPGLPYPDGKPLIPINTAAPEQGSSCQTRAP
P0A9N0	MTVKQTVEITNKLGMHARPAMKLFELMQGFDAEVLLRNDEGTEAEANSVIALLMLDSAKGRQIEVEA TGPQEEEALAAVIALFNSGFDED
P58606	AEKDCIAPGAPCFGTDKPCCNPRAWCSSYANKCL
P13402	MADKTIFNDHLNTNPKTNLRLWVAFQMMKGAGWAGGVFFGTLLLIGFFRVVGRMLPIQENQAPAPN ITGALETGIELIKHLV
A0QZ48	MAQEQTKRGGGGGEDDDLPGASAAGQERREKLTEETDDLLDEIDDVLEENAEDFVRAYVQKGGQ
P85772	GASGLIPVMRN
P85580	GSSGGLITFGRT
P85693	GSSGLIPMGRT
P84439	GSSSGLISMPRV

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P85700	GSSGGMIPFPRV
P85582	GSSGMISFPRT
P86660	GKTKGVTSGLIAFPRL
Q8K459	MQLRHIGDSVNHRVIQEHLAQEVGDVLAPFVALVFVRGQVLLRFFWNNHLL
P80558	MSRLFKITALVPSLSRTRTQRELQNTYFTKLVPYENWFREQQRIQKAGGKIIKVELATGKQGTNAGLQ
Q8WXC3	MGTKREAILKVLENLTPEELKKFKMKLGTVPLREGFERIPRGALGQLDIVDLTDKLVASYYEDYAAEL VVAVLRDMRMLEEAARLQRAA
Q56P42	MASSAELDFNLQALLEQLSQDELSKFKSLIRTISLGKELQTVPQTEVDKANGKQLVEIFTSHSCSYWAG MAAIQVFEKMNQTHLSGRADEHCVMPPP
P73202	MLGQSSLVGYSNTQAANRVFVYEVSGLRQTDANENSAHDIRRSGSVFIKVPYARMNDEMRRISRLGGT IVNIRPYQADSNEQN
Q8VUS8	MNALVGCTTSFDPGWEVDAFGAVSNLCQPMEADLYGCADPCWWPAQVADTLNTYPNWSAGADDV MQDWRKLQSVFPETKGSS
P07552	MLTRFLGPRYRQLARNWVPTASLWGAVGAVGLVWATDWRLILDWVPYINGKFKKDD
P37299	MAYTSHLSSKTGLHFGRLSLRSLTAYAPNLMLWGGASMLGLFVFTEGWPKFQDTLYKKIPLLGPTLE DHTPPEDKPN
P00126	MGLEDEQRMLTGSGDPKEEEEEEEELVDPLTTVREQCEQLEKCVKARERLELCDERVSSRSQTEEDC TEELLDFLHARDHCVAHKLFNSLK
P48504	MSDEEVVDPKATLEVSCKPKCVRQLKEYQACTKRVEGDESGHKHCTGQYFDYWHCIDKCVAAKLFD HLK
P13271	MGRQFGHLTRVRHVITYSLSPFEQRAFPHYFSKGIPNVLRRTRACILRVAPPFVAFYLVYTWGTQEFE KSKRKNPAAYENDR
P46269	MGKQPVKLKAVVYAISPFQQKIMPGLWKDLPGKIHHKVSENWISATLLLGPLVGTYSYVQHFLEKEKL EHRY
P08525	MGPPSGKTYMGWWGHMGGPKQKGITSYAVSPYAQKPLQGIFHNAVFNSFRRFKSQFLYVLIPAGIY WYWWKNGNEYNEFLYSKAGREELERVNV
P00130	MVAPTLTARLYSLLFRRTSTFALTIVVGALFFERAFDQGADAIYEHINEGKLWKHIKHKYENKE
P46270	MESAARRSGGGVLEGFYRLVMRRTPVYVTFVIAGALLGERAVDYGVKTLWEKNNVGKRYEDISVLG QRPVDE
P22289	MSFSSLYKTFFKRNAVFVGTIFAGAFVFQTVFDTAITSWYENHNKGKLWKDVKARIAAGDGDDDDE
P33229	MRYDNVKPCPFCGCPSVTVKAISGYYRAKCNGCESRTGYGGSEKEALERWNKRTTGNNNGGVHV
P08950	MTTIPAIGILPIDFLTILLLFSFISHSVCVEFAEDAGELDKSNAFRRQVPQWAVGHFMGKRSLSDDTEQA TMYSSRFVESTS
P03050	MKGMSKMPQFNLRWPREVLDLVRKVAEENGRSVNSEIYQRVMESFKKEGRIGA
P52119	MPGKIAVEVAYALPEKQYLQRVTLQEGATVEEAIRASGLLELRTDIDLTKNKVGIYSRPAKLSDSVHD GDRVEIYRPLIADPKELRRQRAEKSANK
P33230	MYKITATIEKEGGTPTNWTRYSKSKLTKSECEKMLSGKKEAGVSREQKVKLINFNCEKLQSSRIALYS N
P64530	MSHTIRDKQKLKARASKIQGQVVALKKMLDEPHECAAVLQQIAAIRGAVNGLMREVIKGHLTEHIVHQ GDELKREEDLDVVLKVLDSYIK
P03036	MQTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIAMALNCDPVWLQYGTKR GKAA
P09964	MYKKDVIDHFGTQRAVAKALGISDAAVSQWKEVIPEKDAYRLEIVTAGALKYQENAYRQAA
P03040	MEQRITLKDYAMRFGQTKTAKDLGVYQSAINKAIHAGRKIFLTINADGSVYAEEVKPFPSNKKTTA
P82943	MRVSLAALAFLLTLAVLHSEANEEPAGNMRVCCFSSVTRKIPLSLVKNYERTGDKCPQEAVIFQTRSG RSICANPGQAWVQKYIEYLDQMSK
P04891	MTVITYGKSTFAGNAKTRRHERRRKLAIERDTICNIIDSIFGCDAPDASQEVKAKRIDRVTKAISLAGTR QKEVEGGSVLLPGVALYAAGHRKSKQITAR
P07243	MVTIVWKESKGTAKSRYKARRAELIAERRSNEALARKIALKLSGCVRADKAASLGSLRCKKAEEVERK QNRIYYSKPRSEMGVTCVGRQKIKLGSKPLI
Q9AA08	MADGFDIHIDQEQAARLKVVADRLGMSVSEYAVALIDAGLTGAAPKAIDPDPAIDEAIADAIERGDEPA ISRDEFRAHIRRVTAGLG
Q9A5D6	MAICYARFMVPEPSIFEIDAEAEEAADAEGMADIAAGRVVPHEEVSAWLDTWGTPEEKPAPETWRK
P0CL56	MRLKKRFKKFFISRKEYEKIEEILDIGLAKAMEETKDDELLTYDEIKELLGDK
P0C079	MGSINLRIDDELKARSYAALEKMGVTPSEALRLMLEYIADNERLPFKQTLLSDEDAELVEIVKERLRNP KPVRVTLDEL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9AA09	MTFTVLVSVRAKRDFNRLIVWLVERDPRAAARLGPLLEAALDSLTEAPSRGRSVGPTTREISIPFGQSA YVIRYRLLGSSVHVTRIWHGLEQR
Q9A5D7	MAQVVWTWRALADLTAIRDYIGQFSPLAAQRMALRLKTAADSLAEYPERGRLATATLRELVVVPPY VIRYYVADGLVHIVRIRHAARL
Q9A4F4	MKSVELGPRARRDLTKLRRWLLNRAPSAADRAIDLILSRAEQLAQHSDLGRRKSQNMRELYVSFGAHG YVLQYRVYPDAVVIARIRHSLERR
Q58503	MKVLFAKTFVKDLKHVPGHIRKRIKLIIEECQNSNSLNDLKLDIKKIKGYHNYYRIRVGNYRIGIEVNGD TIIFRRVLHRKSIYDYFP
P0C077	MAYFLDFDERALKEWRKLGSTVREQLKKKLVEVLESPRIEANKLRGMPDCYKIKLRSSGYRLVYQVID EKVVVFVISVGKRERSEVYSEAVKRIL
O50461	MSDDHPYHVAITATAARDLQRLPEKIAAACVEFVFGPLLNNPHRLGKPLRNDLEGLHSARRGDYRVV YAIDDGHHRVEIIHIARRSASYRMNPCRPR
O33347	MRILPISTIKGKLNEFVDAVSSTQDQITITKNGAPAAVLVGADEWESLQETLYWLAQPGIRESIAEADA DIASGRTYGEDEIRAEFGVPRRPH
P65067	MSISASEARQRLFPLIEQVNTDHQPVRITSRAGDAVLMSADDYDAWQETVYLLRSPENARRLMEAVAR DKAGHSAFTKSVDELREMAGGEE
P64528	MRSVNFDPDAWEDFLFWLAADRKTARRITRLIGEIQRDPFSGIGKPEPLQGELSGYWSRRIDDEHRLV YRAGDDEVTMLKARYHY
P11184	QSTNDLIKACGRELVRLWVEICGSVRWGQSALRMTLSEKCCQVGCIRKDIARLC
P81191	WPRGPDYTERRVMCGLQYVRAAISICGPNMQTMRPRNGSGPIVPPPDFLAMYGMARYKPPLSKKCC STGCNREDFRGYCYL
P20465	MVAYPEISWTRNGCTVAKYPEISWTRNGCTVSKYPEISWTRNGCTVSKYPEISWTRNGCTVSKYPEIS WTRNGCTVA
A6MWS7	MKIIVVLAVLMLVSAQVCLVSAAEMGHSSDNELSSRDLVKRFFLPPCAYKGTCNH
P56568	PRGSPRTEYEACRVRCQVAEHGVERQRRCQQVCEKRLREREGRREVD
P86129	IVSYPDDAGEHAHKMG
P06391	MDGIKYAVFTDKSIRLLGKNQYTSNVESGSTRTEIKHWVELFFGVKVIAMNSHRLPGKSRRMGPIMGH TMHYRRMIITLQPGYSIPPLRKKRT
P28804	MAVPKKRTSIYKKRIRKNIWKKKGYWAALKAFSLAKSLSTGNSKSFFVRKISNQTLE
P28805	MAKGKDVRVKVILECTGCVRKSVNKGSRGVSRYITQKNRHNTPSRLELRKFCPYCYKHTIHGEIKK
P12230	MKIRASVRPICEKCRLIRRRGRIIVICSNPKHKQRQG
Q980C1	MPAIEVGRICVKVKGREAGSKCVIVDIIDDNFVLVTGPKDITGVKRRRVNILHLEPTDKKIDIQKGASD EEVKKKLEESNLTEYMKEKIKIRMPTL
Q9RY64	MFAIIQTGGKQYRVSEGDVIRVESLQGEAGDKVELKALFVGGEQTVFGEDAGKYTVQAEVVEHGRG KKIYIRKYKSGVQYRRRTGHRQNFTAIKILGIQG
Q9HQG8	MPNSNGPLSNSGGKLQNDPRDRGTSPPQRAIADYDDGESVHLTLDPSVQDGRFHPRFSGLTGTVVGT QGDAFKVEVNDGGMDKTLIVGAAHLRTQRQEE
P0ADZ2	MIREERLLKVLRAPHVSEKASTAMEKSNTIVLKVAKDATKAEIKAAVQKLFEVEVEVVNTLVVKGKV KRHGQRIGRRSDWKKAYVTLKEGQNLDFVGGAE
P12732	MSWDVIKHPHVTEKAMNDMDFQNKLQFAVDDRASKGEVADAVEEQYDVTVEQVNTQNTMDGEKK AVVRLSEDDDAQEVASRIGVF
P10143	MDAFDVIKTPIVSEKTMKLIEEENRLVFYVERKATKEDIKEAIKQLFNAEVAEVNTNITPKGQKKAYIK LKDEYNAGEVAASLGIY
A0QSD3	MATITDPRDIILAPVISEKSYGLIEDNVYTFVVHPDSNKTQIKIAIEKIFDVKVDSVNTANRQGKRKRTR TGFGKRKSTKRAIVKLAAGSKPIDLFGAPA
P14116	MPRTRECDYCGTDIEPGTGTMFVHKDGATTHFCSSKCENNADLGREARNLEWTDTARGEAGEAEDE A
P68919	MFTINAEVRKEQGKGASRRLRAANKFPAIIYGGKEAPLAIELDHDKVMNMQAKAEFYSEVLTIVVDGK EIKVKAQDVQRHPYKPKLQHIDFVRA
P66133	MLKLNLQFFASKKGVSSTKNGRDSESKRLGAKRADGQFVTGGSILYRQRGTKIYPGENVGRGGDDTL FAKIDGVVKFERKGRDKKQVSVYAVAE
P0A7M2	MSRVCQVTGKRPVTGNNRSHALNATKRRFLPNLHSHRFWVESEKRFVTLRVSAKGMRVIDKKGIDTV LAELRARGEKY
Q9WY96	MAKRCEVCGKAPRSGNTVSHSDKKSGRWFRPNLQKVRVVLPDGTIKRMRVCTSCLKSGKVKKYVGQ VSEV
Q72G84	MSKVCEISGKRPIVANSIQRRGKAKREGGVGKKTTGISKRRQYPNLQKVRVRVAGQEITFRVAASHIP KVYELVERAKGLRLEGLSPKEIKKELLKLL
Q24154	MAKSKNHTNHNQNKKAHRNGIKRPLRKRHESTLGMDVKFLINQRYARKGNLSREESVKRYNERIASQ KGKPKPVTL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P10971	MTVLHVQEIRDMTPAEREAELDDLKTELLNARAVQAAGGAPENPGRIKELRKAIARIKTIQGEEGDLQ ENE
P95057	MAVGVSPGELRELTDEELAERLRESKEELFNLRFQMATGQLNNNRRLRTVRQEIARIYTVLRERELGL ATGPDGKES
Q6N4U2	MAEMKTADIRAMSEDQMDDAILSLKKERFNLRFQRATGQLENTSRLREARRDIARIKTIAAQKRAGKT K
Q72I12	MKLSEVRKQLEEARKLSPVELEKLVREKKRELMELRFQASIGQLSQNHKIRDLKRQIARLLTVLNEKRR QNA
O83227	MGRGGCAQLSYSELLSRRRELERKYLDLRFQLVVEHVDNKLMKRILRRQIAAVNTFLRHKELTELEKR GVRE
Q9HWF3	MATVKVTLVKSLNGRLANHKACVKGLGLRRINHTVEVQDTPENRGMINKAYYLLRVEG
Q6N4V1	MAKAANMIKVEQIGSPIRRHHSQRETLIGLKLNKIGRVAELQDTPEVRGMIGKVQHLVRVVDEK
O34967	MKEGIHPKNHKVIFQDVNSGYRFLSTSTKTSNETAEWEDGNTYPVIKVEVSSDTHPFYTGRQKFNEK GGRVEQFKKRYNMGK
P0A7N1	MKPNIHPEYRTVVFHDTSVDEYFKIGSTIKTDREIELDGVTYPYVTIDVSSKSHPFYTGKLRTVASEGN VARFTQRFGRFVSTKKGA
Q03223	MKAGIHPNFKKATVKCACGNEFETGSVKEEVRVEICSECHPFYTGRQKFASADGRVDRFNKKYGLK
Q9RW44	MQKDLHPKAVPCKIIYQGQVVMETMSTRPEIHVDVWSGVHPFWTGEERFLDTEGRVDKFNKRFGDS YRRGSKK
P18138	MSASDFEERVVTIPLRDARAEPNHKRADKAMILIREHLAKHFSVDEDAVRLDPSINEAAWARGRANTP SKIRVRAARFEEEGEAIVEAETAE
Q6NBB0	MKTEIHPDYHTITVVMTDGTEYQTRSTWGKEGDKLNLDIDPKSHPAWTGGTQQVLDRGGRVSRFQK KFSGFLKKD
P49228	MAKHPVPKKKTSKSKRDMRRSHHALTAPNLTECPQCHGKKLSHHICPNCGYYDGRQVLAV
Q9HZN4	MAVQQNKKSRSARDMRRSHDALESNALSVEKSTGEVHLRHHVSPDGFYRGRKVVDKGSDE
Q6NCE6	MAVPRRKTSPSRRGMRRSADAIKRPTYVEDKDSGELRRPHHLDLKTGMYKGRQVLKKKDS
P78015	MAVKRSTRLGCNDCREINYLTFKNVKKNPEKLALNKFCSRCRKVVVHKEVKRK
P66231	MRVNVTLACTECGDRNYITTKNKRNNPERIEMKKYCPRLNKYTLHRETK
Q9VR93	MSDHFNFNEAFNSQTMRGRANVAKATWASLGLVYVLVKMHRRNTKRRETKLYCKGCQQAMLHG
P0A7N9	MAKGIREKIKLVSSAGTGHFYTTTKNKRTKPEKLELKKFDPVVRQHVIYKEAKIK
P80340	MKRTWQPNRRKRAKTHGFRARMRTPGGRKVLKRRRQKGRWRLTPAVRKR
Q8TZV6	MRIKGVVLSYRRSKENQHNNVMIIKPLDVNSREEASKLIGRLVLWKSPSGKILKGKIVRVHGTKGAVRA RFEKGLPGQALGDYVEIV
Q9RSW6	MPKMKTHKMAKRRIKITGTGKVMAFKSGKRHQNTGKSGDEIRGKGKGFVLAKAEWARMKLMLPRG K
P0A7Q1	MPKIKTVRGAAKRFKKTGKGGFKHKHANLRHILTKKATKRKRHLRPKAMVSKGDLGLVIACLPYA
Q6NDR5	MPKLKTKSGAKKRFKVTATGKVMSAQRGKRHGMIKRTKKQIRQLRGTRAIFKTDGDNIKKYFLPNA
Q2EEQ2	MKVLNSLRTAKERHPDCQIVKRKGRLYVICKSNPRFKAVQGRKKKR
P05745	MTVKTGIAIGLNKGKKVTSMTPAPKISYKKGAASNRTKFVRSLVREIAGLSPYERRLIDLIRNSGEKRA RKVAKKRLGSFTRAKAKVEEMNNIIAASRRH
Q9VMU4	MAKRTKKVGIVGKYGTRYGASLRKMVKKMEITQHSKYTCSFCGKDSMKRAVVGIWSCKRCKRTVA GGAWVYSTTAAASVRSAVRRLRETKEQ
P60619	MASKSGKTGSSGRFGARYGRVSRRRVAEIESEMNEDHACPNCGEDRVDRQGTGIWQCSYCDYKFTG GSYKPETPGGKTVRRSIRAALSEDEE
P32410	MTGAGTPSQGKKNTTTHTKCRRCGEKSYHTKKKVCSSCGFGKSAKRRDYEWQSKAGE
P61927	MTKGTSSFGKRRNKTHTLCRRCGSKAYHLQKSTCGKCGYPAKRKRKYNWSAKAKRRNTTGTGRMR HLKIVYRRFRHGFREGTTPKPKRAAVAASSSS
P63173	MPRKIEEIKDFLLTARRKDAKSVKIKKNKDNVKFKVRCSRYLYTLVITDKEKAEKLKQSLPPGLAVKE LK
Q22SV3	MPKEITDIKKFMKLWQNNKDTPATAGAKKVVYVKTNKRITKFKLRGKKYLYTFKTADPKIAKGIKD AIPATYSKIEIKGKNTKKATKRS
P49167	MAREITDIKQFLELTRRADVKTATVKINKKLNKAGKPFRQTKFKVRGSSSLYTLVINDAGKAKKLIQSL PPTLKVNRL
P22452	MGKKSKATKKRLAKLDNQNSRVPAWVMLKTDREVQRNHKRRHWRRNDTDE
P0DJ61	MGANKTLNMKKRFGRKIKQNRPLPNWYRYKSDTNIRYNSKRRNWRRTKLKIY
Q980V5	MPLTDPAKLQIVQQRVFLKKVCRKCGALNPIRATKCRRCHSTNLRLKKKELPTKKG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P32411	MQMPRRFNTYCPHCNEHQEHEVEKVRSGRQTGMKWIDRQRERNSGIGNDGKFSKVPGGDKPTKKT DLKYRCGECGKAHLREGWRAGRLEFQE
P14125	MSTYTVRGSFPARDGPQQFEKEVEAPNENVAEERVYSDFGSQHNLKRTQITIEEVA
O27647	MKMKTKIFRVKGKFLMGDKLQPFTKELNAIREEEIYERLYSEFGSKHRVPRSKVKIEEIEEISPEEVQD PVVKALVQR
P38613	MAEVKIFMVRGTAIFSASRFPTSQKFTKYVRALNEKQAIEYIYSQLGGKNKIKRYNIHIQEIKEVKEDEI TDKTIRDLAKLDKIIM
Q3SZ47	MFLSAVTFAKSKSKTILVKMVSQAGTGFSFNTKRSRLWEKLTLLHYDPVVKKKVLFVEQKKIRSL
P20084	MVFYKVTLSRSLIGVPHTTKSIVKSLGLGKRGSIVYKKVNPAIAGSLAKVKELVKVEVTEHELTPSQQR ELRKSNPGFIVEKRTID
A8NN94	MAFSVGSVGCLLGPVSRSAGLLGGRWLQGSRAWLGLPDTRRLPVIQQTRGRTRGNEYQPSNIKRKNK HGWIRRLSTPNGVQVILRRMHKGRKSLSH
P36533	MVKVKSKNSVIKLLSTAASGYSRYISIKKGAPLVTQVRYDPVVKRHVLFKEAKKRKVAERKPLDFLRT AK
P19956	MITKYFSKVIVRFNPFGKEAKVARLVLAAIPPTQRNMGTQIQSEIISDYNKVKPLVKVTYKDKKEMEV DPSNMNFQELANHFDRHSKQLDLKHMLEMH
P03049	MARDDPHFNFRMPMEVREKLKFRAEANGRSMNSELLQIVQDALSKPSPVTGYRNDAERLADEQSELV KKMVFDTLKDLYKKTT
O26119	MITPRNIFRHELIGLSVRIARSVHRDIQGISGRVVDETRNTLRIEMDDGREITVPKGIAVFHFRTPQGEL VEIDGRALVARPEERIKKKFRKP
P05798	DVSGTVCLSALPPEATDTLNLIASDGPFPYSQDGVVFQNRESVLPTQSYGYYHEYTVITPGARTRGTR RIITGEATQEDYYTGDHYATFSLIDQTC
P0AFW8	MNDTYQPINCDDYDNLELACQHHLMLTLELKDGEKLQAKASDLVSRKNVEYLVVEAAGETRELRLDK ITSFSHPEIGTVVVSES
P03051	MTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARFGDDGENL
P86508	SVSNIPESIGF
P86511	ASAAGAVRAGDDETLLKPVLNSLDNLVSGL
P86515	AEILFGDVRPPWMPPPIFPEMP
P68317	MVFQLVCSTCGKDISHERYKLIIRKKSLKDVLVSVKNECCRLKLSTQIEPQRNLTVQPLLDIN
P48011	MNHPTSTGGTAFNPPRPATMIYLCADCGARNTIQAKEVIRCRECGHRVMYKMRTKRMVQFEAR
P40422	MSREGFQIPTNLDAAAAGTSQARTATLKYICAECSSKLSLSRTDAVRCKDCGHRILLKARTKRLVQFE AR
O13877	MIIPIRCFSCGKVIGDKWDTYLTLLQEDNTEGEALDKLGLQRYCCRRMILTHVDLIEKLLCYNPLSKQK NL
P03042	MVRANKRNEALRIESALLNKIAMLGTEKTAEAVGVDKSQISRWKRDWIPKFSMLLAVLEWGVVDDD MARLARQVAAILTNKKRPAATERSEQIQMEF
P18681	MMHFQLAGSGVMSAFYPHESELSRRVKQLIRAAKKQLEALCAMK
P03044	MQYAIAGWPVAGCPSESLLERITRKLRDGWKRLIDILNQPGVPKNGSNTYGYPD
Q57839	MRACLKCKYLTNDEICPICHSPTSENWIGLLIVINPEKSEIAKKAGIDIKGKYALSVKE
P11521	MRSSSKKKIDISNHELVPKHEILQLEEAYKLVKELGIKPEQLPWIRASDPVAKSIGAKPGDIIKITRKSPF TGESVTYRYVITG
Q97ZJ9	MGLERDEILSQDLHFNEVFISLWQNRLTRYEIARVISARALQLAMGAPALIDINNLSSTDVISIAEEEFRR GVLPITIRRRLPNGKIILLSLRKS
Q57832	MEIKILERKDNLVEIELINEDHSLPNLLKDILLTKEGVKMASYSIDHPLLHPETGRYISNPKITIITEEGTD PLEVLKEGLRDIIKMCDTLLDELKEKK
Q980K0	MEIRILKSESNYLELEIEGEDHTLGNLIAGTLRRISGVSFASYYQPHPLSDKIIVKILTDGSITPKDALLKAI ENIRGMTSHYIDEIKGLTK
P59283	MVEYKCLNCKKIIKLEELGKRARCPHCSYKILVKLRPKVVKHVKAR
Q97ZX7	MAVYRCGKCWKTFTDEQLKVLPGVRCPYCGYKIIFMVRKPTIKIVKAI
P0A800	MARVTVQDAVEKIGNRFDLVLVAARRARQMQVGGKDPLVPEENDKTTVIALREIEEGLINNQILDVRE RQEQQEQEAAELQAVTAIAEGRR
P66726	MLNPPLNQLTSQIKSKYLIATTAAKRAREIDEQPETELLSEYHSFKPVGRALEEIADGKIRPVISSDYYG KE
Q9EVV4	MAEPGIDKLFGMVDSKYRLTVVVAKRAQQLLRHRFKNTVLEPEERPKMRTLEGLYDDPNAVTWAM KELLTGRLFFGENLVPEDRLQKEMERLYPTEEEA
Q24475	MAVAFYIPDQATLLREAEQKEQQILRLRESQWRFLATVVLETLRQYTSCHPKTGRKSGKYRKPSQ

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9GGE2	MAKKSMIQRELKRQKLVMKYATKRAALKEQIKQTTFLKEKLSLHRKLQQLPRNSSAVRLHNRCMITG RPKGYFRDFGLSRHVLREMAHQGLLPGVCKSSW
Q9M3I4	MKKNSFISVISDEKKEENKGSVEFQVFCFTNKIRRLTLHLELHKKDYSSQRGLRKTLGKRQRLLAYLLK INGVRYKELISKLNIRELKTR
P28807	MVKLRLKRCGRKQRAVYRIVAIDVRSRREGRDLQKVGFYDPIKSQTYLNVPAILDFLEKGAQPTETVY DILKRAEVFKEFRLNQTKFN
Q10360	MSNPPTRPTLYNLVNADGELSYRARDLVQDFSVPIPHELPDPDLVHSIQDFATEYMLATGKKSFECLD ESALLGLGYLVTEWMDAMVTDCLEEFEER
P54798	MAKKSMIIKQKRTPKFKVRAYTRCERCGRPHSVYRKFKLCRICFRELAYKGQLPGIKKASW
Q0PA13	MALDSAKKAEIVAKFAKKPGDTGSTEVQVALLTARIAELTEHLKIYKKDFSSRLGLLKLVGQRKRLLS YLKRKDYNSYSKLITELNLRDK
P0A7T3	MVTIRLARHGAKKRPFYQVVVADSRNARNGRFIERVGFFNPIASEKEEGTRLDLDRIAHWVGQGATIS DRVAALIKEVNKAA
P66439	MAVKIRLTRLGSKRNPFYRIVVADARSPRDGRIIEQIGTYNPTSANAPEIKVDEALALKWLNDGAKPT DTVHNILSKEGIMKKFDEQKKAK
Q5V5R5	MAIKPAYVKKTGTLLMERYPDAFGADFEHNKDVVEELTNIESKGVRNRIAGYVTRKMNNPVEA
O26894	MGNIRTSFVKRIAKEMIETHPGKFTDDFDTNKKLVEEFSTVSTKHLRNKIAGYITRIISQQK
A0QSE0	MADQKGPKYTPAAEKPRGRRKTAIGYVVSDKMQKTIVVELEDRKSHPLYGKIIRTTKKVKAHDENGE AGIGDRVSLMETRPLSATKRWRLVEILEKAK
A0R549	MMAVKKSRKRTAATELKKPRRNQLEALGVTTIDYKDVAVLRTFLSERGKIRSRHVTGLTPQQQRQVA TAIKNAREMALLPMAGPR
Q6N5A3	MAEAGARRPFFRRRKTCPFTGANAPKIDYKDSKLLMRYVSERGKIVPSRITAVSAKKQRELARAIKRA RFLGLLPYVIR
P62659	MSTKNAKPKKEAQRRPSRKAKVKATLGEFDLRDYRNVEVLKRFLSETGKILPRRRTGLSAKEQRILA KTIKRARILGLLPFTEKLVRK
O83223	MSRSVKKGPFVDKKLYKRVVEMNKAANQRNKKVIKSYSRCSTIIPEMVGFTISVHNGKSWIPVYITEE FVGHKLGEFSPTRVFRGHSGSDKKVGR
P75237	MANIKSNEKRLRQNIKRNLNNKGQKTKLKTNVKNFHKEINLDNLGNVYSQADRLARKGIISTNRARRL KSRNVAVLNKTQVTAVEGK
A0R102	MANIKSQIKRIRTNERRRLRNQSVKSSLRTAIRGFREAVDAGDKDKASELLHATSRKLDKAASKGVIHP NQAANKKSALALALNKL
Q9HVM1	MANTPSAKKRAKQAEKRRSHNASLRSMVRTYIKNVVKAIDAKDLEKAQAAFTAAVPVIDRMADKGII HKNKAARHKSRLSGHIKALSTAAA
Q6N0C7	MANTSSAKKATRKIARRTAVNKSRRTQMRGSVRIVEEAIASGDRDAALKAMARAEPELMRAAQRNIIH RNAASRKVSRLTHSIAKLAK
P0A4B8	MTQIVVGENEHIESALRRFKREVSKAGIFQDMRKHRHFETPIEKSKRKKLALHKQSKRRFRT
P49197	MQNDAGQTVELYVPRKCSSSNRIIGPKDHASVQIDFVDVDPETGRMIPGKSTRYAICGAIRRMGESDD AILRLAQKDGLVPRDDVKSN
P68679	MPVIKVRENEPFDVALRRFKRSCEKAGVLAEVRRREFYEKPTTERKRAKASAVKRHAKKLARENAR RTRLY
Q6N274	MQVLVRDNNVDQALKALKKKMQREGIFREMKLRGHYEKPSEKKAREKAEAVRRARKLARKKLQRE GLLPSKPKPAFGADRRPSAAAR
Q9UY20	MEIKITEVKENKLIGRKEIYFEIYHPGEPTPSRKDVKGKLVAMLDLNPETTVIQYIRSYFGSYKSKGYA KYYYDKDRMLYIEPEYILIRDGIIEKKEGE
Q9HJ79	MDLIIKEKRDNPILKRKEIKYVLKFDSSRTPSREEIKELIAKHEGVDKELVIVDNNKQLTGKHEIEGYTKI YADKPSAMLYEPDYELIRNGLKQKEAK
Q8L953	MVLQNDIDLLHPPPELEKRKHKLKRLVQSPNSFFMDVKCQGCFNITTVFSHSQTVVVCGNCQTVLCQ PTGGKARLQEGCSFRKK
Q06125	MPHNEYYNDDGELDRETCPRCGDTVLAEHEDRQHCGKCGYTEWK
Q9HJ78	MQKRELYEIADGKLVRKHRFCPRCGPGVFLAEHADRYSCGRCGYTEFKKAKKSKS
O28935	MHSRFVKVKCPDCEHEQVIFDHPSTIVKCIICGRTVAEPTGGKGNIKAEIIEYVDQIE
P61030	MAEDEGYPAEVIEIIGRTGTTGDVTQVKVRILEGRDKGRVIRRNVRGPVRVGDILILRETEREAREIKS RR
P41057	MAHENVWFSHPRRYGKGSRQCRVCSSHTGLIRKYGLNICRQCFREKANDIGFNKFR
A0R7F9	MVILDPTLDERTVAPSLETFLNVIRKDGGTVDKVDIWGRRRLAYEIAKHAEGIYAVIDVKAEPATVSEL DRQLNLNESVLRTKVLRTDKH
P66595	MRTYEVMYIVRPNIEEDAKKALVERFNGILATEGAEVLEAKDWGKRRLAYEINDFKDGFYNIVRVKS DNNKATDEFQRLAKISDDIIRYMVIREDEDK
P62611	MGKGDRRTRRGKIWRGTYGKYRPRKKK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
C0H3R4	MDGQFEQKKKQKDETYDIEHLIACFSPMIRKKLSNTSYQEREDLEQELKIKMFEKADMLLCQDVPGF WEFILYMVDENS
P53733	MQPAARLLSRSVWKGPNIVPLPIREAMTKGTPIRTNARAATILPQFVGLKFQIHNGKEYVPIEISEDMV GHKLGEFAPTRKRFSYTQTKNK
P82920	MAKHLKFIARTVMVQEGNVEGAYRTLNRILTMDGLIEDIKRRRYYEKPCRRRQRESYETCRRIYNME MARKINFLMRKNRADPWQGC
O14464	MFLQTLRLTMPRMFLHMKPSPITITRACTVPSLLSVAAPQPALVAANRPLVFNRGFKVRTSVKKFCSD CYLVRRKGRVYIYCKSNKKHKQRQG
O31466	MVIATDDLEVACPKCERAGEIEGTPCPACSGKGVILTAQGYTLLDFIQKHLNK
P04170	MQKYVCNVCGYEYDPAEHDNVPFDQLPDDWCCPVCGVSKDQFSPA
Q93PP8	MAEPQDMWRCQMVNCGYVYDPDRGDKRRKVPAGTKFEDLPEDWRCPVCGAGKKSFRRLSDEA
P00267	MQKFECTLCGYIYDPALVGPDTPDQDGAFEDVSENWVCPLCGAGKEDFEVYED
Q12330	MSNKVKTKAMVPPINCIFNFLQQQTPVTIWLFEQIGIRIKGKIVGFDEFMNVVIDEAVEIPVNSADGKE DVEKGTPLGKILLKGDNITLITSAD
P62306	MSLPLNPKPFLNGLTGKPVMVKLKWGMEYKGYLVSVDGYMNMQLANTEEYIDGALSGHLGEVLIRC NNVLYIRGVEEEEEDGEMRE
O74966	MSKAGAPDLKKYLDRQVFVQLNGSRKVYGVLRGYDIFLNIVLEDSIEEKVDGEKVKIGSVAIRGNSVIM IETLDKMT
P46350	MSYDKVSQAKSIIIGTKQTVKALKRGSVKEVVVAKDADPILTSSVVSLAEDQGISVSMVESMKKLGKAC GIEVGAAAVAIIL
Q37935	MLKLKMMLCVMMLPLVVVGCTSKQSVSQCVKPPPPPAWIMQPPPDWQTPLNGIISPSERG
P29377	MSTKKSPEELKRIFEKYAAKEGDPDQLSKDELKLLIQAEFPSLLKGPNTLDDLFQELDKNGDGEVSFEE FQVLVKKISQ
Q8WXG8	MPTQLEMAMDTMIRIFHRYSGKERKRFKLSKGELKLLLQRELTEFLSCQKETQLVDKIVQDLDANKD NEVDFNEFVVMVAALTVACNDYFVEQLKKKGK
P82978	SADEQKLRERFEALDKDKSGTLSVDELYEGVHAVHPKVSRNDIVKIIEKVDTNKDGQVSWQEFIEAFK RLADLKL
P29034	MMCSSLEQALAVLVTTFHKYSCQEGDKFKLSKGEMKELLHKELPSFVGEKVDEEGLKKLMGSLDENS DQQVDFQEYAVFLALITVMCNDFFQGCPDRP
P04163	MPSQMEHAMETMMFTFHKFAGDKGYLTKEDLRVLMEKEFPGFLENQKDPLAVDKIMKDLDQCRDG KVGFQSFFSLIAGLTIACNDYFVVHMKQKGK
P31950	MAKRPTETERCIESLIAIFQKHAGRDGNNTKISKTEFLIFMNTELAAFTQNQKDPGVLDRMMKKLDLD SDGQLDFQEFLNLIGGLAIACHDSFIKSTQK
P79105	MTKLEDHLEGIINIFHQYSVRVGHFDTLNKRELKQLITKELPKTLQNTKDQPTIDKIFQDLDADKDGAV SFEEFVVLVSRVLKTAHIDIHKE
Q99584	MAAEPLTELEESIETVVTTFFTFARQEGRKDSLSVNEFKELVTQQLPHLLKDVGSLDEKMKSLDVNQD SELKFNEYWRLIGELAKEIRKKKDLKIRKK
P81552	MAVEEIVKVSRNYQVTIPAKVRQKFQIKEGDLVKVTFDESGGVVKIQLLDSKTDAH
P19707	EWYSFLGEAAQGAWDMWRAYSDMREANYIGADKYFHARGNYDAAQRGPGGAWAAKVISDARENS QRVTDFFRHGNSGHGAEDSKADQEWG
P89114	MNYLETQLNKKQKQIQEYESMNGNLIKMFEQLSKEKKNDETPKKISSTYIKELKEYNELRDAGLRLAQ IIADEKQCKIKDVFEEIGYSMKD
P76136	MHATTVKNKITQRDNYKEIMSAIVVVLLLTLTLIAIFSAIDQLSISEMGRIARDLTHFIINSLQG
D4GUF6	MEWKLFADLAEVAGSRTVRVDVDGDATVGDALDALVGAHPALESRVFGDDGELYDHINVLRNGEAA ALGEATAAGDELALFPPVSGG
D4GZE7	MNVTVEVVGEETSEVAVDDDGTYADLVRAVDLSPHEVTVLVDGRPVPEDQSVEVDRVKVLRLIKGG
P11760	CVTGAPGCVGGGRL
O00631	MGINTRELFLNFTIVLITVILMWLLVRSYQY
P21887	MSQHLVPEAKNGLSKFKNEVANEMGVPFSDYNGDLSSRQCGSVGGEMVKRMVEKYEQSMK
P02960	MANYQNASNRNSSNKLVAPGAQAAIDQMKFEIASEFGVNLGPDATARANGSVGGEITKRLVQLAEQN LGGKY
P02961	MAKQTNKTASGTSTQHVKQQNAQASKNNFGTEFGSETNVQEVKQQNAQAAANKSQNAQASKNNFG TEFASETSAQEVRQQNAQAQAKKNQNSGKYQG
Q9SK39	MEFTAEQLSQYNGTDESKPIYVAIKGRVFDVTTGKSFYGSGGDYSMFAGKDASRALGKMSKNEEDVS PSLEGLTEKEINTLNDWETKFEAKYPVVGRVVS
P59899	MNSLLMITACLALVGTVWAKEGYLVNHSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGC WCTHLYEQAVVWPLPKKTCNGK

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P60467	MPGPTPSGTNVGSSGRSPSKAVAARAAGSTVRQRKNASCGTRSAGRTTSAGTGGMWRFYTEDSPGL KVGPVPVLVMSLLFIASVFMLHIWGKYTRS
P35179	MARASEKGEEKKQSNNQVEKLVEAPVEFVREGTQFLAKCKKPDLKEYTKIVKAVGIGFIAVGIIGYAI KLIHIPIRYVIV
P52871	MAASVPPGGQRILQKRRQAQSIKEKQAKQTPTSTRQAGYGGSSSSILKLYTDEANGFRVDSLVVLFLSV GFIFSVIALHLLTKFTHII
P24471	DNYWCPQSGEAFECFESDPNAKFCLNSGKTSVVICSKCRKKYEFCRNGLKVSKRPDYDCGAGWESTP CTGDNSAVPAVF
P81761	MKFLYGVILIALFLTVMTATLSEARCGPCFTTDPQTQAKCSECCGRKGGVCKGPQCICGIQY
Q7Z0H4	MKIFFAVLVILVLFSMLIWTAYGTPYPVNCKTDRDCVMCGLGISCKNGYCQGCTR
P0C175	MKFSCGFLLIFLVLSAMIATFSEVEATVKCGGCNRKCCAGGCRSGKCINGKCQCYGRSDLNEEFENYQ
C9X4K7	MLKFAIAVALLLFIGLELREARDGYPQSKVNYCKIYCPNTTVCQWTCKNRAGATDGDCRWSSCYCFN VAPDTVLYGDPGTKPCMA
P01159	SDNNQQGKSAQQGGY
P0CG07	MRCVVLFMVSCLLIVLLINHFEEVEAQKWNKCFLRDIFPGKCEHDANAKLRCKEDDAKKTLA
O75711	MKLMVLVFTIGLTLLLGVQAMPANRLSCYRKILKDHNCHNLPEGVADLTQIDVNVQDHFWDGKGCE MICYCNFSELLCCPKDVFFGPKISFVIPCNNQ
Q45966	MERQNVQQQRGKDQRPQRPGASNPRRPNQR
P63019	RDGYPLASNGCKFGCSGLGENNPTCNHVCEKKAGSDYGYCYAWTCYCEHVAEGTVLWGDSGTGPC RS
Q5MJP5	MKTIPLLFLLFIYFECDGKFIRHKDESFYECGQLIGYQQYCVDACQAHGSKEKGYCKGMAPFGLPGGC YCPKLPSNRVKMCFGALESKCA
Q86SE0	MNYLITISLALLLMTGVASGVRDGYIADAGNCGYTCVANDYCNTECTKNGAESGYCQWFGRYGNAC WCIKLPDKVPIKVPGKCNGR
P56743	GRDGYVVKNGTNCKYSCEIGSEYEYCGPLCKRKNAKTGYCYAFACWCIDVPDDVKLYGDDGTYCSS
B7SNV8	MNYLLVLTLASLLALGVNGKKDGYPVDHANCKYECWYDDKYCDDLCKKRKADSGYCYKLNISCYCL GLPDNAAIKDYGRCRP
Q8WRY4	MNSLLMITACLILIGTVWAEDGYLFDKRKRCTLACIDKTGDKNCDRNCKKEGGSFGHCSYSACWCKG LPGSTPISRTPGKTCKK
P58296	ARDGYPVDEKGCKLSCLINDKWCNSACHSRGGKYGYCYTGGLACYCEAVPDNVKVWTYETNTC
B8QG00	MKTSSLTIIFIAVIITIICLNIHDIEAREIEFNAGRVVRSEKDCIKHLQRCRENKDCCSKKCSRRGTNPEKR CR
Q7WY62	MNWVPSMRKLSDELLIESYFKATEMNLNRDFIELIENEIKRRSLGHIISVSS
Q3E785	MPKRLSGLQKEVLHLYRASIRTAHTKPKENQVNFVNYIHEEFGKYRNLPRKDFTTIEHLLRVGNKKIA TFSHPELTNIH
D0VWU4	MEKLKEFLKGVRDELKRVVWPSRELVVKATISVIIFSLAIGVYLWILDLTFTKIISFILSLRGSL
Q57817	MKTDFNQKIEQLKEFIEECRRVWLVLKKPTKDEYLAVAKVTALGISLLGIIGYIIHVPATYIKGILKPPT TPRV
Q8TZK2	MAELQERIRHFWKESRRAFLVTKKPNWATYKRAAKITGLGIILIGLIGMLIRIVGILILGG
P35874	MEKLRKFFREVIAEAKKISWPSRKELLTSFGVVLVILAVTSVYFFVLDFIFSGVVSAIFKALGIG
P38383	MFARLIRYFQEARAELARVTWPTREQVVEGTQAILLFTLAFMVILGLYDTVFRFLIGLLR
O66505	MYYALLTLFVIIAVVLIISTLLQKGRGDVGAAFGGGMGQSIFGVGGVETILTKATYWLGALFLVLALLL SVIPKEKGSVVEKSVQTEQSEGKGTTQESGK
P60460	MSKREETGLATSAGLIRYMDETFSKIRVKPEHVIGVTVAFVIIEAILTYGRFL
P02852	MKNYSKNATHLITVLLFSFVVILLIIPSKCEAVSNDMQPLEARSADLVPEPRYIIDVPPRCPPGSKFIKNR CRVIVP
Q9XIR8	MAAEPKAATAEVVKMDLFEDDDEFEEFEINEDWLEKEEVKEVSQQWEDDWDDDDVNDDFSRQLRK ELENGTDKK
O94742	MSTDVAAAQAQSKIDLTKKKNEEINKKSLEEDDEFEDFPIDTWANGETIKSNAVTQTNIWEENWDDV EVDDDFTNELKAELDRYKRENQ
O88892	MARGNQREIARQKNMKKTQEISKGKRKEDSLTASQRKQRDSEIMQQKQKIANEKKSMQTTEK
Q9R2C1	MVAKQRIRMANEKHSKNITQRGNVAKTSRNAPEEKASVGPWLLALFIFVVCGSAIFQIIQSIRMGM
Q9BWJ5	MTDRYTIHSQLEHLQSKYIGTGHADTTKWEWLVNQHRDSYCSYMGHFDLLNYFAIAENESKARVRFN LMEKMLQPCGPPADKPEEN
P61095	KECMTDGTVCYIHNHNDCCGSCLCSNGPIARPWEMMVGNCMCGPKA
P43682	MSNSRYSQTESNNDRKLEGLANKLATFRNINQEIGDRAVSDSSVINQMTDSLGSMFTDIKNSSSRLTRSL KAGNSIWRMVGLALLIFFILYTLFKLF

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q6UW10	MGSGLPLVLLLTLLGSSHGTGPGMTLQLKLKESFLTNSSYESSFLELLEKLCLLLHLPSGTSVTLHHARS QHHVVCNT
Q4GWU5	MATTMAKLITLVVLAILAFVEVSVSGYKTSISTITIEDNGRCTKSIPPICFPDGRP
Q7M742	MKGSSALLLVALSLLCVCGLTRAEDDNEFFMEFLQTLLVGTPEELYEGPLGKYNVNDMAKSALRELKS CIDELQPVHKEQLVKLLVQVLDAQEDT
O95968	MRLSVCLLLLTLALCCYRANAVVCQALGSEITGFLLAGKPVFKFQLAKFKAPLEAVAAKMEVKKCVD TMAYEKRVLITKTLGKIAEKCDR
O75556	MKLLMVLMLAALLLHCYADSGCKLLEDMVEKTINSDISIPEYKELLQEFIDSDAAAEAMGKFKQCFLN QSHRTLKNFGLMMHTVYDSIWCNMKSN
P02780	MKLVFLFLLVTIPICCYASGSGCSILDEVIRGTINSTVTLHDYMKLVKPYVQDHFTEKAVKQFKQCFLD QTDKTLENVGVMMEAIFNSESCQQPS
Q96PL1	MKLVTIFLLVTISLCSYSATAFLINKVPLPVDKLAPLPLDNILPFMDPLKLLLKTLGISVEHLVEGLRKCV NELGPEASEAVKKLLEALSHLV
Q03067	MTEETITIDSISNGILNNLLTTLIQDIVARETTQQQLLKTRYPDLRSYYFDPNGSLDINGLQKQQESSQYI HCENCGRDVSANRLAAHLQRCLSRGARR
O46163	MAKLLAVFLVLLIAALVCEQALACTPGSRKYDGCNWCTCSSGGAWICTLKYCPPSSGGGLTFA
C1P5Z7	MRQFYQHYFTATAKLCWLRWLSVPQRLTMLEGLMQWDDRNSES
Q9H299	MSGLRVYSTSVTGSREIKSQQSEVTRILDGKRIQYQLVDISQDNALRDEMRALAGNPKATPPQIVNGDQ YCGDYELFVEAVEQNTLQEFLKLA
P04852	MENTSITIEFSSKFWPYFTLIHMITTIISLLIIISIMIAILNKLCEYNVFHNKTFELPRARVNT
P22109	MPAIQPPLYLTFLLLILLYLIITLYVWTILTINHKTAVRYAALYQRSCSRWGFDQSL
P86128	GIAEFLNYIKSKA
P23308	MKNAKQEHFELDQEWVELMVEAKEANISPEEIRKYLLLNKKSAHPGPAARSHTVNPF
P50263	MSNMMNKFAEKLQGNDDSHQKGKNAKSSNKERDDMNMDMGMGHDQSEGGMKMGHDQSGTKMNA GRGIANDWKTYENMKK
P42579	MLSSVALRYLLVLSLAFLAVVTSSRTQSRFASYELMGTEGTECVTTKTISQICYQCATRHEDSFVQVYQ ECCKKEMGLREYCEEIYTELPIRSGLWQPNK
P56637	MKFFLMCLIIFPIMGVLGKKNGYPLDRNGKTTECSGVNAIAPHYCNSECTKVYYAESGYCCWGACYC FGLEDDKPIGPMKDITKKYCDVQIIPS
Q8GWU7	MCYGHNQSLSSRSSLRRRSHDGEDDSVVDDLRDRLAETEARLRRARAREAELSRRLEHMKRFVSVMEI IETFLERRFQEQKDRIARLFSPVSTK
P84826	RHHRKRIGHTVKQLAKLVKHIHEY
P84828	ILCINVAGRRIC
P84830	ITKDFDALMKYIKRI
P63293	YADAIFTNSYRKVLGQLSARKLLQDIMNRQQGERNQEQGAKVRL
P42692	HADGMFNKAYRKALGQLSARKYLHTLMAKRVGGGSMIEDDNEPLS
Q3E784	MSAENISTGSPTGKQPSSEVNLGEREAGTKNERMMRQTKLLKDTLDLLWNKTLEQQEVCEQLKQEN DYLEDYIGNLMRSSNVLEK
P0C8M5	MKTHVKKDLDKGWHMLIQEARSIGLGIHDVRQFLESETASRKKNHKKTVRQD
Q8PAH9	MHEQLSPRDQELEARLVELETRLSFQEQALTELSEALADARLTGARNAELIRHLLEDLGKVRSTLFADA ADEPPPPHY
Q0VAQ4	MTSLLTTPSPREELMTTPILQPTEALSPEDGASTALIAVVITVVFLTLLSVVILIFFYLYKNKGSYVTYEP TEGEPSAIVQMESDLAKGSEKEEYFI
Q9UUC6	MSLCIKLLHETQGHIVTMELENGSTYRGKLIEAEDNMNCQMRDISVTARDGRVSHLDQVYIRGSHIRFL IVPDMLRNAPMFKVGPGRSVPLPTRGRR
P21779	ALRAAAVAGSPQQLLPLGQRERKAGCKNFFWKTFSSC
P56508	MHARDWFLVFIAIFIPPLAVWLKRGFFTKDLLINFLLFLLGFFPGLIHALYVISCHPYEENEARYSHLSSS DDNYGSLA
O75324	MSIMDHSPTTGVVTVIVILIAIAALGALILGCWCYLRLQRISQSEDEESIVGDGETKEPFLLVQYSAKGP CVERKAKLMTPNGPEVHG
O75971	MLSRLQELRKEEETLLRLKAALHDQLNRLKVEELALQSMISSRRGDEMLSSHTVPEQSHDMLVHVDNE ASINQTTLELSTKSHVTEEEEEEEEEESDS
P86444	SNRSPSLRLRF
P86981	GYKHQCRDRCYGTCGTSCKYGASRPSCACALHGGYCCVARPYHYPYPHPRPVPLPAPAPRPAPHYP VHHPKWPHWRPHYKA
Q9Y675	MERARDRLHLRRTTEQHVPEVEVQVKRRRTASLSNQECQLYPRRSQQQQVPVVDFQAELRQAFLAE TPRGG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q05676	MAPPTTIRTRDQALAPLATLDSQTNCRLKELVQWECQFKGAEYVCSPFKRLFEHCIAPDKSATNYEV TDTYTNS
P0C2P6	DCSQDCAACSILARPAELNTETCILECEGKLSSNDTEGGLCKEFLHPSKVDLPR
P83567	PKDSMLLLQVPVY
P05043	MGGSSEQERLLVSIDEKRKLMIDAARKQGFTGHDTIRHSQELDCLINEYHQLMQENEHSQGIQGLVKK LGLWPRRDVMPAYDANK
Q81DD0	MAEVNVQKSSFFKEKKEESNTDFSLVKGALTENINRLEKLMNNSSSKYIQVKRTKENA
O34800	MERAFQNRCEPRAAKPFKILKKRSTTSVASYQVSPHTARIFKENERLIDEYKRKKA
P15281	MHDYIKERTIKIGKYIVETKKTVRVIAKEFGVSKSTVHKDLTERLPEINPDLANEVKEILDYHKSIRHLR GGEATKLKYKKDEILEGEPVQQS
P24056	MSKVSGGSRRTRARRPMSNRRGRRSQSAAHRSRAQRRRRRTGTTRRARTSTARRARTRTARRSDLT RMMARDYGSDYRS
Q99WA5	MKVTDVRLRKIQTDGRMKALVSITLDEAFVIHDLRVIEGNSGLFVAMPSKRTPDGEFRDIAHPINSDMR QEIQDAVMKVYDETDEVVPDKNATSEDSEEA
P39230	MKKFIFATIFALASCAAQPAMAGYDKDLCEWSMTADQTEVETQIEADIMNIVKRDRPEMKAEVQKQL KSGGVMQYNYVLYCDKNFNNKNIIAEVVGE
P46965	MSEILQDVQRKLVFPIDFPSQRKTEKFQQLSLMIGALVACILGFAQQSLKVLLTAYGISCVITLICVLPAY PWYNKQKLRWAQPKIEINVDQYD
Q9VAL0	MLDIQTHMDFAGQGKAERWSRFIITFFGIVGLVYGAFVQQFSQTVYILGAGFVLSSLITIPPWPLYRRN ALKWQKPIDTDAKSSSSESGDEGKKKKKQ
P80304	ARSFSSYCVRCRRKTPSFNSKTVTFRNKRRAIRSHCAYCQVKKFRIIGHGG
P50088	MKLDSGIYSEAQRVVRTPKFRYIMLGLVGAAVVPTAYMRRGYTVPAHSLDNINGVDTTKASVMGTEQ RAAMTKGKSLQEMMDDDEVTYLMFSSIM
B9A8D7	MKVAVVVALLCFVCYTAAETCSADGDCKNTICDASHDLECHRGQCTCVNHATACSSAADCSGSCTIF GRHGRWHCVDAKCRCFFV
Q8CEK3	MSSTWIKFLFILTLVLLPYSVFSVNIFAGPENVIKEPNCTMYKSKSECSNIAENPVCADDRNTYYNECYF CIEKVVEKLKYRYHGICIYK
P85064	SEDDGSVSPESQEMSYLELPLPSISPLIYAPVCVEDSNSDFYLFVNECEVRKCGCEAGFVYTFVPREMC KATTSLCPMQTKSS
P81728	MVDLWPVSTRPRVLFRAAPSLLNSASAASMLWAELIAMAPSIPSALPPFS
Q8NFR3	MDLRRVKEYFSWLYYQYQIISCCAVLEPWERSMFNTILLTIIAMVVYTAYVFIPIHIRLAWEFFSKICGY HSTISN
P68191	MKSNRQARHILGLDHKISNQRKIVTEGDKSSVVNNPTGRKRPAEK
P21262	MAQYQTWEEFSRAAEKLYLADPMKARVVLKYRHSDGSLCIKVTDDLVCLVYRTDQAQDVKKIEKFH SQLMRLMVAKESRSVAMETD
Q58440	MIIWPSYIDKKKSRREGRKVPEELAIEKPSLKDIEKALKKLGLEPKIYRDKRYPRQHWEICGCVEVDYK GNKLQLLKEICKIIKGKN
Q8TZT9	MGRFVVWPSELDSRLSRKYGRIVPRSIAVESPRVEEIVRAAEELKFKVIRVEEDKLNPRLSGIDEELRTF GMIVLESPYGKSKSLKLIAQKIREFRRRSA
Q3UN54	MEKYLLLLLLGIFLRVGFLQALTCVSCGRLNSSGICETAETSCEATNNRKCALRLLYKDGKFQYGFQG CLGTCFNYTKTNNNMVKEHKCCDHQNLCNKP
P86608	KPENENEEALHE
Q7WY59	MSENRHENEENRRDAAVAKVQNSGNAKVVVSVNTDQDQAQAQSQDGED
O31552	MNIQRAKEIVESPDMKKVTYNGVPIYIQHVNEETGTARIYPLDEPQEEHEVQLNSLKED
P94537	MDLNLRHAVIANVTGNNQEQLEHTIVDAIQSGEEKMLPGLGVLFEVIWQHASESEKNEMLKTLEGGLK PAE
Q7WY58	MGFFNKDKGKRSEKEKNVIQGALEDAGSALKDDPLQEAVQKKKNNR
Q7WY75	MVRNKEKGFPYENENKFQGEPRAKDDYASKRADGSINQHPQERMRASGKR
Q7WY66	MKKKDKGRLTGGVTPQGDLEGNTHNDPKTELEERAKKSNTKR
Q7WY65	MKTRPKKAGQQKKTESKAIDSLDKKLGGPNRPST
P71031	MVKRKANHVINGMNDAKSQGKGAGYIENDQLVLTEAERQNNKKRKTNQ
P11020	VKSRYHQRQYRARKRYAKARRTKKPKRRPKPPRKLRYAPSKKQPKIMKLKLDNEVDNTLKAKNKSL NEALKNRLSLRKHV
P02808	MKFLVFAFILALMVSMIGADSSEEKFLRRIGRFGYGYGPYQPVPEQPLYPQPYQPQYQQYTF
P17306	MSTSRKLKSHGMRRGKNRAPHKGVKRGGSKRKYRKGSLKSRKRCDDANRNYRSHL
Q7M2P1	AKVTEKSWQPQTTSTKRWKKRKTPSQPRSRGKVRKIYKKVKRPLHVCSRKKYSPKVITTSRRQKRA RRANKFETIP

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P09386	MKKMFMAVLFALASVNAMAADCAKGKIEFSKYNEDDTFTVKVDGKEYWTSRWNLQPLLQSAQLTG MTVTIKSSTCESGSGFAEVQFNND
P55852	MSANQEEDKKPGDGGAHINLKVKGQDGNEVFFRIKRSTQLKKLMNAYCDRQSVDMNSIAFLFDGRRL RAEQTPDELDMEDGDEIDAMLHQTGGSGGGATA
Q6WNK7	MTMDTAQLKSQIQQYLVESGNYELISNELKARLLQEGWVDKVKDLTKSEMNINESTNFTQILSTVEPK ALEMVSDSTRETVLKQIREFLEEIVDTQ
P0C0A9	MGLCFPCPAESAPPSPSPEEKRAKLAEAAERRQKEAASRGILDIQSVEAKKKKKEQLEKQMETSGPPA GGLRWTVS
Q8MMH5	MRCVAIFLVVICAFVLQALAEVQCPASGASDQDNLDFCMEYPELIEKCELKSCEGFIKVVS
Q8MMH4	MKFIVLLGALLALLVAVSADRIAREAPEMESVDEAVLTRQAREAEDPAVVEDAIRKFVRWLVQKYGIN IHDLIDHFKH
Q8K3D3	MIDENNDVSEEALSSDIKKLKEKHDMLDKEISQLIAEGYRVIELEKHISLLHEYNDIKDVSQMLLGKLAV TRGVTTKELYPDFDLNLND
Q9UUB7	MEKSQLESRVHLLEQQKEQLESSLQDALAKLKNRDAKQTVQKHIDLLHTYNEIRDIALGMIGKVAEHE KCTSVELFDRFGVNGSE
B5KM66	MADSDPGERSYDNMLKMLSDLNKDLEKLLEEMEKISVQATWMAYDMVVMRTNPTLAESMRRLEDAF LNCKEEMEKNWQELLTETKRKQ
P0C1R2	SIKEKICKIIEAKIGKKPPFCP
P80220	MDLVKNHLMYAVREEVEILKEQIRELVEKNSQLERENTLLKTLASPEQLEKFQSRLSPEEPAPETPEAP EAPGGSAV
Q5D233	MLKFAVLCFLVIMASTFAQKCGDQVCGAGTCCAEYPEIHCKRVGQLYDICVDSEATKDSGNHLFFCP CDEGMYCDMNSWSCQKKTSGRSE
P01371	EHDPSAPGNGYC
P01370	EGGGNRGDPSGVC
P34070	DSGSSRDPGASSGGC
Q5Y4V3	MRAIISLLLISTMVFGVIEAVSLEEGLKIFEGERGDCVGESQQCADWSGPYCCKGYYCTCRYFPKCICV NDNGK
G2TRQ6	MEREGKEDVRKADHKIVYGIERVRHGINNFFDDVGKAVKSESDTADSKRSAEAKADEAPAKM
P17726	YNRLCIKPRDWIDECDSNEGGERAYFRNGKGGCDSFWICPEDHTGADYYSSYRDCFNACI
O31467	MFSNIGIPGLILIFVIALIIFGPSKLPEIGRAAGRTLLEFKSATKSLVSGDEKEEKSAELTAVKQDKNAG
D4GWC8	MFETITPLFPGLPGGPELLVVLLIVVLLFGANKIPKLARSSGQAIGEFQRGREEIEDELQDMTGDDDED DATSESSADSVSTDSVSTESSN
P69428	MGGISIWQLLIIAVIVVLLFGTKKLGSIGSDLGASIKGFKKAMSDDEPKQDKTSQDADFTAKTIADKQA DTNQEQAKTEDAKRHDKEQV
P04613	MDPVDPNLEPWNHPGSQPRTPCNKCYCKKCCYHCQMCFITKGLGISYGRKKRRQRRRPPQGNQAHQ DPLPEQPSSQHRGDHPTGPKE
Q5EPH2	MAATLPVFAVVFFAMVLASSQANECVSKGFGCLPQSDCPQEARLSYGGCSTVCCDLSKLTGCKGKGG ECNPLDRQCKELQAESASCGKGQKCCVWLH
P0CI08	ARCEQCPSYCCQSDSPPECDGCE
P0CI14	MMTKTGLVLLFAFLLVFPVSSLPMDAEAGHARLEMDKRDAGNEAWTRLLKRYEENCGTEYCTSKIG CPGRCVCKEYNYNGEITRRCRA
P0CI11	NEVCPPGECQQYCCDLRKCKCINLSFYGLTCNCDS
P0CI06	GECCTDCAQTAAANYC
P56587	ALCNCNRIIIPHMCWKKCGKK
P04445	MNKTELIKAIAQDTELTQVSVSKMLASFEKITTETVAKGDKVQLTGFLNIKPVARQARKGFNPQTQEA LEIAPSVGVSVKPGESLKKAAEGLKYEDFAK
Q6ZYL4	MVNVLKGVLIECDPAMKQFLLYLDESNALGKKFIIQDIDDTHVFVIAELVNVLQERVGELMDQNAFSL TQK
Q3E7C1	MARARKGALVQCDPSIKALILQIDAKMSDIVLEELDDTHLLVNPSKVEFVKHELNRLLSKNIYNPMDEE ENQ
P04155	MATMENKVICALVLVSMLALGTLAEAQTETCTVAPRERQNCGFPGVTPSQCANKGCCFDDTVRGVP WCFYPNTIDVPPEEECEF
Q62395	METRALWLMLLVVLVAGSSGIAADYVGLSPSQCMVPANVRVDCGYPSVTSEQCNNRGCCFDSSIPNV PWCFKPLQETECTF
P04289	MGLSFSGARPCCCRNNVLITDDGEVVSLTAHDFDVVDIESEEEGNFYVPPDMRGVTRAPGRQRLRSSD PPSRHTHRRTPGGACPATQFPPPMSDSE
Q57755	MSKVKIELFTSPMCPHCPAAKRVVEEVANEMPDAVEVEYINVMENPQKAMEYGIMAVPTIVINGDVE FIGAPTKEALVEAIKKRL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
O31617	MLQLNGKDVKWKKDTGTIQDLLASYQLENKIVIVERNKEIIGKERYHEVELCDRDVIEIVHFVGGG
O32583	MQILFNDQAMQCAAGQTVHELLEQLDQRQAGAALAINQQIVPREQWAQHIVQDGDQILLFQVIAGG
P01249	GQFLEDPSVLTKEKLKSELVANNVTLPAGEQRKDVYVELYLQHLTALKR
P81946	LICYNQLGTKPPTTETCGDDSCYKMIWTYDGVIRRGCGCFTPRGDMPRPRCCKSDKCNL
P62072	MDPLRAQQLAAELEVEMMADMYNRMTSACHRKCVPPHYKEAELSKGESVCLDRCVSKYLDIHERMG KKLTELSMQDEELMKRVQQSSGPA
P87108	MSFLGFGGGQPQLSSQQKIQAAEAELDLVTDMFNKLVNNCYKKCINTSYSEGELNKNESSCLDRCVAK YFETNVQVGENMQKMGQSFNAAGKF
Q9XH48	MDSYSSPPMGGSGSSVSPEVMMESVKTQLAQAYAEELIETLRTKCFDKCVTKPGSSLGGSESSCISRCV ERYMEATAIISRSLFTQR
Q9Y5L4	MEGGFGSDFGGSGSGKLDPGLIMEQVKVQIAVANAQELLQRMTDKCFRKCIGKPGGSLDNSEQKCIA MCMDRYMDAWNTVSRAYNSRLQRERANM
Q7SBR3	MSDSTSETVKKAIIKQVLIESQSANARTLMEKIGENCFTSCVPKPGSSLSNSEKTCVTQCTEKYMAAWN VVNTTYLRRIQQEMGNQ
O60220	MDSSSSSSAAGLGAVDPQLQHFIEVETQKQRFQQLVHQMTELCWEKCMDKPGPKLDSRAEACFVNCV ERFIDTSQFILNRLEQTQKSKPVFSESLSD
Q9Y8C0	MDIPQADLDLLNEKDKNELRGFISNETQRQRVQGQTHALTDSCWKKCVTSPIKTNQLDKTEAVCMAD CVERFLDVNLTIMAHVQKITRGGSK
P57744	MSSLSTSDLASLDDTSKKEIATFLEGENSKQKVQMSIHQFTNICFKKCVESVNDSNLSSQEEQCLSNCVN RFLDTNIRIVNGLQNTR
Q9XGX9	MDASMMAGLDGLPEEDKAKMASMIDQLQLRDSLRMYNSLVERCFVDCVDSFTRKSLQKQEETCVMR CAEKFLKHTMRVGMRFAELNQNAPTQD
Q9Y5J7	MAAQIPESDQIKQFKEFLGTYNKLTETCFLDCVKDFTTREVKPEETTCSEHCLQKYLKMTQRISMRFQ EYHIQQNEALAAKAGLLGQPR
Q8J1Z1	MDGLTAAESRELDQRLQKRQVKEFMSVFGNLVDNCFTACVDDFTSKALSGRESGCISRCVLKSMSTQ TRLGERFGELNAAMTAEMQRR
P0C171	METRQVSRSPRVRLLLLLLLLLVVPWGVRTASGVALPPVGVLSLRPPGRAWADPATPRPRRSLALAD DAAFRERARLLAALERRHWLNSYMHKLLVLDAP
P41518	GLGNNAFVGVR
Q8I6S2	MIRVGLILCCIFIVGVFEASSADDILTAHNLIKRSEVKPPSSSEFVGLMGRSEELTRRLIQHPGSMSETSK RGPPKKGDFNPNELKPESNIC
P08609	SPSNSKCPDGPDCFVGLM
P42634	MNMFITVQIVIVLVLAVLSEAASLPTATERKDAMDEGPNQSDEPEGSVADPSTKDDDYSDSLKQDEKY YKVRLLNTGDKFYGLMG
P85079	MKLSIFFVLFFIAIAYCQPEFLDDEEDEVEETLPVAEEGREKSCITWRNSCMHNDKGCCFPWSCVCWS QTVSRNSSRKEKKCQCRLW
P86308	QKNDKKDRFYGLM
P85879	QPPTPPEHRFPGLM
P85880	ASEPTALGLPRIFPGLM
Q45060	MTKNQNQYQQPNPDDRSDNVEKLQDMVQNTIENIEEAEASMEFASGEDKQRIKEKNARREQSIEAFR NEIQDESAARQNGYRS
Q9Y6G1	MDLIGFGYAALVTFGSIFGYKRRGGVPSLIAGLFVGCLAGYGAYRVSNDKRDVKVSLFTAFFLATIMG VRFKRSKKIMPAGLVAGLSLMMILRLVLLLL
Q06177	MTRTSKWTVHEAKSNPKYFTHNGNFGESPNHVKRGGYGKGNWGKPGDEINDLIDSGEIKTVFNKTR RGSNSQNNERRLSDLQQYHI
Q9BN12	RCHFVVCTTDCRRNSPGTYGECVKKEKGKECVCKS
Q9FNC9	MAAKRIGAGKSGGGDPNILARISNSEIVSQGRRAAGDAVEVSKKLLRSTGKAAWIAGTTFLILVVPLIIE MDREAQINEIELQQASLLGAPPSPMQRGL
P33034	SCIDIGGDCDGEKDDCQCCRRNGYCSCYSLFGYLKSGCKCVVGTSAEFQGICRRKARQCYNSDPDKC ESHNKPKRR
P37045	MKLCMTLLITAIAVVTFVVATQEESAEFNEVEESREDNCIAEDYGKCTWGGTKCCRGRPCRCSMIGT NCECTPRLIMEGLSFA
Q9SD80	MVNNVVSIEKMKALWHSEVHDEQKWAVNMKLLRALGMFAGGVVLMRSYGDLMGV
Q8N4H5	MFRIEGLAPKLDPEEMKRKMREDVISSIRNFLIYVALLRVTPFILKKLDSI
P80967	MFGLPQQEVSEEEKRAHQEQTEKTLKQAAYVAAFLWVSPMIWHLVKKQWK
Q9XIA7	MFPGMFMRKPDKAEALKQLRTHVALFGSWVVIIRAAPYVLSYFSDSKDELKIDF
Q96B49	MASSTVPVSAAGSANETPEIPDNVGDWLRGVYRFATDRNDFRRNLILNLGLFAAGVWLARNLSDIDL MAPQPGV

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P33448	MDGMFAMPGAAAGAASPQQPKSRFQAFKESPLYTIALNGAFFVAGVAFIQSPLMDMLAPQL
Q9ASY8	MESTISLKVNKGKGKGSKGASSSDDKSKFDVVKEWTNWSLKKAKVVTHYGFIPLVIFVGMNSDPKPH LFQLLSPV
Q9P0U1	MVKLSKEAKQRLQQLFKGSQFAIRWGFIPLVIYLGFKRGADPGMPEPTVLSLLWG
P53507	MSFLPSFILSDESKERISKILTLTHNVAHYGWIPFVLYLGWAHTSNRPNFLNLLSPLPSV
Q2EES9	MQHELQPDSLVDLKFIMADTGFGKTFIYDRIKSGDLPKAKVIHGRARWLYRDHCEFKNKLLSRANG
P83580	MNTATGFIVLLVLATVLGAIEAEDAVPDFEGGFASHAREDTVGGKIRRSSVCIPSGQPCPYNEHCCSGS CTYKENENGNTVQRCD
P0C8M0	DWECLPLHSSCDNDCVCCKNHHCHCPYSNVSKLEKWLPEWAKIPDALKRCSCQRNDKDGKINTCDK YKN
P18033	MISKRRFSLPRLDITGMWVFSLGVWFHIVARLVYSKPWMAFFLAELIAAILVLFGAYQVLDAWIARVS REEREALEARQQAMMEGQQEGGHVSH
P81281	TSQPDCSVGCDTSYCPDTSSCNCGTFADYCKCCQYCNACAGKTCNMIAGQSCEDGYLCRPPEGYSYI DVVTGRISSLCLRI
O43715	MNSVGEACTDMKREYDQCFNRWFAEKFLKGDSSGDPCTDLFKRYQQCVQKAIKEKEIPIEGLEFMGH GKEKPENSS
P81735	NGERAPGSKKAPSGFLGTR
P85806	APACVGFQGMR
P86593	APSSMGFMGMR
P85807	GPSSSAFFGMR
P85808	SPATMGFAGVR
P85809	QERRAMGFVGMR
P81753	SGLDSLSGATFGGNR
Q3E790	MTQHKSSMVYIPTTKEAKRRNGKSEGILNTIEEVVEKLYWTYYIHLPFYLMASFDSFFLHVFFLTIFSLS FFGILKYCFL
O97373	MKFFTVLFFLLSIIYLIVAAPGEPGAPIDYDEYGDSSEEVGGTPLHEIPGIRL
P02944	RLLDDTPEVKVLGAVADAIETPKAEPCIDLDVAGEATFAREDDLPDYVLYAEVTFHEICRDGGSESEG KNGSQMRLIADVGPESATVAK
P0A892	MTDLFSSPDHTLDALGLRCPEPVMMVRKTVRNMQPGETLLIIADDPATTRDIPGFCTFMEHELVAKE TDGLPYRYLIRKGG
P45530	MLHTLHRSPWLTDFAALLRLLSEGDELLLLQDGVTAAVDGNRYLESLRNAPIKVYALNEDLIARGLTG QISNDIILIDYTDFVRLTVKHPSQMAW
C5J895	MKIACTLVLFVMLRCYVNARNIPGTCRTHTGIILLSGEEWKDPNHCSTYRCSIFDGEAELEGVTCAAY HVPPHCRLVSAANELYPQCCPTVICSDKSR
P60979	AECLMIGDTSCVPRLGRRCCYGAWCYCDQQLSCRRVGRKRECGWVEVNCKCGWSWSQRIDDWRAD YSCKCPEDQ
P83997	QWIPGQSCTNADCGEGQCCTGGSYNRHCQSLSDDGKPCQRPNKYDEYKFGCPCKEGLMCQVINYCQ KK
P84062	AELKSCFPVGHECDDDASNCNCCGDDVYCACGWGRWNCKCKVADQSYAYGICKDKVNCPNRHLWP AKECKMPCRRNCG
Q4QQV1	MSYIPGQPVTAVVQRVEIHKLRQGENLILGFSIGGGIDQDPSQNPFSEDKTDKVNGWDMTMVTHDQA RKRLTKRSEEVVRLLVTRQSLQKAVQQSMLS
P83905	CGDINAPCQSDCDCCGYSVTCDCYWSKDCKCRESNFVIGMALRKAFCKNK
P86465	NIVDVPCRDDYYRDSSGNGVYDQLGGCGAA
P29423	MKVAILILSILVLAVASETIEEYRDDFAVEELERATCAGQDKPCKETCDCCGERGECVCALSYEGKYR CICRQGNFLIAWHKLASCKK
P0DJ96	KSPPQALNKPLPAPSAPSL
P83909	ECADVYKECWYPEKPCCKDRACQCSLGMNCKCKATLGDIF
P29426	SFCIPFKPCKSDENCCKKFKCKTTGIVKLCRW
P60993	ECRYFWGECNDEMVCCEHLVCKEKWPITYKICVWDRTF
O76200	MWFKIQVLVLAITLITLGIQAEPNSSPNNPLIVEEDRAECAAVYERCGKGYKRCCEERPCKCNIVMDNC TCKKFISELFGFGK
P81791	MKLCILLVVLLITVVRAEEDILENEAEDISPAIKERSARGCIGRNESCKFDRHGCCWPWSCSCWNKEGQ PESDVWCECSLKIGK
P81792	MKCAVLFLSVIALVHIFVVEAEEEPDSDALVPQERACIPRGEICTDDCECCGCDNQCYCPPGSSLGIFK CSCAHANKYFCNRKKEKCKKA

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q6B4T3	MTFKLFVVVTLVLAIYVATAEEAMKDDSEPAERGCIKYGDRCGSPHGLPSNCCNDWKYKGRCGCTM GVCTCGPNCPSRGCDWSKKG
P25683	RICYTHKSLQAKTTKSCEGNTCYKMFIRTSREYISERGCGCPTAMWPYQTECCKGDRCNK
P01535	RSCCPCYWGGCPWGQNCYPEGCSGPKV
P09949	AGGKSTCCPCAMCKYTAGCPWGQCAHHCGCS
Q4U4N3	MLKLICIAFLVTVLTLVAGQDSLDPAEFGCADDVNQAELLKNNDICLQCEDLHKEGVVFSLCKTNCFT TEYFQHCVKDLEEAKKEPPE
P69930	MNKVLFLCLVVLCATSAFAAEEEYVERAPVKRALLSCRCEGKTEYGDKWLFHGGCPNNYGYNYKCF MKPGAVCCYPQNGR
P49125	MSKLFFVVFLCLIISVFAISPADIGCTDISQADFDEKNNNCIKCGEDGFGEEMVNRCRDKCFTDNFYQS CVDLLNKVYEEKDTPPVQE
B3FIV1	MNTIQVIIFAVVLVLTVTVGQADEDSAETSLLRKLEEAEAAMFGQYLEESKNSREKRCIGESVPCDKDD PRCCREYECLKPTGYGWWYASYYCYRKKSG
P81694	SCIPKHEECTNDKHNCCRKGLFKLKCQCSTFDDESGQPTERCACGRPMGHQAIETGLNIFRGLFKGK KKNKKTK
P81658	WQPPWYCKEPVRIGSCKKQFSSFYFKWTAKKCLPFLFSGCGGNANRFQTIGECRKKCLGK
P14531	MKTQVLALFVLCVLFCLAESRTTLNKRNDIEKRIECKCEGDAPDLSHMTGTVYFSCKGGDGSWSKCN TYTAVADCCHQA
P0C1Y9	TMCYSHTTTSRAILTNCGENSCYRKSRRHPPKMVLGRGCGCPPGDDYLEVKCCTSPDKCNY
P68424	MNMKILVLVAVLCLVVSTHAERHSKTDMEDMEDSPMIQERKCLPPGKPCYGATQKIPCCGVCSHNKC T
P83303	MVNMKASMFLALAGLVLLFVVCYASESEEKEFSNELLSSVLAVDDNSKGEERECLEIFKACNPSNDQC CKSSKLVCSRKTRWCKYQIGK
P68425	MGIARILSAVLFLSVLFVVTFPALLSADHHDGRIDTCRLPSDRGRCKASFERWYFNGRTCAKFIYGGCG GNGNKFPTQEACMKRCAKA
P58425	DCGTIWHYCGTDQSECCEGWKCSRQLCKYVIDW
P58427	ECGTLFSGCSTHADCCEGFICKLWCRYERTW
O46166	MKLQLMICLVLLPCFFCEPDEICRARMTHKEFNYKSNVCNGCGDQVAACEAECFRNDVYTACHEAQK G
P49126	MKVFVVLLCLSLAAVYALEERLDKDADIMLDSPADMERAKDGDVEGPAGCKKYDVECDSGECCQKQ YLWYKWRPLDCRCLKSGFFSSKCVCRDV
Q0EAE5	MKGQMIICLVLIALCMSVVVMAQNLRAEELEKANPKDERVRSFERNQKRACKDYLPKSECTQFRCRT SMKYKYTNCKKTCGTC
P86399	MSTFIVVFLLLTAILCHAEHAIDETARGCNRLNKKCNSDADCCRYGERCISTGVNYYCRPDFGP
Q75WH3	MKVFSFTVVVVMILSLSAFVLAGDEGDVMKKIVAMEEAVEERACLAEYQKCEGSTVPCCPGLSCSAG RFRKTKLCTK
Q75WH2	MKAPATIVILIMSLISVLWATADTEDGNLLFPIEDFIRKFDEYPVQPKERSCKLTFWRCKKDKECCGW NICTGLCIPPGKK
Q75WG6	MKLTLFILIVFVVLANVYAAGISERNIIGGRVIKLCGGGAQKCCDREPRCDPCRKCVQSFHSGVYMCSD KKSNCS
Q75WG5	MKILEKALLENDSAAEEESRNLRTKRCARKRAWCEKTENCCCPMKCIYAWYNGQSSCDHTISTIWTS CPK
P0C2V1	ECSKQLGESCKCNKQCCGATVICGTIYVGGKEENLCIEKTSNNAILNFFGKIAHVVENGLSFSCD
P81030	LTCVTSKSIFGITTENCPDGQNLCFKKWYYIVPRYSDITWGCAATCPKPTNVRETIRCCETDKCNE
P0CI24	CCRLACGLGCHPCC
P83562	MRTIVFLIVSILLLSSAVLMLAEGNAASHELQEYPIEESLEEQRKCVDGSCDPYSSDAPRCCGSQICQCIF FVPCYCKYRGK
Q75WH5	MKYVMVILGLLVLAAICCASEVEELDWRQEIARAILEAELQPEERDCKGLFRQCKKSSECCKGSSCESD LTGLCIFNLPGR
Q75WH4	MKFATFAFTLCVVISLSVLVLADEEEKDFLMNLIQPLKESEERQFCGTNGKPCVNGQCCGALRCVVTY HYADGVCLKMNP
P82464	TICYNHLTRTSETTEICPDSWYFCYKISLADGNDVRIKRGCTFTCPELRPTGIYVYCCRRDKCNQ
P36990	GCKGFLVKCDSNSECCKTAIVKGKKKQLSCLCGAWGAGCSCSFRCGNRC
P49269	EIPQNLGSGIPHDRIKLPNGQWCKTPGDLCSSSSECCKAKHSNSVTYASFCSREWSGQQGLFINQCRTC NVESSMC
P36983	MKHLIFSSALVCALVVCTFAEEQVNVPFLPDERAVKCIGWQETCNGNLPCCNECVMCECNIMGQNCR CNHPKATNECESRRR
P49266	SCVDFQTKCKKDSDCCGKLECSSRWKWCVYPSPF

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P36989	CAKHSETCKNGNCCTCTQYRGKDEPMACRRGTHGQRCQCVMKIMKH
P0DJ08	MNFATKVSLLLLAIAVIVIVEGGEGDSWFEEHEESDTERDFPLSKEYESCVRPRKCKPPLKCNKAQICV DPNKGW
P25680	LRCLNCPEVFCRNFHTCRNGEKICFKRFDQRKLLGKRYTRGCAVTCPVAKPREIVECCSTDGCNR
P25678	LECYQMSKVVTCKPEETFCYSDVFMPFRNHIVYTSGCSSYCRDGTGEKCCTTDRCNGARGG
P04096	QEKPYWPPPIYPM
P69968	MAYDLSEFMGDIVALVDKRWAGIHDIEHLANAFSLPTPEIKVRFYQDLKRMFRLFPLGVFSDEEQRQN LLQMCQNAIDMAIESEEEELSELD
O97420	MSPKNNHDPSSSGDSGNTNVQEADLQEMEDVNNSLDALSCALDAVEQRTDDIMSQLRELLNSNREIRR LIAEENDNAPESGDDNMDGQAGSEAAPK
Q9W3T5	MSPYSGSVRRLLDSWPGKKRFGVYRFLPLFFLLGAGLEFSMINWTVGETNFYRTFKRRQAKNYVEEQ QHLQARAANNTN
O97172	MVCVPCIIIPLLLYIWHKFVQPILLRYWNPWEKKDDDGNVIKKGPDFPFECKGGVCPFVPGGKKTEN VSDDDAEESENPPLNATAMAAETEVDESKKEI
P48505	MTSPAAAGNGLFKFLRPKLRPQSTDIQAAAGWGVAAVTGALWVIQPWDFLRKTFIEKQEEEK
P43266	MQTHVRRVALQALRPCLRAGLMAPKFPVRFATTAVSGELLTKTPYTRPGYAAQWTCLVVLFLKNQL LMRLFFAFVAYVVAMKVFGARFHVDHDEDATPAE
P34661	MSGGTAATTAGSKVTFKITLTSDPKLPFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNDGVGVNP AQPAGNIFLKHGSELRLIPRDRVGH
Q9EXD8	MENKNLHIIAACGNGMGTSMLIKIKVEKIMKELGYTAKVEALSMGQTKGMEHSADIIISSIHLTSEFNPN AKAKIVGVLNLMDENEIKQALSKVL
P14739	MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWAL VIQDSNGENKIKML
P86968	MKYVALAFVLSLVILQISAQGGGLTSLLLQKEYMPDSWFDYKLAQMILGGPTGRKSRTQSGRNQRKSN SDSWLWLALAN
P82036	FQFVNPSDIVFGS
P82037	GLAGAISSALDKLKQSQLIKNYAKKLGYPR
Q2YPD7	MNLTPREKDKLLIAMAAMVARRRLERGVKLNHPEAIALVSDFVVEGARDGRTVAELMEAGAHVITRE QVMDGVAEMIRDIQVEATFPDGTKLVTVHEPIR
P40554	MVNVKVEFLGGLDAIFGKQRVHKIKMDKEDPVTVGDLIDHIVSTMINNPNDVSIFIEDDSIRPGIITLIND TDWELEGEKDYILEDGDIISFTSTLHGG
P06481	MTSRLSDPNSSARSDMSVPLYPTASPVSVEAYYSESEDEAANDFLVRMGRQQSVLRRRRRRTRCVGM VIACLLVAVLSGGFGALLMWLLR
Q3ZBI7	MAGPEADAQFHFTGIKKYFNSYTLTGRMNCVLATYGSIALIVLYFKLRSKKTPAVKAT
H2A0P3	MQCVCLCVFVLLLAGCVTSQEEVEVDCYCVYFKYDCPEDYIESNRYNNQCQIRHKCCVPPPKYFFFV FPKGYRMP
H2A0M9	MKNAVLVFLLLSVFALSVNAYGYGCGWYGCPYGSKCICPYYGKCYCVPVYGKNCYIYGCPYPKVCV YGVCKWRYKY
H2A0N0	MYDTWFVLTAVVLFVLVLIGNVHGYPPWAWSGFMGSKWYNSWGPWAWRGWDDDWFDD
H2A0N6	MKYVALAFVLSLVILQISAQVGAAYIPGMGSIGSVGRAGAAAGASAGIGNQGRGAGLLRFFTLILENLM KNNQQAQPKQDNFGAQLQNLLKKKMILEMIN
H2A0N8	MQKLTIFVIFLCCWLVLTMADDYVKKCQHHCDLQLKRCARECHFPFHPYHKECLKICEKTKLHCFRW CAHA
Q06318	MKIAITITVVMLSICCSSASSDICPGFLQVLEALLMESESGYVASLKPFNPGSDLQNAGTQLKRLVDTLP QETRINIMKLTEKILTSPLCKQDLRF
P01145	NDDPPISIDLTFHLLRNMIEMARIENEREQAGLNRKYLDEV
P04558	GSGADCFWKYCV
P40388	MSAQQFYGDKGYAPAPPQQAYGGPNYYPPQQNYPQQGYAPPQGYPQGGYPAQQPMYVQQPQASD PGGDLCCGLLTGLACCCCLDAMF
P81103	MAYTGLTVPLIVMSVFWGIVGFLVPWFIPKGPNRGVIITMLVTCSVCCYLFWLIAILAQLNPLFGPQLK NETIWYLKYHWP
Q3E7B6	MSSFYTVVGVFIVVSAMSVLFWIMAPKNNQAVWRSTVILTLAMMFLMWAITFLCQLHPLVAPRRSDL RPEFAE
Q15836	MSTGPTAATGSNRRLQQTQNQVDEVVDIMRVNVDKVLERDQKLSELDDRADALQAGASQFETSAAK LKRKYWWKNCKMWAIGITVLVIFIIIIVWVVSS
Q9BV40	MEEASEGGGNDRVRNLQSEVEGVKNIMTQNVERILARGENLEHLRNKTEDLEATSEHFKTTSQKVAR KFWWKNVKMIVLICVIVFIIILFIVLFATGAFS

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
E4QWH3	MLTKVFQSGNSQAVRIPMDFRFDVDTVEIFRKENGDVVLRPVSKKTDDFLALFEGFDETFIQALEARD DLPPQERENL
O07227	MSDVLIRDIPDDVLASLDAIAARLGLSRTEYIRRRLAQDAQTARVTVTAADLRRLRGAVAGLGDPELM RQAWR
O06416	MLSRRTKTIVVCTLVCMARLNVYVPDELAERARARGLNVSALTQAAISAELENSATDAWLEGLEPRST GARHDDVLGAIDAARDEFEA
Q8ZZP2	MSEVISIRVRRGLKKELEELGINYAEAVRKFLEELVARERRRRALERARALREELRKKGAFPPSAELIR EDRDEASR
O07782	MSATIPARDLRNHTAEVLRRVAAGEEIEVLKDNRPVARIVPLKRRRQWLPAAEVIGELVRLGPDTTNL GEELRETLTQTTDDVRW
P96916	MSEVASRELRNDTAGVLRRVRAGEDVTITVSGRPVAVLTPVRPRRRRWLSKTEFLSRLRGAQADPGL RNDLAVLAGDTTEDLGPIR
Q7CPV2	MHTTLFFSNRTQAVRLPKSISFPEDVKHVEIIAVGRSRIITPVGESWDSWFDGEGASTDFMSTREQPAV QEREGF
O05728	MYALAFDLKIEILKKEYGEPYNKAYDDLRQELELLGFEWTQGSVYVNYSKENTLAQVYKAINKLSQIE WFKKSVRDIRAFKVEDFSDFTEIVKS
P85800	SDQGDVAEPKMHKTAPPFDFEAIPEEYLDDES
P84843	MKFALFSVLVVLLIATFVAADECPRICTADYRPVCGTPSGGRRSANRTFGNQCSLNAHNCLNKGDTY DKLHDGECK
E6Z0R4	MLSEEEIEYRRRDARNALASQRLEGLEPDPQVVAQMERVVVGELETSDVIKDLMERIKREEI
P21680	MSRDELRIVLGAMIPNMEEGFEIKTRDGAILRVDPEWECCKEFKDGLKAEIIKQLKSKPAVVFGYS
P06922	YYVLHLCLAATKYPLLKLLGSTWPTTPPRPIPKPSPWAPKKHRRLSSDQDQSQTPETPATPLSCCTET QWTVLQSSLHLTAHTKDGLTVIVTLHP
P0CK45	MPNLWFLLFLGLVAAMQLLLLLFLLLFFLVYWDHFECSCTGLPF
P06927	MTNLDTASTTLLACFLLCFCVLLCVCLLIRPLLLSVSTYTSLIILVLLLWITAASAFRCFIVYIIFVYIPLFLI HTHARFLIT
P04020	MHGRLVTLKDIVLDLQPPDPVGLHCYEQLEDSSEDEVDKVDKQDAQPLTQHYQILTCCCGCDSNVRL VVECTDGDIRQLQDLLLGTLNIVCPICAPKP
P06465	MVGEMPALKDLVLQLEPSVLDLDLYCYEEVPPDDIEEELVSPQQPYAVVASCAYCEKLVRLTVLADHS AIRQLEELLLRSLNIVCPLCTLQRQ
P59637	MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEG VPDLLV
P0C2R0	MFNLFLTDTVWYVGQIIFIFAVCLMVTIIVVAFLASIKLCIQLCGLCNTLVLSPSIYLYDRSKQLYKYYNE EMRLPLLEVDDI
Q65212	MGGRRRKKRTNDTKHVRFAAAVEVWEADDIERKGPWEQVAVDRFRFQRRIASVEELLSTVLLRQKK LLEQQ
P15236	MLKSEPSFASLLVKQSPGMHYGHGWIAGKDGKRWHPCRSQSELLKGLKTKSPKSSGFLIIRIVHFVIKG VKHVTR
P03679	MGKIFDQEKRLEGTWKNSKWGNQGIIAPVDGDLKMIDLELEKKMTKLEHENKLMKNALYELSRMEN NDYATWVIKVLFGGAPHGAK
P04532	MSEQTVEQKLSAEIVTLKSRILDTQDQAARLMEESKILQGTLAEIARAVGITGDTIKVEEIVEAVKNLTA ESADEAKDEE
P13848	MPLKPEEHEDILNKLLDPELAQSERTEALQQLRVNYGSFVSEYNDLTKSHEKLAAEKDDLIVSNSKLFR QIGLTDKQEEDHKKADISETITIEDLEAK
P03654	MKPKTTLLLQELLLLTYELNRSGLLVENEEIQSQLKKLEVVLLCNLSPSSQRAGKN
P68660	MTRQEELAAARAALHDLMTGKRVATVQKDGRRVEFTATSVSDLKKYIAELEVQTGMTQRRRGPAGF YV
P83627	YNIPLGWGRRDMPGCLGVLGNRDLYDDVSRICSDCQNVFRDKNVESKCRSDCFSTSYFETCIMALDLA EKISDYKLHASILKE
P0A3W4	MKYCLLCLVVALSGCQTNDTIASCKGPIFPLNVGRWQPTPSDLQLRNSGGRYDGA
P08063	MVIIKLNANKNMPVLAVEKPQEIHKEELSDHHQSNGFTSLDLEMIELENFVLHCPLPEENLAG
P18378	MLAFCYSLPNAGDVIKGRVYEKDYALYIYLFDYPHFEAILAESVKMHMDRYVEYRDKLVGKTVKVKV IRVDYTKGYIDVNYKRMCRHQ
P41806	MAVDVPRAVINKLMLFTAAMVVLPVLTFFIIQQFTPNTLISGGLAAAMANVVLIVYIVVAFREDTEDH KVDGNKKED
Q8WQK0	MSKIILAIFLIVLCGLIFVTVDAMIDAPCKDNDDCDRFTEYCAIYADENGNEAGKRCEDAIGLLV
P06817	MNNATFNCTNINPITHIRGSIIITICVSLIVILIVFGCIAKIFINKNNCTNNVIRVHKRIKCPDCEPFCNKRD DISTPRAGVDIPSFILPGLNLSEGTPN
P83230	SDSKIGNGCFGFPLDRIGSVSGLGCNRIMQNPPKKFSGE

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P28374	EVKYDPCFGHKIDRINHVSNLGCPSLRDPRPNAPSTSA
P80909	MKKAYPVINRVECKACERCIIACPRKVLYMSNKINERGYHYVEYRGEGCNGCGNCYYTCPEINAIEVH IERCEDGDTDG
P27383	MMGKGFEMMVASAIRAAGINPDELMEKANTLVHNLNYQLDRFGQRLDSIDSRLSVIEKALDISPAEKP DNQPELTGITFEGDNNDQ
P86825	KHFGKDSNFPFGT
P27388	MQLITDMAEWSSKPFRPDMSLTGWLAFVGLIIVAIILWQQIIRFIIE
P83717	AVSCFKACMKKRGIPFVCSVDC
P20224	MKWVQKAIKRPGRVHRYLMRLYGKRAFTKDGDIKASYLDKAIKHVKKAKIPKEKKRSLLSALLLAKR LKRMHRK
P27391	MALINPQFPYAGPVPIPGPAPTETMPLLNYRVEGRIAGIQQARQFMPFLQGPHRAVAEQTYHAIGTGI QMGQTFNQPLINTQEG
P27390	MNDFVGPIVTVLTAIIGVAILAVLVSRNSNTAGVIKAGSGGFSSMLGTALSPVTGGTGFAMTNNYSGF
P20225	MEISLKPIIFLVVFIIVGIALFGPINSVVNNVTTSGTYTTIVSGTVTTSSFVSNPQYVGSNNATIVALVPLF YILVLIIVPAVVAYKLYKEE
O70791	MSSQQETNDKSNTQGHPETDPEGKTGTDTGNTEDSPPDTDNVPITDDAIMDDVMDEDVKEEDIDYS WIEDMRDEDVDAEWLFELIDECNGWPD
Q10848	MYSGVVSRTNIEIDDELVAAAQRMYRLDSKRSAVDLALRRLVGEPLGRDEALALQGSGFDFSNDEIESF SDTDRKLADES
P95006	MRTQVTLGKEELELLDRAAKASGASRSELIRRAIHRAYGTGSKQERLAALDHSRGSWRGRDFTGTEY VDAIRGDLNERLARLGLA
P95003	MLVAYICHVKRLQIYIDEDVDRALAVEARRRRTSKAALIREYVAEHLRQPGPDPVDAFVGSFVGEADL SASVDDVVYGKHE
O28071	MPKIIEAVYENGVFKPLQKVDLKEGERVKIKLELKVEPIDLGEPVSVEEIKKIRDGTWMSS
O33300	MHRGYALVVCSPGVTRTMIDIDDDLLARAAKELGTTTKKDTVHAALRAALRASAARSLMNRMAENA TGTQDEALVNAMWRDGHPENTA
P71622	MTATEVKAKILSLLDEVAQGEEIEITKHGRTVARLVAATGPHALKGRFSGVAMAAADDDELFTTGVS WNVS
P0CW33	MRTTVSISDELLATAKRRARERGQSLGAVIEDALRRELAAARTGGARPTVPVFDAGTGPRPGIDLTSN TVLSEVLDEGLELNSRK
O53778	MDKTTVYLPDELKAAVKRAARQRGVSEAQVIRESIRAAVGGAKPPPRGGLYAGSEPIARRVDELLAG FGER
O07779	MKAVVDAAGRIVVPKPLREALGLQPGSTVEISRYGAGLHLIPTGRTARLEEENGVLVATGETTIDDEV VFGLIDSGRK
O07770	MALNIKDPSVHQAVKQIAKITGESQARAVATAVNERLARLRSDDLAARLLAIGHKTASRMSPEAKRLD HDALLYDERGLPA
P96913	MALSIKHPEADRLARALAARTGETLTEAVVTALRERLARETGRARVVPLRDELAAIRHRCAALPVVD NRSAEAILGYDERGLPA
O06565	MRTTVTVDDALLAKAAELTGVKEKSTLLREGLQTLVRVESARRLAALGGTDPQATAAPRRRTSPR
O50456	MRTTLTLDDDVVRLVEDAVHRERRPMKQVINDALRRALAPPVKRQEQYRLEPHESAVRSGLDLAGFN KLADELEDEALLDATRRAR
P0CW28	MNEVSIRTLNQETSKVLARVKRGEEINLTERGKVIARIIPASAGPLDSLISTGSVQPARVHGPAPRPTIP MRGGLDSGTLLERMRAEERY
Q8VJG5	MRTTLQIDDDVLEDARSIARSEGKSVGAVISELARRSLRPVGIVEVDGFPVFDVPPDAPTVTSEDVVRA LEDDV
P65027	MRTTIDVAGRLVIPKRIRERLGLRGNDQVEITERDGRIEIEPAPTGVELVREGSVLVARPERPLPPLTD EIVRETLDRTRR
Q8VJF3	MKTTLDLPDELMRAIKVRAAQQGRKMKDVVTELLRSGLSQTHSGAPIPTPRRVQLPLVHCGGAATRE QEMTPERVAAALLDQEAQWWSGHDDAAL
P0A5G9	MRTTIRIDDELYREVKAKAARSGRTVAAVLEDAVRRGLNPPKPQAAGRYRVQPSGKGGLRPGVDLSS NAALAEAMNDGVSVDAVR
P65077	MRATVGLVEAIGIRELRQHASRYLARVEAGEELGVTNKGRLVARLIPVQAAERSREALIESGVLIPARR PQNLLDVTAEPARGRKRTLSDVLNEMRDEQ
P86166	NVCDGDACPDGVCRSGCTCDFNVAQRKDTCFYPQ
P35967	MEQAPEDQGPQREPHNEWTLELLEELKREAVRHFPRPWLHGLGQHIYETYGDTWAGVEAIIRILQQL LFIHFRIGCQHSRIGIIQQRRARRNGASRS
P69699	MQPIQIAIVALVVAIIIAIVVWSIVIIEYRKILRQRKIDRLIDRLIERAEDSGNESEGEISALVEMGVEMGHH APWDVDDL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P17309	MIKQLQHALELQRNAWNNGHENYGASIDVEAEALEILRYFKHLNPAQTALAAELQEKDELKYAKPLAS AARKAVRHFVVTLK
P03704	MSNVNTGSLSVDNKKFWATVESSEHSFEVPIYAETLDEALELAEWQYVPAGFEVTRVRPCVAPK
P03715	MASTRGYVNIKTFEQKLDGNKKIEGKEISVAFPLYSDVHKISGAHYQTFPSEKAAYSTVYEENQRTEW IAANEDLWKVTG
P84746	PKVSPRWPPIPP
P68927	MYLTLQEWNARQRRPRSLETVRRWVRECRIFPPPVKDGREYLFHESAVKVDLNRPVTGSLLKRIRNG KKAKS
P60590	MKAQIFVVVLGLAALSVLCYGSEADESALHEEIFQLLAASDEVPKPQERDCVRFWGKCSQTSDCCPHL ACKSKWPRNICVWDGSVGK
P85875	FLSALLGMLKNL
Q09022	MRYAIVFFLVCVITLGEALKCVNLQANGIKMTQECAKEDTKCLTLRSLKKTLKFCASGRTCTTMKIMS LPGEQITCCEGNMCNA
Q912Z9	MSSDLRLTLLELVRRLNGNATIESGRLPGGRRRSPDTTTGTTGVTKTTEGPKECIDPTSRPAPEGPQE EPLHDLRPRPANRKGAAVE
P44465	MTIENDYAKLKELMEFPAKMTFKVAGINREGLAQDLIQVVQKYIKGDYIPKEKRSSKGTYNSVSIDIIA ENFDQVETLYKELAKVEGVKMVI
Q9KGL3	MNHYVYILECKDGSWYTGYTTDVDRRIKKHASGKGAKYTRGRGPFRLVATWAFPSKEEAMRWEYE VKHLSRRKKEQLVSLKGGPYENTTKLSTT
O53766	MKAKVGDWLVIKGATIDQPDHRGLIIEVRSSDGSPPYVVRWLETDHVATVIPGPDAVVVTAEEQNAA DERAQHRFGAVQSAILHARGT
O30176	MDYFRLAEKFLREMHAKYMKRVSRPGNTPRPWFDFSEERLLSRLFEEMDELREAVEKEDWENLRDE LLDVANFCMYLWGKLSVKNIYDKGEEQ
Q70LE8	MDTHEFHKLLIKVVDLFLEDRIKEFELKLNTTLDELEFEELIGKPDSSNSAENNGIFIDEYSYDASENAIK KLFVEYVRQPEFKYTVLSIKGVNDWVRE
Q58452	MIFMRKVVAEVSIIPLGKGASVSKYVKKAIEVFKKYDLKVETNAMGTVLEGDLDEILKAFKEAHSTVL NDVDRVVSSLKIDERKDKENTIERKLKAIGEL
P45026	MELQKIEQILKDTLNIAEVYAQGENAHFGVIVVSDEIAALSRVKQQQTIYAPLMPYFSTGEIHALTIKTY TVEKWKRDRALNQFN
Q7DDI1	MYFEIYKDAKGEYRWRLKAANHEIIAQGEGYTSKQNCQHAVDLLKSTTAATPVKEV
Q931U1	MAVFKVFYQHNRDEVIVRENTQSLYVEAQTEEQVRRYLKDRNFNIEFITKLEGAHLDYEKENSEHFNV EIAK
P67358	MEYEYPIDLDWSNEEMISVINFFNHVEKYYESGVTAGDFMGAYKRFKEIVPAKAEEKQIFNTFEKSSG YNSYKAVQDVKTHSEEQRVTAKK
Q9HI35	MDVKPDRVIDARGSYCPGPLMELIKAYKQAKVGEVISVYSTDAGTKKDAPAWIQKSGQELVGVFDRN GYYEIVMKKVK
P60076	MSNSDLNIERINELAKKKKEVGLTQEEAKEQTALRKAYLESFRKGFKQQIENTKVIDPEGNDVTPEKIK EIQQKRDNKN
P67291	MATWLAIIFIVAALILGLIGGFLLARKYMMDYLKKNPPINEEMLRMMMMQMGQKPSQKKINQMMTM MNKNMDQNMKSAKK
Q7A5S4	MQIELTDAAVTWFKNELELPENNKVLVFFVRYGGEFQLKQGFSPAFTVEPKEDVDIGYEQQYDDLNV VVAEKDLWYFEDDHIIVNVVDHEDEISYSTK
O27255	MITAELTVIPLGTCSTSLSSYVAAAVEALKKLNVRYEISGMGTLLEAEDLDELMEAVKAAHEAVLQAGS DRVYTTLKIDDRRDADRGLRDKVESVKEKI
Q7A0W3	MKNYSFYQFVMTVRGRHDDKGRLAEEIFDDLAFPKHDDDFNILSDYIETHGDFTLPMSVFDDLYEEY TEWLKF
P71376	MTTLSTKQKQFLKGLAHHLNPVVMLGGNGLTEGVLAEIENALNHHELIKVKVAGADRETKQLIINAIV RETKAAQVQTIGHILVLYRPSEEAKIQLPRK
O28889	MEDERIKLLFKEKALEILMTIYYESLGGNDVYIQYIASKVNSPHSYVWLIIKKFEEAKMVECELEGRTKI IRLTDKGQKIAQQIKSIIDIMENDT
Q9X1D7	MIVRVCMGSSCHLKGSYEVVRRFQELQKKYNFKLYGSLCFGNCSQGVCVEIDGRLFSRVTPENAEEIL KKVLQNG
O25966	MRIDKFLQSVGLVKRRVLATDMCNVGAVWLNGSCAKASKEVKAGDTISLHYLKGIEEYTILQIPALKN VPRKDTHLYIAPKTKE
Q58830	MNKPVKKQQPKKVIPNFEYARRLNGKKVKIFLRNGEVLDAEVTGVSNYEIMVKVGDRNLLVFKHAID YIEY
O28758	MEIMDEIKVNLQKEVSLEEAERYAKNIASKYGDGILLSVHDSKTGYRAPEVYCCGEKPWEVYACNRG ANLKISVNQFEFYFRIEVEGQAKY
P64665	MFEKIRKILADIEDSQNEIEMLLKLANLSLGDFIEIKRGSMDMPKGVNEAFFTQLSEEVERLKELINALN KIKKGLLVF

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P44254	MLSKDPKVLIKLGELEKDKSKAKKYFGDACDLRSQEGCDKYRELNQKQDTNK
Q5XA23	MIYKVFYQETKDQSPRRESTKALYLNIDATDELDGRIKARRLVEDNTYYNVEFIELLSDKHLDYEKETGVFELTEF
Q9PM34	MDFSKMGELLNQVQEKAKNIELELANREFSAKSGAGLVKVSANGKGEIIDVSIDDSLLEDKESLQILLISAINDVLAMVAQNRSSMANDVLGGFGGMKL
Q7A4S2	MAMTNEEKVLAIREKLNIVNQGLLDPEKYKNANEEELTDIYDFVQSRERLSPSEVTAIADALGQLRHD
O26271	MKFMVEVRIRLKKGMLNPEAATIERALALLGYEVEDTDTTDVITFTMDEDSLEAVEREVEDMCQRLLCNPVIHDYDVSINEMEG
Q5X9I2	MGFTDETVRFKLDDGDKRQISETLTAVYHSLDEKGYNPINQIVGYVLSGDPAYVPRYNDARNQIRKYERDEIVEELVRYYLQGNGIDVK
Q1ICG2	MFKVNEYFDGTVKSIAFEGQEGPATVGVMAPGEYEFGTAKREIMHVVSGALTVKLPGSDNWEKFAAGSQFNVPADSKFQLKVAVDTAYLCEYRD
Q99S93	MAMTVKKDNNEVRIQWRVADIKIPTSEIKNITQDQDIHAVPKLDSKDVSRIGSTFGKTNRVIIDTEDHEYIIYTQNDQKVYNELTK
Q7WKU6	MESRLLDILVCPVCKGRLEFQRAQAELVCNADRLAFPVRDGVPIMLEAEARSLDAEAPAQPS
O28186	MDLICMYVFKGEESFGESIDVYGDYLIVKVGTEFLAVPKKSIKSVEDGRIVIGEFDEEEARELGRKWLEEKSKPVTLEELKSYGFGEEGE
Q87MV2	MPITSKYTDEQVEKILAEVALVLEKHAASPELTLMIAGNIATNVLNQRVAASQRKLIAEKFAQALMSSLETPKTH
O64818	MGGGNAQKSAMARAKNLEKAKAAGKGSQLEANKKAMSIQCKVCMQTFICTTSEVKCREHAEAKHPKADVVACFPHLKK
Q8UIR1	MYKFEIYQDKAGEYRFRFKASNGETMFSSEGYKAKASAIHAIESIKRNSAGADTVDLTTMTA
O27953	MPAYVFSKESFLKFLEGHLEDDVVVVVSSDVTDFCKKLSESMVGEKEYCFAEFAFPADIFDADEDEIDEMMKYAIVFVEKEKLSEAGRNAIR
P67382	MKAIITVVGKDKSGIVAGVSGKIAELGLNIDDISQTVLDEYFTMMAVVSSDEKQDFTYLRNEFEAFGQTLNVKINIQSAAIFEAMYNI
P60876	MKKLAVILTLVGGLYYAFKKYQERVNQAPNIEY
Q57696	MIEKLAEIRKKIDEIDNKILKLIAERNSLAKDVAEIKNQLGIPINDPEREKYIYDRIRKLCKEHNVDENIGIKIFQILIEHNKALQKQYLEETQNKNKK
O53672	MSTTAELAELHDLVGGLRRCVTALKARFGDNPATRRIVIDADRILTDIELLDTDVSELDLERAAVPQPSEKIAIPDTEYDREFWRDVDDEGVGGHRY
Q481E4	MPIVSKYSNERVEKIIQDLLDVLVKEEVTPDLALMCLGNAVTNIIAQVPESKRVAVVDNFTKALKQSV
P65033	MTDSEHVGKTCQIDVLIEEHDERTRAKARLSWAGRQMVGVGLARLDPADEPVAQIGDELAIARALSDLANQLFALTSSDIEASTHQPVTGLHH
P71951	MSGHALAARTLLAAADELVGGPPVEASAAALAGDAAGAWRTAAVELARALVRAVAESHGVAAVLFAATAAAAAAVDRGDPP
Q830S9	MENKKSHYFYVLLCQDGSFYGGYTTEPERRLTEHNSGTGAKYTRLAKRRPVIMIHTEKFETRSEATKAEAAFKKLTRKQKEQYLKTFH
P67236	MSAVVVDAVEHLVRGIVDNPDDVRVDLITSRRGRTVEVHVHPDDLGKVIGRGGRTATALRTLVAGIGGRGIRVDVVDTDQ
P43979	MKTITLNIKGIHCGCCVKNLTQVLTELDGVQSADVQLEGKANITFDENRVNVAQLIEVIEDAGFDATE
O07225	MTKEKISVTVDAAVLAAIDADARAAGLNRSEMIEQALRNEHLRVALRDYTAKTVPALDIDAYAQRVYQANRAAGS
O07226	MIAPGDIAPRRDSEHELYVAVLSNALHRAADTGRVITCPFIPGRVPEDLLAMVVAVEQPNGTLLPELVQWLHVAALGAPLGNAGVAALREAASVVTALLC
P0CG96	MCSGPKQGLTLPASVDLEKETVITGRVVDGDGQAVGGAFVRLLDSSDEFTAEVVASATGDFRFFAAPGSWTLRALSAAGNGDAVVQPSGAGIHEVDVKIT
Q8VJ11	MSFGFPTFSQNRFTEQYSGLCPIAPGRGAGLQPCRRDCPVARWLVADHPVFGSDCRCRMMVGVNRVRIGRHELTGA
Q9HYE3	MLIPHDLLEADTLNNLLEDFVTREGTDNGDETPLDVRVERARHALRRGEAVILFDPESQQCQLMLRSEVPAELLRD
Q57812	MPLVGFMKEKKRATFYLYKNIDGRKLRYLLHKLENVENVDIDTLRRAIEAEKKYKRSITLTEEEEVIIQRLGKSANLLLNCELVKLDEGERA
Q9EXD3	MSKDKKNKVEQLEPVDLFERTKLEDTQVLNDVELDDIKKLEELKKELENTFEPRTRIEIKREIKELERKLRRNR
Q5B601	MPHLPELSKQEPSANTLVDNYRAKGEDLENSHHNNESRLAEGVHYDRNKAPALQEREKASTEKVNVEGGGASSSMVDNIRRGNPSGVA
O25237	MPSDSKKPTIIYPCLWDYRVIMTTKDTSTLKELLETYQRPFKLEFKNTSKNAKFYSFNVSMEVSNESERNEIFQKISQLDKVVQTL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P44746	MALLITSKCTNCDMCLPECPNEAISIGDEIYVIDPILCTECVGHYDTPTCQKVCPITNCIKPDPEHQETE EQLWERFVMIHHSDKL
Q8QL46	MQTQEQSQKKKQKAVFGIYMDKDLKTRLKVYCAKNNLQLTQAIEEAIKEYLQKRNG
Q8QHM9	MKKEIQVQGVRYYVESEDDLVSVAHELAKMGYTVQQIANALGVSERKVRRYLESC
Q9HKJ8	MVRVDQNLFNEVMYLLDELSQDITVPKNVRKVAQDSKAKLSQENESLDLRCATVLSMLDEMANDPNV PAHGRTDLYTIISKLEALSKS
Q8VKH3	MGSDCGCGGYLWSMLKRVEIEVDDDLIQKVIRRYRVKGAREAVNLALRTLLGEADTAEHGHDDEYD EFSDPNAWVPRRSRDTG
O26773	MSLRKLTEGDLDEISSFLHNTISDFILKRVSAKEIVDIDITVLVEYTDELKVDISAELYLDELSDADPGIVD EAVDAAYRSLESFLDGFRE
P44045	MLFSGDELANNTVLELRAQGQLSAFNKQPNLTFETDAPAILQQAVAQTRE
Q7A6L9	MADESKFEQAKGNVKETVGNVTDNKNLENEGKEDKASGKAKEFVENAKEKATDFIDKVKGNKGE
P44887	MYYVIFAQDIPNTLEKRLAVREQHLARLKQLQAENRLLTAGPNPAIDDENPSEAGFTGSTVIAQFENL QAAKDWAAQDPYVEAGVYADVIVKPFKKVF
P44897	MAQHSKYSDAQLSAIVNDMIAVLEKHKAPVDLSLIALGNMASNLLTTSVPQTQCEALAQAFSNSLINAV KTR
O05901	MGILDKVKNLLSQNADKVETVINKAGEFVDEQTQGNYSDAIHKLHDAASNVVGMSDQQS
O25581	MPFINIKLVPENGGPTNEQKQQLIEGVSDLMVKVLNKNKASIVVIIDEVDSNNYGLGGESVHHLRQKN
O83939	MMIDKLSGLDPVQNLRASCASEHVARAPAGDEITVSAEAQKKAELYLALEAVRSAPDVREYKIAAAEQ KLADPAYLERALSHVVERFLEEQNL
Q9X078	MAKYQVTKTLDVRGEVCPVPDVETKRALQNMKPGEILEVWIDYPMSKERIPETVKKLGHEVLEIEEV GPSEWKIYIKVK
B2RUZ4	MQPQESHVHYSRWEDGSRDGVSLGAVSSTEEASRCRRISQRLCTGKLGIAMKVLGGVALFWIIFILGYL TGYYVHKCK
Q3E791	MCPRTVLLIININHWFYDKNIVRIILTFRLDSGHISDICFINKNLANALITADISLLKRHDIRCTKYIITYYQ RYRNKEKGKFISLCKNTIISSSV
P0A8H8	MSETITVNCPTCGKTVVWGEISPFRPFCSKRCQLIDLGEWAAEEKRIPSSGDLSESDDWSEEPKQ
P0A8K5	MEKYCELIRKRYAEIASGDLGYVPDALGCVLKVLNEMAADDALSEAVREKAAYAAANLLVSDYVNE
Q47156	MHRILAEKSVNITELRKNPAKYFIDQPVAVLSNNRPAGYLLSASAFEALMDMLAEQEEKKPIKARFRPS AARLEEITRRAEQYLNDMTDDDFNDFKE
Q47149	MIQRDIEYSGQYSKDVKLAQKRHKDMNKLKYLMTLLINNTLPLPAVYKDHPLQGSWKGYRDAHVEP DWILIYKLTDKLLRFERTGTHAALFG
P0AAN5	MPTKPPYPREAYIVTIEKGKPGQTVTWYQLRADHPKPDSLISEHPTAQEAMDAKKRYEDPDKE
P39563	MLIDFCCSYIAGTHGRERAPSFTGTFVSHVSGENNCRPRRSEITQPCASGTEKKHFAATEKPCTNSLEG SRKDFLSLPLGHSYLFLFCFWRMICSEPKL
Q3E756	MSFIPIVCGMKSFDSSYDTVPGHQNLYCPNCHNYSVGPIKRKEFFTIWFIPLVPVFWGKQLHCPICNW RQDFKNDEQLNKVIQEQQNLRQKQPN
O43137	MSSALYKQSTNFTHSTGSFLQSAPVELTTVSGYQEFLKKQEKKNYEIQTVLSEDKSHGYVLKDGEVIA NIIGEAKDYLLDLAGQA
P0C5L4	MKSKTSSFPFCLVMFKKLVLLNQLSRQLVKQLLQEWSIMKLWVTLLPEFTNF
Q3E781	MFSHFEVSENRPRKQPRRKRISLGMINTVVSLDR
Q8TGK4	MELFIPCPERLKKMMLKEELRKELLILRCLYHPTIQIMLPTLGTLGTEKRKEKYALSLFEPILNCVGSAK TSG
P0C5K9	MWGLNRWLTFTMLILLITSHCCYWNKR
Q3E778	MWVVLSKEKILLKKAYYAKTILFSALVLRGVRGE
P0AAS7	MATFSLGKHPHVELCDLLKLEGWSESGAQAKIAIAEGQVKVDGAVETRKRCKIVAGQTVSFAGHSVQ VVA
P68661	MADLRKAARGRECQVRIPGVCNGNPETSVLAHIRLTGLCGTGTKPPDLIATIACSACHDEIDRRTHFV DAGYAKECALEGMARTQVIWLKEGVIKA
P56100	MWYFAWILGTLLACSFGVITALALEHVESGKAGQEDI
P0AAW9	MLELLKSLVFAVIMVPVVMAIILGLIYGLGEVFNIFSGVGKKDQPGQNH
P41039	MASGWANDDAVNEQINSTIEDAIARARGEIPRGESLDECEECGAPIPQARREAIPGVRLCIHCQQEKDL QKPAYTGYNRRGSKDSQLR
Q3E830	MEAEKQSDIKGTIAFDTHGNVIESTGVGSQRIEDIGDLSKVTLDAEGFAQVQGDSLLVHLYKRNDITLA VYTSAQ
Q3E7Z8	MCVCAIPFFEFFLPFIPHYAFLLFVSSVRFTVNERCYYLVCVLKLNCAFFFMVMIFELKRVCVSYLDRS RKIQIVSFFPFITIIFFHS

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
P24244	MWYLLWFVGILLMCSLSTLVLVWLDPRLKS
P0AB14	MPTQEAKAHHVGEWASLRNTSPEIAEAIFEVAGYDEKMAEKIWEEGSDEVLVKAFAKTDKDSLFWGE QTIERKNV
P64442	MRPFLQEYLMRRLLHYLINNIREHLMLYLFLWGLLAIMDLIYVFYF
Q8DJI1	MGIFNGIIEFLSNINFEVIAQLTMIAMIGIAGPMIIFLLAVRRGNL
P0AB43	MFCVIYRSSKRDQTYLYVEKKDDFSRVPEELMKGFGQPQLAMILPLDGRKKLVNADIEKVKQALTEQ GYYLQLPPPPEDLLKQHLSVMGQKTDDTNK
P0AB61	MNKETQPIDRETLLKEANKIIREHEDTLAGIEATGVTQRNGVLVFTGDYFLDEQGLPTAKSTAVFNMF KHLAHVLSEKYHLVD
P94425	MIVLQAYIKVKPEKREEFLSEAQSLVQHSRAEEGNAQYDLFEKVGEENTFVMLEKWKDEAAMKFHNE TAHFQGFVAKGKELLSAPLDVVRTELSE
Q3E7Z7	MTCGIENSYKSAEKKKKYRSFRFFESRDYSELCIIVGTYY
Q3E829	MLSKEALIKILSQNEGGNDMKIADEVVPMIQKYLDIFIDEAVLRSLQSHKDINGERGDKSPLELSHQDLE RIVGLLLMDM
P0C5L7	MHREGIGGWKKRIPSVSKRKTKCLTHRQMTSSFFLVFA
P0C5M0	MRVLHVMLSFLNSLLFLPICFCLLQLKATCAVRVKKYSMKKKKKR
Q6Q5X2	MRHHQNMHYAPQQQPVYVQQPPPRRESGGCCRTCCHFLCCLCLINLCCDVF
P0C5M1	MGSMILDITGNSMSAIVKVVSNIIRPLVLKNFII
P76061	MVHYEVVQYLMDCCGITYNQAVQALRSNDWDLWQAEVAIRSNKM
P67700	MKMANHPRPGDIIQESLDELNVSLREFARAMEIAPSTASRLLTGKAALTPEMAIKLSVVIGSSPQMWLN LQNAWSLAEAEKTVDVSRLRRLVTQ
P29009	MDFDTIMEKAYEEYFEGLAEGEEALSFSEFKQALSSSAKSNG
P64463	MTTYDRNRNAITTGSRVMVSGTGHTGKILSIDTEGLTAEQIRRGKTVVVEGCEEKLAPLDLIRLGMN
A5A617	MVGRYRFEFILIILILCALITARFYLS
Q3E7B0	MQDLEIFLSIFAFIFVFYFGAHRTVMNRNKSDVPYLQ
P0C5M8	MPLEVLGHLSKAFLFLARNNEHSHKKYNQ
P0AA31	MKNIVPDYRLDMVGEPCPYPAVATLEAMPQLKKGEILEVVSDCPQSINNIPLDARNHGYTVLDIQQDG PTIRYLIQK
P12049	MYKVKVYVSLKESVLDPQGSAVQHALHSMTYNEVQDVRIGKYMELTIEKSDRDLDVLVKEMCEKLLA NTVIEDYRYEVEEVVAQ
O31537	MLETVPVRCVERKITSLVVDLSGVPIVDTMVAQQLYNLSKTLFLLGVKAVFSGIRPDVAQTSIQLGLDF SEYETYGTLKQALENMGVRCIVEELEENK
P0C5M9	MLNVSMITKWFTESTCKSLLTNTDTMLPNHRKLNQELRNWKNCPFWSHLNKTKPLISNSLNVINCLHQ LSNCKTFPLVMMKTTYY
P0C5N0	MYTFSYSTHNELLEFFHLFVTIQWLALIGQKTLSQFCLYRNAAVVGFFIRFTFGTPIFLQLL
O35019	MSSPNTETLTQMIEEISQKLNMLNVGVIKAEDFSDEKIEDLTYLHRMVMKKESFSPSEMQAIAQELASL RK
P0C5N1	MRNLHHLRVGHEAHSTHHLWMWKHLRRRAISELRILRNFIRKCIRGGVFLWII
Q45U48	MLLYLYILYIALKTVSHLFFCNPCFRNLSVGDMLNPRLSLSFFTNHLLCPEAPLIIRGKGSYSAVGNFFSR KK
Q3E772	MGKRFSESAAKKAAGLARKRDQAHAKQRAQMEQLEAEEASKWEQGSRKENAKKLEEEQKRQEKAR AKKERDALLTAEEEQLGKGGKGKRKMK
P77295	MTELAQLQASAEQAAALLKAMSHPKRLLILCMLSGSPGTSAGELTRITGLSASATSQHLARMRDEGLID SQRDAQRILYSIKNEAVNAIIATLKNVYCP
P64559	MDINNKARIHWACRRGMRELDISIMPFFEHEYDSLSDDEKRIFIRLLECDDPDLFNWLMNHGKPADAE LEMMVRLIQTRNRERGPVAI
Q8XCU6	MSAVTVNDDGLVLRLYIQPKASRDSIVGLHGDEVKVAITAPPVDGQANSHLVKFLGKQFRVAKSQVVI EKGELGRHKQIKIINPQQIPPEVAALIN
Q07074	MHRKKRKKEKKRTEKDNTTNLPPLFLFPCSLSLPTLLAPVHYIPTRLTHHQAENQLFLLLFQPIIVKPL RS
Q3E7Z6	MTAFASLREPLVLANLKIKVHIYRMKR
Q3E758	MHDIWVITTSPACFEILYKYCKQKGRARMGGLIVKIIRFNHASV
P0C5N6	MKLLQGRMTYRGDAHPHPRITPSLLANYVGNFKGFAMWHATGKIIHVPVRHQTGMHFCI
P0C5N7	MDRIIRGKRDHILHCPLAAYSSNPRKYPYVKNSLRQDSLWSRGSATFPVTLWSKVILK
P0C5N4	MYFHSFLDTFSKYLGSTSCPLLRLSR

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
O07520	MLFFPWWVYLCIVGIIFSAYKLVAAAKEEEKVDQAFIEKEGQIYMERMEKERERRSSQQHEEENQNHS IA
P64618	MNIYTFDFDEIESQEDFYRDFSQTFGLAKDKVRDLDSLWDVLMNDVLPLPLEIEFVHLGEKTRRRFGA LILLFDEAEEELEGHLRFNVRH
O07580	MELVRIFKEHNVFGWISVGTAVLSLLLLNLAIISNVTFYSYQMLPFAMAAVPFGVVELFIKRGRTGPGL LGVILNLFVIICVYTIVSVDTNLQFGF
Q3E7Z5	MTRDTPEDVSTAGAKDILDVLNLLKGGEEKISEVELKLDEMEKKMDSLLVQLEDLHRDNNDLAKSSSQ K
Q3E7Z3	MPSDYTSHYPVILIKKKKKKIAGMYRHSKRYLEIMSTASAQFVGN
Q3E739	MSFSVSCKTPKTTKLLVSSISESAVALIIITIRILFSIGKSDFKKIISKEINGAETIYYRNIPESKPQGS
Q3E7Z4	MMCVCIPKKKLMDWRVYYIYSYVVCLYMCGSDCACICVLACVVQCVCFNVEMRL
P0A8C8	MAPPLSPGSRVLIALIRVYQRLISPLLGPHCRFTPTCSSYGIEALRRFGVIKGSWLTVKRVLKCHPLHPG GDDPVPPGPFDTREH
P0ADP9	MKCKRLNEVIELLQPAWQKEPDLNLLQFLQKLAKESGFDGELADLTDDILIYHLKMRDSAKDAVIPGL QKDYEEDFKTALLRARGVIKE
P32162	MKDVVDKCSTKGCAIDIGTVIDNDNCTSKFSRFFATREEAESFMTKLKELAAATSSADEGASVAYKIK DLEGQVELDAAFTFSCQAEMIIFELSLRSLA
Q9H374	MFYLTLESYLVLSHLHGSIAYFTNYTIKSRSSYLHLKICTENVAVNMFKKVYLKQKSKEVYLNQSKKSH FQMAIKHVLHRISA
Q3E7A3	MIAQSTRLAAAVSSSAASAGVSRIAASAMASTIFKRSPGNSFNSFKEYRENAKTYGPLSASLATRRHLAH APKL
P0C5P0	MEKKDLSLSVTLIDVYCSISLYCSNSTSGYFSTNNGKASAKSVRFPTGSGNGIYPPLKWTNLFCNFILIA PGLFNSSSFNVIKKGTPNNIAFL
P68206	MNKDEAGGNWKQFKGKVKEQWGKLTDDDMTIIEGKRDQLVGKIQERYGYQKDQAEKEVVDWETR NEYRW
O31639	MNKDKLRYAILKEIFEGNTPLSENDIGVTEDQFDDAVNFLKREGYIIGVHYSDDRPHLYKLGPELTEKG ENYLKENGTWSKAYKTIKEIKDWIK
C1P621	MRFYFYSQAVDESGVTR
Q3E7A7	MLGMIRWVVEGTLVAMLLSAIRRETGMIFFYNQYQLGGWIHRYLSWGEMCYTRTLKMVKRSKFFRK QLNEDGFGRINDSGPKRRGRDQSQYSSRFVELD
P0C5P2	MIKVPLPDVIFVAHRNKHTRQGNITQTKY
P0C5P3	MGHLVLVRHYVLVLLLIELMQLLLGSLGLS
P75677	MTQSVLLPPGPFTRRQAQAVTTTYSNITLEDDQGSHFRLVVRDTEGRMVWRAWNFEPDAGEGLNRY IRTSGIRTDTATR
C1P5Z8	MKENKVQQISHKLINIVVFVAIVEYAYLFLHFY
O34588	MSQLMGIITRLQSLQETAEAANEPMQRYFEVNGEKICSVKYFEKNQTFELTVFQKGEKPNTYPFDNID MVSIEIFELLQ
O31683	MKTLRLNNVTLEMAAYQEESEPKRKIAFTLNVTSETYHDIAVLLYEKTFNVEVPERDLAFRGEMTNYS TSLTNLYEPGAVSEFYIEITEIDKNADS
P0C5P6	MMTSLSLSIALLSKTDLVKISLRISTAFGISSCRDLA
Q3E825	MSLRNISMITKNLQTTAKCYVPKSSPTSTTIPVIRDASTTQCRRITTVINITSLKGYSPSPRTVHDKPIVI CTDNEEVETVSEHVKV
Q3E795	MAVSNNNNNNNSKERTQNIKEVEEKLGENPKITLKGGGKTKIMDFEQLRKPHCVRPSARFPVEDTAG GLLRTGGHRPQISDEEVSKRHHEQSHGQEDH
Q3E747	MPQKPLKVTKKAKDPRRVTKKQKNLRKAAPLQLKSKKKSLQHLKKLKKSSSLTETTERLVASKVGHL ELLRGTRKELEKGKKNSK
P0C5Q1	MIRVFIGSLPMLDLKNRVSSYWHFSSTPVARRITDHTCLM
Q5UP30	MNSYIREKVDEYRERYDLPDLKVTRSDKEGKRLKAVYTDKDGHRKKIYFGQEGAYTYADGAPDYVR NAYHARASGQYTKKGKQAISIPGSAASLSYNILW
O07638	MASEMIVDHRQKAFELLKVDAEKILKLIRVQMDNLTMPQCPLYEEVLDTQMFGLSREIDFAVRLGLVD EKDGKDLLYTLERELSALHDAFTAK
Q47272	MMFEFNMAELLRHRWGRLRLYRFPGSVLTDYRILKNYAKTLTGAGV
P32729	MSGMEWFPLLGLANRARKVVSGEDLVIKEIRNARAKLVLLTEDASSNTAKKVTDKCNYYKVPYKKVE SRAVLGRSIGKEARVVVAVTDQGFANKLISLLD
Q3E7A6	MVHFIFIALRSMRFMRRLVRNLQYLLLPITSSLLFI
Q3E842	MASSTSTSASASSSIKTNSALVSNNVVAASSVSATSTASSSAAKNTTSSSKNAAPGMVANPVSSKYGIIMA AFAAVSFVLGTGI
P0C5Q3	MWAGILEILSAFIRILFKLLYCWALFFTVLKGFSRGPLLPLIYLINKSL

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q3E766	MNCLCLCSLYSKSISAYFSEFSSTNIYKSYLRLPSVLYYVCMMHTMMPNQLDAVGIQSSESLLM
Q3E782	MAHKCASAKLLSGIMALLFNGKSLLRPICLHVHNHLVSNSDTNIVWP
P58034	MNNSNNLDYFTLYIIFSIAFMLITLLVILIAKPSTGLGEVLVTINLLNALVWLAINLVNRLRERLVNHRDQQ
P0CB62	MPSGNQEPRRDPELKRKAWLAVFLGSALFWVVVALLIWKVWG
O31797	MFESEAELRRIRIALVWIAVFLLFGACGNQDTIIETDNGNSDYETPQPTSFPLEHNHFGVMEDGYIKIYEYNESRNEVKLKKEYADDELE
Q3E841	MKSSIPITEVLPRAVGSLTFDENYNLLDTSGVAKVIEKSPIAEIIRKSNAELGRLGYSVYEDAQYIGHAFKKAGHFIVYFTPKNKNREGVVPPVGITN
Q3E7Z2	MILALGDFLPKQEDKACERPWVQFPARPVIFFHHQGGIFLFSINQPNLSCFSKLKEVNSLYVRVATYICQKNESRFRTNRLKGDQ
Q8TGJ2	MPIIGVPRCLENPFCAPAKFPLSVKKKIRI
Q3E7Z1	MISVCFVFPHSLALDFKSRCKKNRTKLCSAYYVSQVLRICKEMPYRDLILFSTVRKGVYMRLYY
Q3E7A8	MSDKPDSQVFCPNCNERLQKCLVQQNYAIIICPSLVCGYPFNQREVLENLTYVDDNDVLKVAKKRLSSRSKP
Q3E7Z0	MTLAYYGQPVKMCHILPPLRSLPVLVGKKKLKKKKSQTTNNHVIFLFTLFIKLLKTHNRMSL
Q3E767	MCKLMWCTGVVSKTALLTGNFFFSSSEFFFKATHRKSENYLNGRQT
C1P600	MKYINCVYNINYKLKPHSHYK
A5A615	MNVSSRTVVLINFFAAVGLFTLISMRFGWFI
A5A616	MLGNMNVFMAVLGIILFSGFLAAYFSHKWDD
Q45069	MDEILKQYMVLYKKMSNMINGPDYPGKEKDIQHQKDQIEVYEKQLQQGFSTDYDYDVFADSVIKCAYGDMTLEDLEAVYYGLTTPFF
A5A618	MSTDLKFSLVTTIIVLGLIVAVGLTAALH
Q2EES1	MPTKRFDKKHWKMVVVLLAICGAMLLLRWAAMIWG
P53906	MSNTKHTTSHHMELKRIIILTLLFILIMLIFRNSVSFKMTFQELLPRFYKKNSNSVSNNNRPSSIFSENLVDFDDVNMVDKTRLFIFLFFSFIITIPFMV
P40166	MSFNGISIRVITSYFLFFLDNLSKIKFSRLFSFKYRDFCDSCPLDIIIINASIRCLRSVFDFLHILTPRLNGKTTKKPKRNLRTQRVFDEKLHSHNASPN
Q3E735	MNTQELCKIFVAREYPLVVVPFIYFVLFLHQKYHTTLNYVWYPTCSKRIWVREKGRKCSFFFFSKVPRSDGFANNRCQRK
Q3E835	MNDDEDRAQLKARLWIRVEERLQQVLSSEDIKYTPRFINSLLELAYLQLGEMGSDLQAFARHAGRGVVNKSDLMLYLRKQPDLQERVTQE
Q3E7Y9	MQSMICSSEHENLTCKYWPVSFLASWCENGSGTLMQKDGSLLYAVKNFSHIFEKKIFHTNL
Q3E7Y8	MFYGSFNKCVTGYSCRMAIHYYVYRIIKSATRPDYKSNTQILVL
Q3E736	MMIIIFIELCRIADSLSWIPKSLRRTSSTFYIPNIIALLKMESQQLSQNSPTFQKHTPIGHINHDQYNSDSGSYYTLM
Q3E7Y7	MKLLFLNIIVVRRHLHCKSYRLSPWYIYIYGDYLLYTTEIPYKPFTRQP
O14468	MSTEKLEASEEPQAPLANTSETNSIKGDTENIVTVFDLANEIEKSLKDVQRQMKENDDEFSRSIQAIEDKLNKMSR
Q3E769	MQYCELDLSGQWLDTVYCEENFSDFVFIKFLNPSQFEEKIYCYTLHITKRTLENKRLLLYYEDEFKKHGHDINELVGDGIILRSCWNPRQ
Q3E832	MILALGDFLTNRKTKHARGPGFNSQLAPFIFDYLFPIGRVTDFFYFFQGPFVL
P0C5R4	MRSHFDTEYLGDLARKFKCTISWNYDIRFIVQYKVQRVILITHLNV
P64496	MGKATYTVTVTNNSNGVSVDYETETPMTLLVPEVAAEVIKDLVNTVRSYDTENEHDVCGW
C1P603	MKKTTIIMMGVAIIVVLGTELGWW
C1P602	MRIGIIFPVVIFITAVVFLAWFFIGGYAAPGA
P64508	MCGIFSKEVLSKHVDVEYRFSAEPYIGASCSNVSVLSMLCLRAKKTI
C1P604	MYIFITHFFTEYVILKYLLPI
C1P606	MGQFFAYATVITVKENDHVA
Q2EES6	MRIAKIGVIALFLFMALGGIGGVMLAGYTFILRAG
C1P609	MKIILWAVLIIFLIGLLVVTGVFKMIF
O31947	MASKKVHQINVKGFFDMDVMEVTEQTKEAEYTYDFKEILSEFNGKNVSITVKEENELPVKGVE
O34498	MAGYLNNIALNLEIVLKNKADSPEVSETLVTRICENLLLSKEVSFLKADGSVENFKLSDMEYEITNTEELPE

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
O31898	MPKYWSYPVGLAVEINNNARYGCPHHVGRKGKIIEHLHSATYDYAVSDETGDITYFKEHELTPLKGG LAYV
Q3E834	MVLFGLGRLFYVILLLINAVAVLSEERFLRRIGLGRSNDETPVFGQDQNTTKSKVVQLIGAVQTLLRIPL IGINILVIVYELLLG
O31864	MKSFYHYLLKYRHPKPKDSISEFANQAYEDHSFPKTSTDYHEISSYLELNADYLHTMATFDEAWDQYE SEVHGR
Q3E752	MVAFLELTSDVSQPFVIPSLSPVSQPSSRKNSDANVDDLNLAIANAALLDASASSRSHSRKNSLSLL
O13587	MRGETGVSIKNPRPSRPFSCFWRKGDVENIRKSDIGNEKKIDAKFNRLQYNLYYKPLSHHKAGLLYKE LFFRSCFSYTTCSLDFQGKRHQVERKAVDIVL
Q3E751	MHPLVDELTLSRYLTHGTSVLSSSLYSVAFFLFFFPNFLFFCSCPNHKWVSLPFIGMDILEALCFYREGK IRNIFEIGGLLLQSFYN
P0C5S1	MVYVMSMVSLLKRLLTVTRWKLQITG
C1P610	MKYFFMGISFMVIVWAGTFALMI
A5A621	MIERELGNWKDFIEVMLRK
P45906	MVENPMVINNWHDKLTETDVQIDFYGDEVTPVDDYVIDGGEIILRENLERYLREQLGFEFKNAQ
C1P612	MSEENKENGFNHVKTFTKIIFIFSVLVFNDNEYKITDAAVNLFIQI
C1P613	MKDVDQIFDALDCHILREYLILLFYD
C1P614	MNFLMRAIFSLLLLFTLSIPVISDCVAMAIESRFKYMMLLF
P64567	MKKKPVAQLERQHSLLENPCAYGLLSQFQAAIVVNCFTLNKII
P54494	MNTFYDVQQLLKTFGHIVYFGDRELEIEFMLDELKELYMNHMIEKEQWARAAAVLRKELEQTKNGR DFYKG
Q46868	MIDPKKIEQIARQVHESMPKGIREFGEDVEKKIRQTLQAQLTRLDLVSREEFDVQTQVLLRTREKLALL EQRISELENRSTEIKKQPDPETLPPTL
P0CD97	MSKHKHEWTESVANSGPASILSYCASSILMTVTNKFVVNLDNFNMNFVMLFVQSLVCTVTLCILRIVGV ANF
C1P618	MKIADQFHDELCRLAAINFEAHVLHG
P0C074	MAEKQRQLKLQKIYKQKYIGLGDESTTREQWQRNVRNDTLNTLQGHSASLEYVSLSRGDLSIRDTRIH LLKSMSPGYKAYLREER
C1P620	MLESIINLVSSGAVDSHTPQTAVAAVLCAAMIGLFS
P38374	MAVQTPRQRLANAKFNKNNEKYRKYGKKKEGKTEKTAPVISKTWLGILLFLLVGGGVLQLISYIL
A8DYQ1	MIKNFIFDNLIILAVPFMIKTSLKTNLIFFFLCVFVPHMAS
O34365	MGMPVEFNTLIVTKGKEVRIDENIFTLEKDGYRVYPMEIPMDVRKTKFGEKSGTAEVQKLQWEEGR TIITYKLTSLHSVN
O32067	MKPSTNRMLTRIKSVYMFIQEKGLVTTQELVDEFGITPRTIQRDLNVLAYNDLVHSPSRGKWETTRKK VKITS
P71071	MYIDITIDLKHYNGSVFDLRLSDYHPVKKVIDIAWQAQSVSMPPREGHWIRVVNKDKVFSGECKLSDC GITNGDRLEIL
O32127	MILIQNAEFELVHNFKDGFNEEAFKARYSDILNKYDYIVGDWGYGQLRLKGFFDDQNQKATFETKIST LDEYIYEYCNFGCAYFVLKRIRK
P71066	MSELFSVPYFIENLKQHIEMNQSEDKIHAMNSYYRSVVSTLVQDQLTKNAVVLKRIQHLDEAYNKVKR GESK
P70994	MPYVTVKMLEGRTDEQKRNLVEKVTEAVKETTGASEEKIVVFIEEMRKDHYAVAGKRLSDME
P94372	MTQTEIIITVAACLIVLAQGIFLFIDAKKRNHMAWVWGIVGLIQAPMPLICYYFFVIRPDRKKRGIKQ
P94373	MNISWEMILPLIVLQLALAVFALISCIKEERTNGPKWMWAAIIVCINIIGPILFFTVGRKQR
P22469	MSSTNETNQVLQRLNSLKIVETPKEQHEFGKRECYSLDSKKYSLVPATPSSSGHGKFQTELKKRRKNK LNRMYTYEADKNFIKARKSLNF
P94542	MSDGKKTKTTVDIYGQHFTIVGEESRAHMRYVAGIVDDKMREINEKNPYLDINKLAVLTAVNVVHDY VKLQEKCEKLERQLKEKD
P0AF36	MTMSLEVFEKLEAKVQQAIDTITLLQMEIEELKEKNNSLSQEVQNAQHQREELERENNHLKEQQNGW QERLQALLGRMEEV
Q9Y5V0	MARGQQKIQSQQKNAKKQAGQKKKQGHDQKAAAKAALIYTCTVCRTQMPDPKTFKQHFESKHPKT PLPPELADVQA
O73557	MGNKQAKAPESKDSPRASLIPDATHLGPQFCKSCWFENKGLVECNNHYLCLNCLTLLLSVSNRCPICK MPLPTKLRPSAAPTAPPTGAADSIRPPPYSP
Q6UY62	MGNSKSKSKLSANQYEQQTVNSTKQVAILKRQAEPSLYGRHNCRCCWFANTNLIKCSDHYICLKCLNI MLGKSSFCDICGEELPTSIVVPIEPSAPPPED

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
B0I3E2	MEAVRAYELQLELQQIRTLRQSLELKMKELEYAEGIITSLKSERRIYRAFSDLLVEITKDEAIEHIERSRL VYKREIEKLKKREKEIMEELSKLRAPLS
Q5W1E8	MAKKKSKLEIIQAILEACKSGSPKTRIMYGANLSYALTGRYIKMLMDLEIIRQEGKQYMLTKKGEELLE DIRKFNEMRKNMDQLKEKINSVLSIRQ
Q54323	MGRPYKLLNGIKLGVYIPQEWHDRLMEIAKEKNLTLSDVCRLAIKEYLDNHDKQKK
Q9P9J8	MRPGIRKLVVLNPRAYKGGSGHTTFYLLIPKDIAEALDIKPDDTFILNMEQKDGDIVLSYKRVKELKI
Q5UBU1	MIFEDKFIITTADEIPGLQLYSLGIASTISDNVDEIVENLRKQVKAKGGMGLIAFRITCADGKFLGYGTIV KADEAQFTMA
D0VWU5	MLMVVPLSEMGPGDKGIVVNILGGHNARQKLVSMGLTPGATIQVLESHPMGPIIISVGGVRFAIGKGLA GRVMVRKL
Q5JII0	MCLAVPGKVIEVNGPVAVVDFGGVKREVRLDLMPDTKPGDWVIVHTGFAIEKLDEKKAMEILEAWA EVEKAMEGF
O27018	MKTGADRFLEELPEVAESFKNFREAVRSEGKLTEREKLLISVACSVAVRCDACTRRHAEEALEAGITE GELAEAAAVAALIRAGSAMNTASAIFRD
O27908	MFIATLKGIFTLKDLPEEFRPFVDYKAGLEKKKLSDDDEIAIISIKGTQSNHVLFLSSYNSVDEIRKELEE AGAKINHTTLKILEGHL
O26975	MVAKGLIRIVLDILKPHEPIIPEYAKYLSELRGVEGVNITLMEIDKETENIKVTIQGNDLDFDEITRAIESY GGSIHSVDEVVAGRTMVEEVTTPQD
O27725	MDCRERIEKDLELLEKNLMEMKSIKLSDDEEAVVERALNYRDDSVYYLEKGDHITSFGCITYAHGLLDS LRMLHRII
Q8PS17	MNLFGQKDRGNHVSGVDRGKVIMYGLSTCVWCKKTKKLLTDLGVDFDYVYVDRLEGKEEEEAVEE VRRFNPSVSFPTTIINDEKAIVGFKEKEIRESLGF
O27849	MVFYLKVKVEDFGFREDMGLNYVRYRVSGLDEELTEKLIERLDEDTERDDGDLIITVFYEREYFPFGS EESKVKMADFIAREEIEMMVFLSSVLED
O27887	MRIVEEMVGKEVLDSSAKVIGKVKDVEVDIESQAIESLVLGKGGISEGLGLSKGETIVPYEMVKKIGDKI LLKGPEE
O27775	MVIGMKFTVITDDGKKILESGAPRRIKDVLGELEIPIETVVVKKNGQIVIDEEEIFDGDIIEVIRVIYGG
O26981	MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGELKIMGRVAS KEEIKKILS
Q04926	MKMGVKEDIRGQIIGALAGADFPINSPEELMAALPNGPDTTCKSGDVELKASDAGQVLTADDFPFKSA EEVADTIVNKAGL
O26567	MVGRRPGGGLKDTKPVVVRLYPDEIEALKSRVPANTSMSAYIRRIILNHLEDD
Q975W5	MASAIVLINTDAGGEDEVFERLKSMSEVTEVHVVYGVYDIVVKVEADSMDKLKDFVTNTIRKLPKVRS TLTMIIVEGKSLVKK
Q97ZR0	MAIRRLVLDVLKPIRGTSIVDLAERISKLDGVEGVNISVTDMDVETMGLMIIIEGTSLNFDDIRKMLEEE GCAIHSIDEVVSGNRIIEGKIKDDL
Q97U11	MDLELKELQSKMKEMYFEKDSQRGIYATFTWLVEEVGELAEALLSNNLDSIQEELADVIAWTVSIANL EGIDIEEALKKKYKL
Q97ZE1	MMAMSILEDPEFVKLRQFKGKVNFNLVMQILDEIELDLRGSDNIKTSIIYVYSSHLDEIRKNKEFYDMIA EILQRYYKKIGIENVNQLILTTIK
Q8ZYG6	MDVLQEQVFKDLKSRGFKIIEQLDDKIFIAEKKERYLFYVMVEGVEVTIQTLLSVINMGETLSMPVVLA LVSNDGTVTYYYVRKIRLPRNIYAEAV
Q8ZYG5	MASDISKCFATLGATLQDSIGKQVLVKLRDSHEIRGILRSFDQHVNLLLEDAEEIIDGNVYKRGTMVVR GENVLFISPVP
Q8ZYK2	MKKHIIIKTIPKKEEIISRDLCDCIYYYDNSVICKPIGPSKVYVSTSLENLEKCLQLHYFKKLVKNIEIFDE VHNSKPNCDKCLIVEIGGVYFVRRVN
A3MW14	MAVEIRAIENGPYEVKIGGRAIYLCRCGHSGSKPHCDGTHAKVGFKAPGAKIV
O28492	MEEELRRETLKWLERIEERVKEIEGDEGFMRNIEAYISDSRYFLEKGDLVRAFECVVWAWAWLEIGLE VGKLHETA
O29885	MVLPNQMVKSMVGKIIRVEMKGEENQLVGKLEGVDDYMNLYLTNAMECKGEEKVRSLGEIVLRGNN VVLIQPQEE
O29641	MKEIKEITKKDVQDAEIYLYGSVVEGDYSIGLSDIDVAIVSDVFEDRNRKLEFFGKITKKFFDSPFEFHIL TKKEWKMSKRFIRKYRRLDLIT
O28736	MAEVLMYGLSTCPHCKRTLEFLKREGVDFEVIWIDKLEGEERKKVIEKVHSISGSYSVPVVVKGDKHV LGYNEEKLKELIRG
O29588	MPVLIVYGPKLDVGKKREFVERLTSVAAEIYGMDRSAITILIHEPPAENVGVGGKLIADRERE
O30069	MSLEEWIKADSLEKADEYHKRYNYAVTNPVRRKILRMLDKGRSEEEIMQTLSLSKKQLDYHLKVLEAG FCIERVGERWVVTDAGKIVDKIRG
Q9YDN4	MDDETLRLQFGHLIRILPTLLEFEKKGYEPSLAEIVKASGVSEKTFFMGLKDRLIRAGLVKEETLSYRV KTLKLTEKGRRLAECLEKCRDVLGS

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9HQM9	MHKDELLELHEQMVNIKDQFLGFDHVDETAFAAYEELDVEPSHVHKSKSEHKHAVFLLGNALAAAMS EDEFSSAGRISKRMEELADDASNQL
Q9HPW4	MPAPAECCHMVFKKVLLTGTSEESFTAAADDAIDRAEDTLDNVVWAEVVDQGVEIGAVEERTYQTE VQVAFELDGSQ
Q8TIQ6	MAEVKVKLFANLREAAGTPELPLSGEKVIDVLLSLTDKYPALKYVIFEKGDEKSEILILCGSINILINGNN IRHLEGLETLLKDSDEIGILPPVSGG
Q6M142	MAFGKPAMKNVPVEAGKEYEVTIEDMGKGGDGIARIDGFVVFVPNAEKGSVINVKVTAVKEKFAFA ERV
Q6LY05	MFSAKKLSPADKLKNISSMLEEIVEDTTVPRNIRAAADNAKNALHNEEQELIVRSATAIQYLDDISEDPN MPIHTRTQIWGIVSELETIKN
Q8U440	MSEKACRHCHYITSEDRCPVCGSRDLSEEWFDLVIIVDVENSEIAKKIGAKVPGKYAIRVR
Q8TZI1	MKAKRVQAKIEIEFPSEDVAKVVYEAVLYEHLSVPYRRSEIDFKLEGKKIILDIKATDSSALRGTVNSYL RWIKAAIDVIEV
Q8U2E0	MSTRGDLIRILGEIEEKMNELKMDGFNPDIILFGREAYNFLSNLLKKEMEEEGPFTHVSNIKIEILEELGG DAVVIDSKVLGLVPGAAKRIKIIK
Q8U1Z3	MKMIKVKVIGRNIEKEIEWREGMKVRDILRAVGFNTESAIAKVNGKVVLEDDEVKDGDFVEVIPVVSG G
Q8U1N0	MDLVEKVKELCLELEEENLAKAIERFITLTHGIEKTRGEAFAKASIYGFLEGILTTLKMKYSNEKIETLL NEVKTAREETEALLRKPRPPLLVDNDL
Q8U362	MATRREKIIELLLEGDYSPSELARILDMRGKGSKKVILEDLKVISKIAKREGMVLLIKPAQCRKCGFVFK AEINIPSRCPKCKSEWIEEPRFKLERK
Q8TZN2	MGMESLLEKVLKEWKGHKVAVSVGGDHSFTGTLEDFDEEVILLKDVVDVIGNRGKQMLIGLEDINWI MLLE
Q8U3C7	MVKVKVKYFARFRQLAGVDEEEIELPEGARVRDLIEEIKKRHEKFKEEVFGEGYDEDADVNIAVNGR YVSWDEELKDGDVVGVFPPVSGG
Q8U0X6	MKWIKFTTNLTPEEAKIVQYELSTRDEFYRVFINPYAKVAEVVIDDSKVNIEELKEKLKGEVIEEKEITL QELIEGSLSWNNVLRSKA
Q8U098	MTTLKLLRKEIDKIDNQIISLLKKRLEIAQAIGKIKKELNLPIEDRKREEEVLRRAGEFREIFEKILEVSKD VQRL
Q9HJR9	MVTVRYYATLRPITKKKEETFNGISKISELLERLKVEYGSEFTKQMYDGNNLFKNVIILVNGNNITSMK GLDTEIKDDDKIDLFPPVAGG
Q9HIX0	MMQIDSIEIGGKVYQFFKSDLGNAPLLFIKGSKGYAMCGYLNMETSNKVGDIAVRVMGVKTLDDMLS AKVVEASQEAQKVGINPGDVLRNVIDKL
Q9HLX9	MREYPVKKGFPTDYDSIKRKISELGFDVKSEGDLIIASIPGISRIEIKPDKRKILVNTGDYDSDADKLAVV RTYNDFIEKLTGYSAKERKKMMTKD
O73966	MTYRVKIHKQVVKALQSLPKAHYRRFLEFRDILEYEPVPREKFDVIKLEGTGDLDLYRARLGDYRVIY SVNWKDKVIKILKLKPRGRAYK
O73967	MRMEKVGDVLKELERLKVEIQRLEAMLMPEERDEDITEEEIAELLELARDEDPENWIDAEELPEPED
O73971	MDVLAKFHTTVHRIGRIIIPAGTRKFYGIEQGDFVEIKIVKYEGEEPKEGTFTARVGEQGSVIIPKALRD VIGIKPGEVIEVLLLGHYKPRN
O58788	MEELKEIMKSHILGNPVRLGIMIFLLPRRKAPFSQIQKVLDLTPGNLDSHIRVLERNGLVKTYKVIADRP RTVVEITDFGMEEAKRFLSSLKAVIDGLDL
Q02039	MKFLVLPLSLAFLQIGLVFSTPDRCRYTLCCDGALKAVSACLHESESCLVPGDCCRGKSRLTLCSYGEG GNGFQCPTGYRQC
Q8TCY0	MNWKVLTGTTYSPWRVRMEFPLCGCLSLILHHFADKEGRTIGRRESCLATIWTISRPWQAGSLWITLS
Q7M4S4	DIVMTQSPGTLSVSPGERAT
Q7Z2R6	MKFLFLFFLRQSLALSPRLECSGAVLAHCKLCLPGLRHCPAPATREAEAREWLETRSRRLQ
Q5H9Q2	MMTSVSSDHCRGAQEKPQISAAQSTQPQKQVVQLYRPLDTS
Q5HYL3	MKDVPGFLQQSQSSGPGQPAVWHRLEELYTKKLWHQLTLQVLDFVQDPCFAQGDGLIKGESSVPRG NHSSRS
Q9BTS9	MNFKWSRRSLRVSVLTFSPRPSCHHITGCHWRVAVNVLRCYIRISAVQASKPVQCIRATGSGGACLSV CPAGEL
Q9XTN3	MDQETRDQMKNAAAEAKDNVHDKIQELKDDVGNKAAEVRDAVSSTVESIKDKLSGGS
Q8IRS7	MAAPAPALKDLPKVAENLKSQLEGFNQDKLKNASTQEKIILPTAEDVAAEKTQQSIFEGITAFNQNNL KHTETNEKNPLPDKEGEGEESVHRRHREL
Q7KMS3	MSQEVEETLKRIQSHKGVVGTIVVNNEGIPVKSTLDNTTTVQYAGLMSQLADKARSVVRDLDPSNDM TFLRVRSKKHEIMVAPDKDFILIVIQNPTD
A1ZBK7	MSLVSDEEWVEYKSKFDKNYEAEEDLMRRRIYAESKARIEEHNRKFEKGEVTWKMGINHLADLTPEE FAQRCGKKVPPN

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
E6PBT9	MNSGRANQRSKPMTGLINDKKEKIDAYLPRKCDWSNKLIFSNDQSSVQIAIAEVGENGQATGSKTNVV LCGSVRSKGEAHIALENILRERGLYPIQE
E6PBU8	MGRMHGTLAKAGKVRKQTPKVEKKDKPRKTPKGRSYKRILYNRRYAPHILATDPKKRKSPNWHAG KKEKMDAAANPVKKD
Q8I5R9	MADRKSNKNAVVKNVDMTEEMQIDAIDCANQALQKYNVEKDIAAHIKKEFDRKYDPTWHCVVGRNF GSYVTHETKNFIYFYIGQVAILLFKSG
Q8IK57	MAQVFEECVSFINGLPRTINLPNELKLDLYKYYKQSTIGNCNIKEPSAHKYIDRKKYEAWKSVENLNRE DAQKRYVDIVSEIFPYWQDGE
A0FKY4	MTKLSIFVVIAMLVMVSSAFRFQSKMKAKTSTDPEEHFDPNTNCDYTNSQDAWDYCTNYIVNSSCGEI CCNDCFDETGTGACRAQAFGNSCLNW
Q9TXD4	AGAEAEKLSGLSKYFNGTTMAGRANVAKATYAVIGLIIAYNVMKPKKK
Q9TWW4	SITNDLRAIADSYLYDQHKLREQQEENLRRRFYELSLRPYPDNL
Q9TXE5	GSPYEPSLKDSFADSLTGKDQIQPEAFLTSGADLRNNLIDLLRPVEVYTTEDSDEEW
B7XBA7	MWKNFTIVLLLISFIGLAWSSVQILRCPDGMQMLRSGQCVATTEPPFDPDSY
Q9TWD2	SDPFFRFGKQQVATDDSGELDDEILSRVSDDDKNI
Q7M3N3	SEETAALPTADDGGTGPLRAAVMAAI
Q9TWV7	SGSDYCETLKEVADEYILLSYKIEEQRAADCGGEPPNSQ
Q7M3P7	GLTPNMNSLFF
Q7M478	GLAPIDSAAVLQGPSGTVVRSGLAGGVVAHSPLAYSAPLATSAPYLHGGIVAAGPVAYAAVPAGSGLE GQWIPDINEKLYDDGSYKPHIYGF
Q7M479	IIPLTYTAGTPLTYSYSSPIAYADYGYGLSPYAAGYYGNYGYGLPLAYRNLYI
O45713	MSGFSVYVGNAPFQTTEDDLGNYFSQAGNVSNVRIVCDRETGRPRGFAFVEFTEEAAAQRAVDQFNG VDFNGRALRVNLAQNRNN
Q22759	MEYIKSRFVSLRDNLMLILDFLSEIGGPSEEQIDNYNIMKIKTNFFIDEIDFHLRGDVTSEQLMEYLGVYA IFYHRSRTELMKLLHEMCPNRHLNW
O18694	MATSTSTNPNTLLPLELIDKCIGSKIWVIMKNDKEIVGTLTGFDDYVNMVLEDVVEYENTADGKRMT KLDTILLNGNHITMLVPGGEGPEI
Q18124	MVDLKQQLQWIDYLGVVAVWLCFFGAILVISITCILWCCVSKDDDPTVFAKYGFGPQPRIPSQRLAAQ EAKAE
Q20670	MESSHNRRQRHFQILEQLCQDTKDDRNCDDGKKEEFEEPPPRPPTRRCMFYRDDTEVVKVLRAKQF
Q9U3B7	MSDEKSTTPTAQLDAPADGNMNDLTSLIQGVLQQTQDRFQHMSDQIIRRIDDMTTRIDDLEKNINDLL QSNQVEHPPSAQ
Q7M4G2	VAIPRLDEFTTHLPDIVQGVGEEDLVEFLEFRFDTMDKDGSLLLELQEFK
Q7M4I2	EVEVEIEEDVAVLTDAAFAEYVAEN
P84512	NSDDPCDELVLQENCDIEVQQCQESGGQDCSTDHEECMARLRLSCDDSNVDFETTASPSR
Q7M3P8	PLDSVYGTHGMSGFA
Q7M4A5	PTLVYFPVRGRAEAMVRLLL
A7AV23	MSADDFDAAVKYVSNTTTMMASNDDKLCFYKYYKQATVGDCNKPKPGMLQLQEKYKWEAWNALR GMSTESAKEAYVKLLDTLAPSWRN
A7ATL3	MVSKSIVEERLRSMLSPQFLKVTDNSGGCGAAFNAYIVSQQFEGKGLLDRQRLVNSAIAAEMPQIHAFT MKCLTPGEWEAKNRPEE
Q7M3M4	NIVTPRTPPPSQGK
Q25163	MFSTKMFVVFVAVCICVTQSIRFGMGKVPCPDGEVGYTCDCGEKICLYGQSCNDGQCSGDPKPSSEF EEFEIDEEEK
Q7M491	QAVFFTYPVPAAKDAPTLKTVTGQDIPTAVHIPNTPQFVYTYPFNPVTPYTIPFNYHGFPFVRAPLRP QSEVTPAAEEGNAVVET
Q7M4A1	ALIGPSGAILDDGTQVQFSKAGVTVLLEGPSGYVFSDGTLVQKKS
G1K3P1	MAAWGKTVDPEHKIRMLADMHGEFTRALGTELDSSKMLGNNRSRRYAMLIDDNKIRSVSTEPDITGL ACLLSIQRQKENKT
D0VWS4	MKFNDLEVGKEDRPMTPPFIKSVDVLWNPFEDLVPRRLPDAPPAQKDERKRARAAGGRTTEGRARS
Q7M481	TSVPASTSLVRESTIVSPYVPSYVAPVVPSTVVASGVVPATGVKGVDVKTVVPFPYTYQSVYPGVYSA GVYPAGVYPTGVYGTGYPIY
Q7M482	AAPAVVAAAPAVRTATLTTVVNNPGHAVSYRVD
P84499	GKGRKMKGKKRRRGRRALRKGGRRRRRRRGRKGRKGKKGRKGRRRRRRARKGRKGRRRRTRRR RTKRSRRAGKRRRRGGKRGRRTRRRRRR
I3NI55	SCSCKRNFLCC

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9BP37	MTYMCSILICLVLILCARGAEADDNGNYGNGMASVRTQGNTYDDLASLISYLTRHSFRRPFHECALCYS ITDPGERQRCIDMYCSYTN
Q16940	MKMLYAIAIMFLLVSLCSARTVRKAYPECGENEWLDDCGTQKPCEAKCNEEPPEEEDPICRSRGCLLP PACVCKDGFYRDTVIGDCVREEECDQHEIIHV
O96050	MKFLLLVAFLSVVVLTTKSVPVGSDCEPKLCTMDLVPHCFLNPEKGIVVVHGGCALSKYKCQNPNHE KLGYTHECEEAIKNAPRP
Q7M4A9	AITAICDAGDYALVPQPGFPHYETVCKAYDFYNCRPENDWEADLYLYNHIGECIGLAPT
Q7M3A7	GTVTAFSPFDARADAEALRKAMKTPAEVQNIK
Q9TRK0	EEEVAVGEGPGPRGDDAETGPRED
Q9TQQ5	MNYALKGQGRTLYGF
Q7M2Q3	FFWKTKPRKHGGRR
Q7M3I2	MDDREDLVYQAKGARRASWRIISSIEQKEENKGGEDKLKESKVFYYKMKGDYHRYLAEFATGNDRK E
Q7M2Q2	AKVGVAGFGRIGRLV
Q58K79	MAGLSTDDGGSPKGDVDPFYYDYETVRNGGLIFAALAFIVGLIIILSKRLRCGGKKHRPINEDEL
Q7M386	SHERAVKYLNSEKKADYQVVDDEIE
D9N169	GSHMDEPKFIPDPNAEKPDDWNEDMDGEWEAPRISNPA
P84494	LPNVLTQVSGPWK
Q9TRD4	ALEIIPRQLCDNAGFDATNILNK
Q7M2K9	SVINFGWMSSVTSTSTRYNGYGYGFGGSTPVD
Q9TRW7	TLDPGLLPGDFAADEAGARLFA
Q7M2K4	ALSALPGDNVGFNVKFALK
Q71T70	MGFDMYDWNIAAKSQEERDKVNVDLAASGVAYKERLNIPVIAEQVAREQPENLRTYFMERLRHYRQ LSLQLPKGSDPAYQKDDAVKK
Q38503	MVQNDFVDSYDVTMLLQDDDGKQYYEYHKGLSLSDFEVLYGNTADEIIKLRLDKVL
Q37964	MKPEELVRHFGDVEKAAVGVGVTPGAVYQWLQAGEIPPLRQSDIEVRTAYKLKSDFTSQRMGKEGH NSGTK
O21880	MAKQLSTARKFKMITGKDLFQQQKAMDTELKKEDGEITDLMEFVQYGLYLALFQDNIVKAKSDFSDF RSSFEFDTDGKGLKELVELWQKEI
C8ZKC7	MSQFQEVRPVAQALYPTHPSTKDALEEARLLLPGGTHHDFMRALMGYHNTLVKVMEEQC
A9J573	MLAEFEDRVAGIPCLIVVTYWEPYVPAKVSGPPEYCYPAEGGCGEWEVRDRRGRPAPWLERKLTEA ERERIDQAVFDRMEGR
O48468	MSESLLYGYFLDSWLDGTASEELLRVAVNAGDLTQEEADKIMSYPWGAWND
Q83VS7	MSNTISEKIVLMRKSEYLSRQQLADLTGVPYGTLSYYESGRSTPPTDVMMNILQTPQFTKYTLWFMTN QIAPESGQIAPALAHFGQNETTSPHSGQKTG
Q8GVZ2	MSFYRAQHRILPRVIRRAPHIRPRSQARAAAAEPSRFFSRPWSSPSGRERKDRGERRAAAMRAKWKK KRMRRLKRKRRKMRQRSK
Q9SE17	MKSAVYALLCFIFIVSGHIQELEANLMKRCTRGFRKLGKCTTLEEEKCKTLYPRGQCTCSDSKMNTHS CDCKSC
Q9SI54	MSGRKETVLDLAKFVDKGVQVKLTGGRQVTGTLKGYDQLLNLVLDEAVEFVRDHDDPLKTTDQTR RLGLIVCRGTAVMLVSPTDGTEEIANPFVTAEAV
Q9SD66	MVSSLLMSFAPATVRVYATSTKGGSGGPKEEKNPIDFVLGFMTKQDQFYETNPLLKKVDEKEGTTTG GRGTVRGGKNSAPTPVPKKSEGGFGGLGSLFKK
Q9M892	MDNKQNASYQAGQATGQTKEKAGGMMDKAKDAAASAQDSLQQTGQQMKEKAQGAADVVKDKTG MNKSH
Q9LE44	MAKNKDDIKYATAQAKLSEDEAIRVSYKHGTPLEGGKIAESEPVELFSSAQRIEKGKEQSAASGDQTQI QRDIKDIKGTRTDDSPR
D0VWX7	WEETKECAFTEFFKLAPLASNPALSVCQDASGWQMLPPAGYPTPEQLKLMCGTAECFTLIDAIKALN PNDCILVFGDVRLNVKKLVTEFEPSCF
Q94FS7	MNFKTCPAVALVAVVATVATAEDPLYCQAIGCPTLYSEANLAVSKECRDQGKLGDDFHRCCEEQCGS TTPASA
Q7M221	SPGEWCWPGMGHPMYPFPRCRAL
G1K3S4	GPTEKAAVKKMAKAIMADPSKADDVYQKWADKGYTLTQLSDFLKSKTRGKYDRVYNGYMTYRDY V
E1C9K7	GVIEEYLEKSKTNKELNDKKRLATTGANFARAYTVEFGSCKFPENFTGCQDLAKQKKVPFLSDDLDL ECEGKDKYKCGSNVFWKW

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
E1C9K9	DLEDLGNTTGQWDSYGSDAPSPYNPLQSKLFETFAAPFTKRGLLLKFLILGGGSTLAYLSATASGDILP
E1C9L3	DFIGSPTNLIMVTSTSLMLFAGRFGLAPSANRKATAGLKLEVRDSGLQTGDPAGFTLADTLACGVVGH IIGVGVVLGLKNIGAL
Q7M1F8	FIIPQQSLAPPAIIP
Q7M1X1	KRDPDWRREQQRREQQRRRREQQQQRRERQQRGERD
Q7M1F9	AFLQPSHHDADE
Q7M1H0	YTPAGPTQYPPY
Q9S8X2	VPVPQLQPQNPSQQQPQEQVPLVQQQQFPGQQQPFPPQQPYPQPQPFPSQQPYL
Q7M1X2	DLVSLSGAHTFGRSRNRFF
Q7M265	WVAQTSEEEQEGSTNAVLEGE
Q58FS3	MASTRGSGRPWSAKENKAFERALAVYDKDTPDRWANVARAVEGRTPEEVKKHYEILVEDIKYIESGK VPFPNYRTTGGNMKTDEKRFRNLKIR
Q9S8Z0	AKGSWLPGLQSPAYL
Q7M1M5	NIQVDPSGQVQWLQQQLV
Q9S9E6	PQSPQQRPPLLQQCCNELHQEEPLCVCPTLKGASKAVKQQVRQQGQQQGQQGQQLQQVISRIYQTAT HLPRVCNIRQVSICPFQKTTPGPY
Q9S9F1	PAAPFRIPKCRKEFQQAQHLRACQQWLHKQAMQFGSGSGPS
Q7M1I5	GYVLGNPAVIDGESNILYLEDPAGIVPGVVQQISLGNR
Q7M1G8	TADMDGPGNPEMT
Q7M1G5	ALESGKQKKPQQVK
Q7M1G7	ALESAKLVPWWIPP
Q54996	MSALTVDDLKKLLAETAGEDDSVDLAGELDTPFVDLGYDSLALLETAAVLQQRYGIALTDETVGRLG TPRELLDEVNTTPATA
B9A7V6	MSNLKWFSGGDDRRKKAEVIITELLDDLEIDLGNESLRKVLGSYLKKLKNEGTSVPLVLSRMNIEISNAI KKDGVSLNENQSKKLKELMSISNIRYGY
Q57468	MIVGNLGAQKAKRNDTPISAKKDIMGDKTVRVRADLHHIIKIETAKNGGNVKEVMDQALEEYIRKYLP DKL
Q04822	MSESIVTKIISIVQERQNMDDGAPVKTRDIADAAGLSIYQVRLYLEQLHDVGVLEKVNAGKGVPGLWR LL
O54465	METKYELNNTKKVANAFGLNEEDTNLLINAVDLDIKNNMQEISSELQQSEQSKQKQYGTTLQNLAKQ NRIIK
P71469	MKKFLVLRDRELNAISGGVFHAYSARGVRNNYKSAVGPADWVISAVRGFIHG
Q79DA1	MKQTDIPIWERYTLTIEEASKYFRIGENKLRRLAEENKNANWLIMNGNRIQIKRKQFEKIIDTLDAI
P71460	MKIKLTVLNEFEELTADAEKNISGGRRSRKNGIGYAIGYAFGAVERAVLGGSRDYNK
Q6WRY9	MRVSYNKLWKLLIDRDMKKGELREAVGVSKSTFAKLGKNENVSLTVLLAICEYLNCDFGDIIEALPET PDKERDS
Q8VL32	MIINNLKLIREKKKISQSELAALLEVSRQTINGIEKNKYNPSLQLALKIAYYLNTPLEDIFQWQPE
Q8KUB7	MEQQHPTIHTLKIETEFFKAVKERRKTFEIRKNDRNFQVGDILILEEYMNGMYLDDECEAEVIYITDYA QREGYVVLGIELH
P83814	MDLETLRARREAVLSLCARHGAVRVRVFGSVARGEAREDSDLDLLVAFEEGRTLLDHARLKLALEGL LGVRVDIVSERGLAPRLREQVLREAIPL
Q9RLG7	MKDLMSLVIAPIFVGLVLEMISRVLDEEDDSRK
P83819	MDWEERENLKRLVKTFAFPNFREALDFANRVGALAERENHHPRLTVEWGRVTVEWWTHSAGGVTE KDREMARLTDALLQR
Q9R5R3	MTKWNYGVFFLNFYHVGQQEPSL
Q9EV85	MPMISCDMRYGRTDEQKRALSAGLLRVISEATGEPRENIFFVIREGSGINFVEHGEHLPDYVPGNAND KALIAKLK
Q9EV84	MPFIECHIATGLSVARKQQLIRDVIDVTNKSIGSDPKIINVLLVEHAEANMSISGRIHGEAASTERTPAVS
Q8GGH0	MESFLLSKVSFVIKKIRLEKGMTQEDLAYKSNLDRTYISGIERNSRNLTIKSLELIMKGLEVSDVVFFEM LIKEILKHD
A7B1J1	MSEHREGTLFYDTETGRYDIRFDLESFYGGLHCGECFDVKVKDVWVPVRIEMGDDWYLVGLNVSRL DGLRVRM
Q3ZDQ9	MMLIDCPNCGPRNENEFKYGGEAHVAYPADPHALSDKQWSRYLFYRQNKKGIFAERWVHAAGCRK WFNALRDTVTYEFKAIYPAGAPRPEIHSAEGGTR
C2TQ79	MEVLNKQNVNIIPESEEVGGWVACVGACGTVCLASGGVGTEFAAASYFL
Q7M0S4	MQEKLDLVIEGGAVINGLGG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q7M1A6	KKIATTVGEARLSGINYRHPDSALVSYPVAAAAPLGRLPAGNYRIAIVGGGAGGIAALYELGRLAATLP AGSGIDVQIYEADPDSFLHDRPG
D0VX16	GVDRDYLQSEYGVLKAGQCYKVVRSFRDYRNINYERGDVMRFLGSNFVPYESGLSLFFDKNGSERQI MLCVRPEFQMEIAHHLDSYFCKLDDN
Q32WH4	MLIRRLKDARLRAGISQEKLGVLAGIDEASASARMNQYEKGKHAPDFEMANRLAKVLKIPVSYLYTPE DDLAQIILTWNELNEQERKRINFYIRKKAK
Q3R0L1	MNAIDIAINKLGSVSALAASLGVRQSAISNWRARGRVPAERCIDIERVTNGAVICRELRPDVFGASPAG HRPEASNAAA
Q9F4H3	MSILFLAIPLTIFVLFVAPIWLWLHYSNRQQSGAQLSHQDMQRLSQLTDDARRMRERIQALEEILDAEH PNWRQS
Q2F9Z1	MKLSELQSHIKEFDYAPEQSEHYFFKLIEEVGELSESIRKGKSGQPTLDELKGSVAEELYDVLYYVCAL ANIHGVNLEKTHELKEVLNKVKYNR
Q52QJ3	MNKKNILPQQGQPVIRLTAGQLSSQLAELSEEALGDAGLEASVTACITFCAYDGVEPSITVCISVCAYD GE
Q7M0N3	MKFNLNGQAVEFNGEPDTPLL
O87799	MATFPIMSNFERDFVIQLVPVDTEDTMDQVAEKCAYHSINRRVHPQPEKILRVRRHEDGTLFPRGMIV SDAGLRPTETLDIIFMDN
Q7M0P8	ATDIFTAPLGYISEYGVRSSML
Q9KJ82	MSLEKAHTSVKKMTFGENRDLERVVTAPVSSGKIKRVNVNFDEEKHTRFKAACARKGTSITDVVNQL VDNWLKENE
D8NA05	MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYAS TGTTTTEWCWDGNGWTKGAYTATN
O85043	MKIMQVEKTLVSTNRIADMGHKPLLVVWEKPGAPRQVAVDAIGCIPGDWVLCVGSSAAREAAGSKSY PSDLTIIGIIDQWNGE
A6B4U8	MDQFLVQIFAVIHQIPKGKVSTYGEIAKMAGYPGYARHVGKALGNLPEGSKLPWFRVINSQGKISLKG RDLDRQKQKLEAEGIEVSEIGKIALRKYKWQP
C4XAL0	MEWWVKKVQDNASASLCRVVLQSGALEMIAEIEACRLRLREGDKLTPLADARYCLNNNPTQTLKIRN ATHYSSERWTNADK
H0USY9	VSGQAQLEQLASVAAGARYLKNKCNRSDLPADEAINRAAINVGKKRGWANIDANLLSQRSAQLYQQL QQDSTPEATKCSQFNRQLAPFIDSLR
Q7SIF7	MKISGRNKLEATVKEIVKGTVMAKIVMDYKGTELVAAITIDSVADLDLVPGDKVTALVKATEMEVLK
Q7M0K1	PVFGHADLPAPDDT
Q46702	MRSLLLMGVLLISACSSGHKPPPEPDWSNTVPVNKTIPVDTQGGRNES
O52685	MKKLLAMTAVAALTMSVNVSAQDAEAIYNKACTVCHSMGVAGAPKSHNTADWEPRLAKGVDNLVK SVKTGLNAMPPGGMCTDCTDEDYKAAIEFMSKAK
Q7X0F0	MQSHHDHYADLVKFGQRLRELRTAKGLSQETLAFLSGLDRSYVGGVERGQRNVSLVNILKLATALDIE PRELFC
Q56GA4	MKALIYETLVNLANQDPEQHATIRQNLYEQLDLPFDKQLALYAGALGPASSGKLENHEAISNAVDSVV QLLEIPEH
D0VWQ8	AADAPAQLDPAGEKLYRSACVVCHASGVANAPKLGDKQAWAPFLAQGADALLATVLKGKGAMPPR GGTAADEATLRAAVAYMMDAAR
Q56446	MKDPKTLLRVSIIGTTLVALCCFTPVLVILLGVVGLSALTGYLDYVLLPALAIFIGLTIYAIQRKRQADA CCTPKFNGVKK
Q8KRK5	MAETNKGTGPMADHSHPAHGHVAGSMDITQQEKTFAGFVRMVTWAAVVIVAALIFLALANA
Q7M0R4	ANDTEDIGKGSDIEIIKRTEDKTSNKWGVTQNIQFDFVKDRNYTVHEIKVKGQN
Q8NM20	MPTKTYSEEFKRDAVALYENSDGASLQQIANDLGINRVTLKNWIIKYGSNHNVQGTTPSAAVSEAEQI RQLKKENALQRARTRHPAESC
Q8NQM9	MSLDPQLLEVLACPKDKGPLRYLESEQLLVNERLNLAYRIDDGIPVLLIDEATEWTPNN
A1KYE3	MTVATDNNPTPEAVADLKKKVRKLNSKAGQMKMDLHDLAEGLPTDYENLVETAEKTYEIFRELDQL KKKLNIWEETLK
O53692	MSLLDAHIPQLVASQSAFAAKAGLMRHTIGQAEQAAMSAQAFHQGESSAAFQAAHARFVAAAAKVNT LLDVAQANLGEAAGTYVAADAAAASTYTGF
Q7D756	MSFVITNPEALTVAATEVRRIRDRAIQSDAQVAPMTTAVRPPAADLVSEKAATFLVEYARKYRQTIAA AAVVLEEFAHALTTGADKYATAEADNIKTFS
O05893	MNDYKLFRCIQCGFEYDEALGWPEDGIAAGTRWDDIPDDWSCPDCGAAKSDFEMVEVARS
P71603	MKEAINATIQRILRTDRGITANQVLVDDLGFDSLKLFQLITELEDEFDIAISFRDAQNIKTVGDVYTSVA VWFPETAKPAPLGKGTA

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q8VK10	MSNHTYRVIEIVGTSPDGVDAAIQGGLARAAQTMRALDWFEVQSIRGHLVDGAVAHFQVTMKVGFRL EDS
Q9K2J6	MATTLPRITARVDVDTQDLLAKAAALAGMSSINSFVLNAAIEKAKQVIEREQALKLSQADAVLLMEAL DNPAVVNAKLKLASERYESKTQ
Q6BBK3	MSDIQEKIEQARQEAHAISEEKGATSPDAAAAWDAVEELQAEAAHQRQQKSETEPFFGDYCSENPDA AECLIYDD
Q8YZH5	MNPENSETYINHPTWGLLYRICMVDESQDLFTTLYAQRLFFLVGNDIKAIKFQPIGRTEARMLLENRL RNLRRNGQSQEYDQLQSVFQRTFQ
Q8NL33	MNTVRWNIAVSPDVDQSVRMFIAAQGGGRKGDLSRFIEDAVRAYLFERAVEQAKAATVGMGETELN DLIDEAVQWAREH
Q99QI5	MKKTLSLKNDFKEIKTDELEIIIGGSGSLSTFFRLFNRSFTQALGK
O68692	MTQLEEQLHNVETVRSITMQLEMALTKLKKDMMRGGDAKQYQVWQRESKALESAIAIIHYVAGDLK
O68691	MSNFSGFTKGTDIADLDAVAQTLKKPADDANKAVNDSIAALKDKPDNPALLADLQHSINKWSVIYNIN STIVRSMKDLMQGILQKFP
Q81SJ3	MYYNFTSNIMGKGSFSEEMEMGQLKNKIENKKKELIQLVARHGLDHDKVLLFSRDLDKLINKFMNVK DKVHK
Q9WYF6	MRYVLYVPDISCNHCKMRISKALEELGVKNYEVSVEEKKVVVETENLDSVLKKLEEIDYPVESYQEV
Q9WZF7	MNIDEIERKIDEAIEKEDYETLLSLLNKRKELMEGLPKDKLSEILEKDRKRLEIIEKRKTALFQEINVIRE ARSSLQKNIWTRGDTLGRG
Q9WY19	MERKKLIAKFVEIASEKMGKDLETVDEENTFKELGFDSIDVIDLVMFFEDEFALRIEDEEISKIRKVKDL IDIVIKKLEEIDDEVSEG
Q9X074	MALVLVKYGTDHPVEKLKIRSAKAEDKIVLIQNGVFWALEELETPAKVYAIKDDFLARGYSEEDSKVP LITYSEFIDLLEGEEKFIG
Q9WYV6	MPKVTVSIKVVPAVEDGRLHEVIDRAIEKISSWGMKYEVGPSNTTVEGEFEEIMDRVKELARYLEQFA KRFVLQLDIDYKAGGITIEEKVSKYR
Q99YR7	MSYEKEFLKDFEDWVKTQIQVNQLAMATSQEVAQEDGDERAKDAFIRYESKLDAYEFLLGKFDNYK NGKAFHDIPDELFGARHY
Q9X0Z3	MEKRFYILTIVVEDREKAYRQVNELLHNFSEDILLRVGYPVREENMAIIFLVLKTDNDTIGALSGKLGQI SGVRVKTVPLKR
Q9X0J6	MEVKIEKPTPEKLKELSVEKWPIWEKEVSEFDWYYDTNETCYILEGKVEVTTEDGKKYVIEKGDLVT FPKGLRCRWKVLEPVRKHYNLF
Q9X0X1	MPLFKFAIDVQYRSNVRDPRGETIERVLREEKGLPVKKLRLGKSIHLEVEAENKEKAYEIVKKACEELL VNPVVEEYEVREL
Q9X1G8	MIKVTVTNSFFEVTGHAPDKTLCASVSLLTQHVANFLKAEKKAKIKKESGYLKVKFEELENCEVKVLA AMVRSLKELEQKFPSQIRVEVIDNGS
Q8XPF0	MHKDIFTSVVRVRGSKKYNVVPVKSNKPVEISKWIDFSNVLSRLYVGVPTKSGNVVCKNIMNTGVDIIC TKNLPKDS
Q8P6W3	MALTLYQRDDCHLCDQAVEALAQARAGAFFSVFIDDDAALESAYGLRVPVLRDPMGRELDWPFDAP RLRAWLDAAPHA
Q7A1E8	MTFYNFIMGFQNDNTPFGILAEHVSEDKAFPRLEERHQVIRAYVMSNYTDHQLIETTNRAISLYMAN
Q6D5X8	MGNIYQITVEEKAEHQRTLSFEFSLHDDLFKLLEKVDGKMDMTPEQTQAFMVGLKLFGEVMMQQRK HPLFKEFSAPFRAFMMNLKKQ
Q18AW3	MIRLTIEETNLLSIYNEGGKRGLMENINAALPFMDEDMRELAKRTLAKIAPLTENEYAELAIFAADEV
Q251Q8	MEKGGLVGMEKNPSNHRLAMDNRQFLSLTGVSKVQSFDPKEILLETIQGVLSIKGEKLGIKHLDLKAG QVEVEGLIDALVYPRNSGSRQNVWAKIFR
Q2RW81	MAKAQPIEIAGHEFARKADALAFMKVMLNRYRPGDIVSTVDGAFLVEALKRHPDATSKIGPGVRNFE VRSADYGTQCFWILRTDGSEERFSYKKCV
Q7VWF8	MKRPGAIPTVQIDNERVKVTEWRFPPGGETGWHRHSMDYVVVPMTTGPLLLETPEGSVTSQLTRGV SYTRPEGVEHNVINPSDTEFVFVEIEIKAA
Q8CZ42	MEVVMDNIIDVSIPVAEVVDKHPEVLEILVELGFKPLANPLMRNTVGRKVSLKQGSKLAGTPMDKIVR TLEANGYEVIGLD
Q2W014	MTDPVIAQKAPYPVTVEAGKTYHWCACGRSKAQPFCDGSHKGTGLAPVAYTPDKAGTAYFCGCKA SKAPPLCDGTHKTL
Q8YJA2	MEFLMVDVIIYTRPGCPYCARAKALLARKGAEFNEIDASATPELRAEMQERSGRNTFPQIFIGSVHVG GCDDLYALEDEGKLDSLLKTGKLI
Q9ZLR7	MELGNKNIKPGRKRVAVDELKRNFSVTFYLSKEEHDVLRRLADEEVESVNSFVKRHILKTIIYKKGTN QDSSINCDSSSRL
Q7DDK1	MMVFDDIAKRKIRFQTRRGLLELDLIFGRFMEKEFEHLSDKELSEFSEILEFQDQELLALINGHSETDKG HLIPMLEKIRRA

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q7P1H6	MDTSNHLLPGLFRQLGLEDEPAAIRAFIDSHPLPPRVPLPEAPFWTPAQAAFLRQALECDAEWSEAAD ELAVLLQQGEA
Q7P0Y9	MLLHSVQTPRGEILNVSEQEARDVFGASEQAIADARKATILQTLRIERDERLRACDWTQVQDVVLTAD QKATWAKYRQALRDLPETVTDLSQIVWPQLPV
Q7NW74	MNVAHYRGYEIEPGHQYRDDIRKYVPYALIRKVGVPDRTPIPATYPEFYDLEADAERVSIACAKIIIDS HLDRHDQGLADLG
A9W9V0	MPMLEVFYSGDRPPDRTRKQAFAAEASAIFQRVIGTPPGRLQLIIQIVSPENTLAVIDLDRPDSDTTAEP DQQ
A9W9U6	MLLLRITMLEGRSTEQKAELARALSAAAAAAFDVPLAEVRLIIQEVPPTHWTVGGISMAELRQQASTS TQGQ
Q3BQX0	MCAHFMRKRPLDAETIRKLIESGLPEARVDVHGDDGVHFEATVVSPAFVGKPPLARHRMVYATLGEL MGGAIHALQLKTLTPEEG
Q2FDC9	MNNNEENSVFFGKKKKVSLHLLVDPDMKDEIIKYAQEKDFDNVSQAGREILKKGLEQIKSNK
Q0SZ80	MNSLAGIDMGRILLDLSNEVIKQLDDLEVQRNLPRADLLREAVDQYLINQSQTARTSVPGIWQGCEED GVEYQRKLREEW
Q97K90	MTKGIGLNEVEIKSKVKVIGIVPESKVRRKIMDMGIVRGTEIYIEGKAPMGDPIALRLRGYSLSLRKSEA KDILVEVL
Q5E7H1	MALIMTQQNNPLHGITLQKLLTELVEHYGWEELSYMVNINCFKKDPSIKSSLKFLRKTDWARERVENI YLKLQRHKERNQ
O50982	MLVQRIEKYAKSKNINATIEAIAETRLSEVVDRFDVVLLAPQSRFNKKRLEEITKPKGIPIEIINTIDYGT MNGEKVLQLAINAFNNKSSV
Q5F924	MGFTNLVSLAALIEKAFPIRYTPAGIPVLDIILKHESWQEENGQQCLVQLEIPARILGRQAEEWQYRQG DCATVEGFLAQKSRRSLMPMLRIQNIKEYKG
Q83DI6	MHIQMTGQGVDISPALRELTEKKLHRIQPCRDEISNIHIIFHINKLKKIVDANVKLPGSTINAQAESDDM YKTVDLLMHKLETQLSKYKAKKGDHR
Q5M594	MKLINTTWTHQELVNNQLDNTDAFLVETYSAGNTDVVFTQAPKHYELLISNKHRAVKDNELEVIREF FLKRKIDKDIVLMDKLRTVHTDKLIEISFPTTV
Q8CSK1	MEIIAISETPNHNTMKVSLSEPRQDNSSTTYTAAQEGQPEFINRLFEIEGVKSIFYVLDFISIDKEDNANW NELLPQIENTFAKSN
Q8CPV7	MPTENPTMFDQVAEVIERLRPFLLRDGGDCTLVDVEDGIVKLQLHGACGTCPSSTITLKAGIERALHE EVPGVIEVEQVF
Q5NHD0	MKVKIYTRNGCPYCVWAKQWFEENNIAFDETIIDDYAQRSKFYDEMNQSGKVIFPISTVPQIFIDDEHI GGFTELKANADKILNKK
Q6N1A7	MSIEVRDCNGALLADGDNVSLIKDLKLKGSSTVLKRGTMIRGIRLTDSEDEIEGRTDKIKGLVLRTEFL KKA
Q6N9A4	MTDTAAEDVRKIATALLKTAIEIVSEEDGGAHNQCKLCGASVPWLQTGDEIKHADDCPVVIAKQILSS RPKLHAV
Q6N7Y3	MMTASDRLGADPTQAASSPGGARAVSIVGNQIDSRELFTVDREIVIAHGDDRYRLRLTSQNKLILTK
Q6N882	MTSTFDRVATIIAETCDIPRETITPESHAIDDLGIDSLDFLDIAFAIDKAFGIKLPLEKWTQEVNDGKAT TEQYFVLKNLAARIDELVAAKGA
Q6NAY9	MRELLRTNDAVLLSAVGALLDGADIGHLVLDQNMSILEGSLGVIPRRVLVHEDDLAGARRLLTDAGLA HELRSDD
Q46TT3	MSDPGNEQNGDGIDPAIVEVLLVLREAGIENGATPWSLPKIAKRAQLPMSVLRRVLTQLQAAGLADVS VEADGRGHASLTQEGAALAAQLFPDPF
Q6N0Y6	MKVMIRKTATGHSAYVAKKDLEELIVEMENPALWGGKVTLANGWQLELPAMAADTPLPITVEARKL
Q6N4D8	MATADDFKLIRDIHSTGGRRQVFGSREQKPFEDLVDLGWLKRSSVDSRATHYQITERGTSAALRS
Q6N5V5	MTGPKQQPLPPDVEGREDAIEVLRAFVLDGGLSIAFMRAFEDPEMWGLLLVDIARHAARSYARESEY TEDEALERIVEMFEAELSRPTDTGATTERTQ
Q6N3W8	MLVTINGEQREVQSASVAALMTELDCTGGHFAVALNYDVVPRGKWDETPVTAGDEIEILTPRQGG
A8Z4R3	MKTLKELRTDYGLTQKELGDLFKVSSRTIQNMEKDSTNIKDSLLSKYMSAFNVKYDDIFLGNEYENFV FTNDKKKSIILAFKEKQTS
A8Z2S1	MQIEKLRGAALDELFDAILTLENREECYQFFDDLCTVNEIQSLSQRLQVAKMIKQGYTYATIEQESGAS TATISRVKRSLQWGNDAYTMILDRMNIETNE
Q9HY51	MSTPLTLIATITAAPGHAEALERELRALVAPSRAEAGCLQYDLHQDRHDSHLFYMIEQWRDDAALERH QNTEHFLRFSRGNEALLQNVKIDQLYRLA
Q9I169	MTSVFDRDDIQFQVVVNHEEQYSIWPEYKEIPQGWRAAGKSGLKKDCLAYIEEVWTDMRPLSLRQH MDKAAG
Q9HXS2	MSLNTPRNKPSRTETEAVAASSGRSAVGRRDYTEQLRRAARRNAWDLYGEHFY

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q9HWT8	MFRSTSHVRTESAARYVNRLCKHWGHKFEVELTPERGFIDFGDSNCELLAHPDHVLMILNSPDEDSLA HMQNVVADHLQRMANSESLEIAWQPAES
Q9I322	MKIESISPVQPSQDAGAEAVGHFEGRSVTRAAVRGEDRSSVAGLARWLARNVAGDPRSEQALQRLAD GDGTPLEARTVRRR
Q9I293	MSIEIDSEQGVCSVEIEGSRHRAPVDSLRIGTDAEARLSVLYIDGKRLHISEEDAQRLVVAGAEDQRRHL MADD
Q9A6G2	MIGVVATLKVQPAKAAEFEKVFLDLAAKVKANEPGCLVYQLTRSKTEEGVYKVLELYASMDALKHH GGTDYFKAAGAAMGPTMAGAPVIEYLDAVE
O66683	MVNRIELSRLIGLLLETEKRKNTEQKESGTNKIEDKVTLSKIAQELSKNDVEEKDLEKKVKELKEKIEK GEYEVSDEKVVKGLIEFFT
Q8XVB9	MADDVVITARNNGPYHIKGSFRIVTQGGRELPVEQGQAWLCRCGHSLNKPFCDGSHKRVEFDSNLDA PAAPEPPA
Q8YWG3	MKTIQPCSVEDIQSWLIDQFAQQLDVDPDDIDMEESFDNYDLNSSKALILLGRLEKWLGKELNPVLIFN YPTIAQLAKRLGELYL
Q8YR53	MPDPLMYQQDNFVVLETNQPEQFLTTIELLEKLKGELEKISFSDLPLELQKLDSLPAQAQHLIDTSCEL DVGAGKYLQWYAVRLEK
Q8YPN9	MSEHNFTDNLRWTSEAKTKLKNIPFFARSQAKARIEQLARQAEQDIVTPELVEQARLEFGQ
Q8YNP7	MNQPIELSLEQQFSIRSFATQVQNMSHDQAKDFLVKLYEQMVVREATYQELLKHQWGLDSGSTPA
Q8YVD1	MAITEFDTSLPSIRQLQNLIKQAAPVEIKLVTGDAITGRVLWQDPTCVCIADENSRQTTIWKQAIAYLQ PKG
A6L747	MKMLKEKAGALAGQIWEALNGTEGLTQKQIKKATKLKADKDFFLGLGWLLREDKVVTSEVEGEIFV KLV
Q5LST8	MMFLRKVEGPRSVTLPDGSIMTRADLPPANTRRWVASRKIAVVRGVIYGLITLAEAKQIYGLSDEEFN SWVSALAEHGKDALKVTALKKYRQL
A5TZS3	MNRFLTSIVAWLRAGYPEGIPPTDSFAVLALLCRRLSHDEVKAVANELMRLGDFDQIDIGVVITHFTDE LPSPEDVERVRARLAAQGWPLDDVRDREEHA
Q5LU41	MTKTLRTPEHVYLCQRLRQARLDAGLTQADLAERLDKPQSFVAKVETRERRLDVIEFAKWMAACEG LDVVSEIVATIAEGRAQA
Q5LL55	MSETWLPTLVTATPQEGFDLAVKLSRIAVKKTQPDAQVRDTLRAVYEKDANALIAVSAVVATHFQTI AAANDYWKD
Q71W18	MANLSELPNIGKVLEQDLIKAGIKTPGELKDVGSKEAFLRIWENDSSVCLSELYALEGAVQGIRWHGLD EAKKIELKKFHQSL
Q3IZ23	MRRQPKTRQESARMSIEAPETVVVSTWKVACDGGEGALGHPRVWLSIPHETGFVECGYCDRRYIHES FAAAK
Q3IYU5	MRDMTEETRKDLPPEALRALAEAEERRRRAKALDLPKEIGGRNGPEPVRFGDWEKKGIAIDF
A3DK08	MKITKDMIIADVLQMDRGTAPIFINNGMHCLGCPSSMGESIEDACAVHGIDADKLVKELNEYFEKKEV
Q64VS8	MDQLKTIKELINQGDIENALQALEEFLQTEPVGKDEAYYLMGNAYRKLGDWQKALNNYQSAIELNPD SPALQARKMVMDILNFYNKDMYNQ
Q47RW6	MGIRHIALFRWNDTVTPDQVEQVITALSKLPAAIPELKNYAFGADLGLAAGNYDFAVVADLDGEDGF RAYQDHPDHRAALAIIAPMLADRVAVQFAL
Q97S59	MKTRKIPLRKSVVSNEVIDKRDLLRIVKNKEGQVFIDPTGKANGRGAYIKLDNAEALEAKKKKVFNRS FSMEVEESFYDELIAYVDHKVKRRELGLE
Q97QU6	MGLMAMLLITIRRENQALVNNKDYPLEMKGTLEIL
Q97RM2	MKLKKEKKMAEFTFEIEEHLLTLSENEKGWTKEINRVSFNGAPAKFDIRAWSPDHTKMGKGITLSNEE FQTMVDAFKGN
Q97N67	MKASIALQVLPLVQGIDRIAVIDQVIAYLQTQEVTMVVTPFETVLEGEFDELMRILKEALEVAGQEADN VFANVKINVGEILSIDEKLEKYTETTH
Q97N65	MSDAFTDVAKMKKIKEEIKAHEGQVVEMTLENGRKRQKNRLGKLIEVYPSLFIVEFGDVEGDKQVNV YVESFTYSDILTEKNLIHYLD
O25025	MRDYSELEIFEGNPLDKWNDIIFHASKKLSKKELERLLELLALLETFIEKEDLEEKFESFAKALRIDEELQ QKIESRKTDIVIQSMANILSGNE
O24902	MSRVQMDTEEVREFVGHLERFKELLREEVNSLSNHFHNLESWRDARRDKFSEVLDNLKSTFNEFDEA AQEQIAWLKERIRVLEEDY
O25552	MLTIETSKKFDKDLKILVKNGFDLKLLYKVVGNLATEQPLAPKYKDHPLKGGLKDFRECHLKPDLLLV YQIKKQENTLFLVRLGSHSELF
O25501	MRNIYVGNLVYSATSEQVKELFSQFGKVFNVKLIYDRETKKPKGFGFVEMQEESVSEAIAKLDNTDFM GRTIRVTEANPKKS
O25010	MEKTENTDETRLRGTKNKLGRKPKADANKKTRAVSLYFSDEQYQKLEKMANEEEESVGSYIKRYILK ALRKIE

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
O25482	MMVEVRFFGPIKEENFFIKANDLKELRAILQEKEGLKEWLGVCAIALNDHLIDNLNTPLKDGDVISLLP PVCGG
O25839	MFHEFRDEISVLKANNPHFDKIFEKHNQLDDDIKTAEQQNASDAEVSHMKKQKLKLKDEIHSMIIEYRE KQKSERA
Q6G326	MADYNIPHFQNDLGYKIIEIGVKEFMCVGATQPFDHPHIFIDMGSTDEKICPYCSTLYRYDPSLSYNQT NPTGCLYNPK
Q9KL30	MSNQTCVENEVCEACGCAGEIGFIIREGDDVAEVSLFGSDKAHLEGKLAEYISLAKQVYANVEYEVAP VADNATELHARFKFEVSAEKLIFELKTRALAR
Q5HD32	MSNLEIKQGENKFYIGDDENNALAEITYRFVDNNEINIDHTGVSDELGGQGVGKKLLKAVVEHARENN LKIIASCSFAKHMLEKEDSYQDVYLG
Q57SJ8	MKTGYKVMLGALAFVVTNVYAAEIMKKTDFDKVASEYTKIGTISTTGEMSPLDAREDLIKKADEKGA DVVVLTSGQTENKIHGTADIYKKK
A6T888	MAQIIFNEEWMVEKALMARTGLGARQIESYRQGAWIEGVHFKRVSPSGEKTLRGTTWYNYPEINKFI RDS
B7J6R7	MSDPRTQPLEIRPLMISRVMEVDWADGHTSRLTFEHLRVECPCAECKGHTPDQAQIVTGKEHVSVVE VVPVGHYAVQLHFSDGHNTGIFTWEYLRRLDAE
Q5KVS1	MDGRQLNRLLLEWIGAWDPFGLGKDAYDVEAASVLQAVYETEDARTLAARIQSIYEFAFDEPIPFPDC LKLARRLLELKQAASCPLP
Q5KZY7	MEFAPRSVVIEEFIDTLEPMMEAYGLDQVGIFEEHGEGNRYYIGYTINKDDEMITIHMPFVKNERGELA LEKQEWTVRKDGREKKGFHSLQEAMEEVIHS
Q8A7U6	MIRLNVFVRVNETNREKAIEAAKELTACSLKEEGCIAYDTFESSTRRDVFMICETWQNAEVLAAHEKT AHFAQYVGIIQELAEMKLEKFEF
Q8A5J2	MKENKLDYIPEPMDLSLVDLPESLIQLSERIAENVHEVWAKARIDEGWTYGEKRDDIHKKHPCLVPYD ELPEEEKEYDRNTAMNTIKMVKKLGFRIEKED
Q8ABY1	MELSNELKVERIRLSLTAKSVAEEMGISRQQLCNIEQSETAPVVVKYIAFLRSKGVDLNALFDRIIVNK
Q8A574	MDKKIVGANAGKVWHALNEADGISIPELARKVNLSVESTALAVGWLARENKVVIERKNGLIEIYNEGH FDFSFG
B1L914	MKTFFLIVHTIISVALIYMVQVQMSKFSELGGAFGSGGLHTVFGRRKGLDTGGKITLVLSVLFFVSCVV TAFVLTR
A9CJC8	MEVQSMLLNDVKWEKPVTISLQNGAPRIFNGVYEAFDFLQHEWPARGDRAHEQALRLCRASLMGDV AGEIARTAFVAASRQAHCLMEDKAEAPNTIAS
A9CH91	MQHPYVGIWVTADGRIRQELLPNGRYDEARGNRKSAYQGRYEVRGAHINYWDDTGFTADGDFVSAN ELHHGGMTFYREK
A9CIQ1	MTDEDSEANALADPDNPPLSAEQLASAPRMPRIKIIRRALKLTQEEFSARYHIPLGTLRDWEQGRSEPD QPARAYLKIIAVDPEGTAAALRKGAT
A9CJD6	MCTAHQHHAETTTAPDNAGLSFHVEDMTCGHCAGVIKGAIEKTVPGAAVHADPASRTVVVGGVSDA AHIAEIITAAGYTPEARA
A9CKF1	MAKGYWIAQVDVRDSERYKDYVSTAKPAFERFGANFLARGGSVTELEGTARARNVVIEFPSVQHAID CYNSPEYQAAAKIRQEVADAEMMIVEGI
Q8ZQ92	MADLFDGMKRRMDALIAERFGMKVNINGTDCIVVESDFLAELGPVEGNGKNVVVFSGNVIPRRGDRV VLRGSEFTVTRIRRFNGKPQLTLEENNGGKGA
Q7CPV8	MKKTAAIISACMLTFALSACSGPNYVMHTNDGRSIVTDGKPQTDNDTGMISYKDANGNKQQINRTDV KEMVALEN
Q5SI58	MPRYQATLLIELKKGILDPQGRAVEGVLKDLGHPVEEVRVGKVLEIVFPAENLLEAEEKAKAMGALLA NPVMEVYALEALKELP
Q8ZN54	MEITCPVCHHALERNGDTAHCETCAKDFSLQALCPDCRQPLQVLKACGAVDYFCQNGHGLISKKRVN FVISDQ
Q5SH17	MPPRPEEVARKLRRLGFVERMAKGGHRLYTHPDGRIVVVPFHSGELPKGTFKRILRDAGLTEEEFHN L
Q5SHL1	MPHSHLHLDPKVREEARRRLLSAKGHLEGILRMLEDEKVYCVDVLKQLKAVEGALDRVGEMVLRAHL KDHVATAHERGDVEEIVEELMEALKYR
Q5SKN0	MRKVRPEELPALLEEGVLVVDVRPADRRSTPLPFAAEWVPLEKIQKGEHGLPRRPLLLVCEKGLLSQV AALYLEAEGYEAMSLEGGLQALTQGK
Q5SHL2	MLKLKVEGMTCNHCVMAVTKALKKVPGVEKVEVSLEKGEALVEGTADPKALVQAVEEEGYKAEVL A
Q5SK02	METLRVSSKSRPNSVAGAIAALLRTKGEVEVQAIGPQAVNQAVKAIAIARGYIAPDNLDLVVKPAFVKL ELENEERTALKFSIKAHPLET
Q5SK07	MITAFVLIRPRGNRVQALGEAIAELPQVAEVYSVTGPYDLVALVRLKDVEELDDVVTQGILSLEGVER TETLLAFRAYPRRLLDQGFALGQG

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q5SIT3	MEKDLLDKLGQHLVWRMGRAEDEDVLVVRVGLASATPRFRELPRLLNLPEAEMRRLVQEGRVRVE WVEE
Q7CR88	MKKRIIAAALLATVASFSTLAAEQVSKQEISHFKLVKVGTINVSQSGGQISSPSDLREKLSELADAKGGK YYHIIAAREHGPNFEAVAEVYNDATK
Q5SKG8	MVWLNGEPRPLEGKTLKEVLEEMGVELKGVAVLLNEEAFLGLEVPDRPLRDGDVVEVVALMQGG
Q53WI4	MVVLKVTLLEGRPPEKKRELVRRLTEMASRLLGEPYEEVRVILYEVRRDQWAAGGVLFSDKEGT
Q5SGN2	MVNPAERLAELDGVLMQYLLEADLLRELPPTYRLVLLPLDEPEVAAQALAWAMEAPNPEGWPSVYA LFLQGRPIRLLLLGKEVEVAPRAA
Q5SGM1	MGYRIEFDPRAEKELEKLDREVARRILRFLRERVATLEDPRSLGEPLRGPELGRFWKYRVGDYRLICH IQDREATVLVLRVGHRRDVYR
Q5SJV1	MAYGKAHLEAQLKRALAEEIQALEDPRLFLLTVEAVRLSKDGSVLSVYVEAFREEEGALRALSRAERR LVAALARRVRMRRLPRLEFLPWRASPA
Q5SKX7	MDLVPLKLVTIVAESLLEKRLVEEVKRLGAKGYTITPARGEGSRGIRSVDWEGQNIRLETIVSEEVALRI LQRLQEEYFPHYAVIAYVENVWVVRGEKYV
Q5SJJ5	MRTLKVQALWDGEAGVWVAESDDVPGLATEAATLEELLAKLAVMVPELLEENGVALELPVELRLEA TRPLVFS
Q5SHH4	MRRRYRVVVERDEEGYFVAHVPELHAHTQAQSFEELLRRLQEAIAVSLEEERAEVVGLEGALEIEAA
Q5SM75	MVPDWEEVLGLWRAGRYYEVHEVLEPYWLKATGEERRLLQGVILLAAALHQRRLGRPGLRNLRKAE ARLEGLPCPLMGLDWRSLLQEARRRLGA
Q5SGW3	MRVEALGKVAPLPQAQTPVSLNEASLEELMALPGIGPVLARRIVEGRPYARVEDLLKVKGIGPATLER LRPYLRP
Q5SLL2	MDGMGTLTRYLEEAMARARYELIADEEPYYGEIPDLPGVWATGKSLKECEANLQAALEDWLLFLLSR GETPPPLGEVRIELPHGEAA
Q87A02	MDRKLLHLLCSPDTRQPLSLLESKGLEALNKAIVSGTVQRADGSIQNQSLHEALITRDRKQVFRIEDSIP VLLPEEAIATIQIANFPDK
Q7WDN8	MAMQDFRPGVYRHYKGDHYLALGLARADETDEVVVVYTRLYARAGLPMSTRLLRIWNETVDTGAG PQPRFAYVGHVTPEQG
Q7WL96	MIQEIASILVQPGREADFEAGVAQARPLFMRARGCHGVALHRSIEAPQRYTLVVDWETVDNHMVDFR QSADFQEWRKLVGECFAEPPQVHHEQKVL
Q65I22	MNMTNNQSNDFVVIKAVEDGVNVIGLTRGTDTRFHHSEKLDKGEVMICQFTEHTSAIKVRGEALIQTA NGEMKSESKK
Q6MX43	MSVYKVIDIIGTSPTSWEQAAAEAVQRARDSVDDIRVARVIEQDMAVDSAGKITYRIKLEVSFKMRPA QPR
Q484V4	MTAYIIVGLTPKDAEKLQQYGARVASTLAKYSGEVLVKGSVEQLHGKFEHKAQVILEFPSREDAYNW YHSEEYQALISTRDLGMDSQFQLIG
Q7WAF1	MITPPDHPPRIAIQYCTQCQWLLRAAWMAQELLSTFGADLGEVALVPGTGGVFRIHYNGAPLWDREV DGGFPEAKVLKQRVRDHLDPGRPLGHIDGRPKP
A3QJE4	MNQSIIFTEQLTWDVQLSAIHFTAQQQGMVIDCYIGQKVLEHLAAEKINNSEQALSLFEQFRFDIEEQA EKLIEQEAFDVQGHIQVERVD
A3QHR8	MSAPVTLINPFKVPADKLEAAIEYWEAHRDFMAQQPGYLSTQLHQSIDEGATYQLINVAIWQSEADFY QAAQKMRQALGHVQVEGLCGNPALYRVIRT
Q2GJE9	MSEEIKAQVMESVIGCLKLNDEQKQILSGTTNLAKDFNLDSLDFVDLIMSLEERFSLEISDEDAQKLETV DDICRYIASKSSDA
Q03GF5	MFITTEGINAGYTIKDVVEATSSLMLASEDIDKYNMFDQLFDEAKQKLKKKADLLEGDGIIGLKYNTE VVEVNGAPKFLVVHGYGTVILIDK
Q4UNB3	MNKAKIFMNGQSQAVRLPKEFRFSVKEVSVIPLGKGIVLQPLPNSWKDVFQEMAEISSDDIFPEGRKDL PPQKRKYFE
Q0P9G0	MAENSILTSLLPLVVLFAIFYFLVIRPQQKQAKAHKQMLESLQKGDKIITNGGLICEVVKPEEDFIKVKL NEDNVTAKISREFIAKKIDA
Q48JU9	MKQRTLPAAFLLALSIASLAGCASPTVITLNDGREIQAVDTPKYDEESGFYEFKQLDGKQTRINKDQV RTVKDL
Q82Y06	MASKAIFYHAGCPVCVSAEQAVANAIDPSKYTVEIVHLGTDKARIAEAEKAGVKSVPALVIDGAAFHIN FGAGIDDLK
Q82SY3	MQHLEAVRNILGDVLNLGERKHTLTASSVLLGNIPELDSMAVVNVITALEEYFDFSVDDDEISAQTFET LGSLALFVEHKLSH
Q82V59	MTHHTEVFEGGTIDIEDDTSLTINGKEISYVHDAVKNKWSSRYLPYTQYDSLLDLARAIIRDTVEFSGV KE
Q82UW4	MKMRSQLLIVLQEHLRNSGLTQFKAAELLGVTQPRVSDLMRGKIDLFSLESLIDMITSIGLKVEINIKDA A

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q82XL7	MGKKKNKKTEVQQPDPMRKNWIMENMDSGVIYLLESWLKAKSQETGKEISDIFANAVEFNIVLKDW GKEKLEETNTEYQNQQRKLRKTYIEYYDREMK
Q82U33	MQIHVYDTYVKAKDGHVMHFDVFTDVRDDKKAIEFAKQWLSSIGEEGATVTSEECRFCHSQKAPDEV IEAIKQNGYFIYKMEGCN
Q82XT5	MANDGYFEPTQELSDETRDMHRAIISLREELEAVDLYNQRVNACKDKELKAILAHNRDEEKEHAAML LEWIRRCDPAFDKELKDYLFTNKPIAHE
Q82SJ4	MNNQVEPRKLVVYGREGCHLCEEMIASLRVLQKKSWFELEVINIDGNEHLTRLYNDRVPVLFAVNED KELCHYFLDSDVIGAYLS
Q82T22	MHVWPVQDAKARFSEFLDACITEGPQIVSRRGAEEAVLVPIGEWRRLQAAARPSLKQLLLSDSARTEM LVPERGKARRRQVEPLR
Q82S47	MTKLALFVRLEAKPGQEAALADFLASALPLANAESGTTAWFALKFGPSTFGVFDAFADEAGRQAHLN GQIAAALMANAATLLSSPPNIEKVELLAAKLPA
Q82WP3	MYVTIVYASVKTDKTEAFKEATRMNHEQSIREPGNMRFDILQSADDPTRFVLYEAYKTRKDAAAHKE TAHYLTWRDTVADWMAEPRKGVIYGGLYPTGDD
Q746F4	MGKRLYAVAYDIPDDTRRVKLANLLKSYGERVQLSVFECYLDERLLEDLRRRARRLLDLGQDALRIYP VAGQVEVLGVGPLPELREVQVL
Q87PC4	MKRQKRDRLERAQSQGYKAGLNGRSQEACPYQQVDARSYWLGGWRDARDEKQSGLYK
Q8KGB1	MKLSELKAGDRAEVTSVAAEPAVRRRLMDLGLVRGAKLKVLRFAPLGDPIEVNCNGMLLTMRRNEA EGITVHILAGDEGHPHGWPGFRRRHRFGKRA
Q99TQ4	MLTGKQKRYLRSLAHNIDPIFQIGKGGINENMIKQIDDTLENRELIKVHVLQNNFDDKKELAETLSEAT RSELVQVIGSMIVIYRESKENKEIELP
Q99UG9	MIANENIQDKALENFKANQTEVTVFFLNGFQMKGVIEEYDKYVVSLNSQGKQHLIYKHAISTYTVETE GQASTESEE
Q2GHF9	MTVTQSQLELLIRNAFPEAEITVTSLVGDNNHYSIKVISSQFQGKSKLEQHRMIYKVLDGLNIHAIQIQT GCK
Q8EF93	MCLSIPSQVVAVDNERQSVTVDTLGVRRDVSSHLMTEPLAIGDYVLIHIGFVMNKIDRNDALQSLELYQ EIVSKLENETTH
Q8EI81	MKTREEKMNKTILLLACLLGLVACSSQYIMSTKDGKMITSDSKPKLDKTTGMYLYYDEDGREVMIKQ EDVTQIIER
Q8EJX2	MRVFPVYAPKLIVKHARIFLTGVIWVKDLGRLEFEKGRFLLPRKSLPKVKQAILELNELIEAQNHQTKT A
Q8EDS4	MMTKKERIAIQRSMAEEALGKLKAIRQLCGAEDSSDSSDMQEVEIWTNRIKELEDWLWGESPIA
P74795	MIFPGATVRVTNVDDTYYRFEGLVQRVSDGKAAVLFENGNWDKLVTFRLSELEAVKPI
P73213	MTIQLTVPTIACEACAEAVTKAVQNEDAQATVQVDLTSKKVTITSALGEEQLRTAIASAGHEVE
Q889N2	MKTIHNARYQALLDLLLEARSAAGITQKELAARLGRPQSFVSKTENAERRLDVIEFMDFCRGIGTDPY ALLSKLEAMTPS
Q887T9	MAVETLYRSTRDLETTFVDRKLADAHDQMLELAELLTDVLIKNVPGLSEKHAEDASIYMAKNRAVFA AAFKNNATALSELSEPAESEG
Q39T60	MPTLDALTPIFRQVFDDDSIVLTRETSANDIDAWDSLSHMNLIVSLEVHYKIKFALGELQKLKNVGDLA DLVDKKLARK
Q39VC5	MNLTVNGKPSTVDGAESLNVTELLSALKVAQAEYVTVELNGEVLEREAFDATTVKDGDAVEFLYFM GGGK
Q748S4	MKLAKKDWFFIVLIVVVVGVFWAISGEVRTKKVPLDTNHKRFYDAFAQGAGKLDLDRQCVECHHEK PGGIPFPKNHPVKPADGPMRCLFCHKFK
Q74DN8	MTQKFTKDMTFAQALQTHPGVAGVLRSYNLGCIGCMGAQNESLEQGANAHGLNVEDILRDLNALA
Q74G82	MRFIPATAALLIILAGTAGAIDKITYPTRIGAVVFPHKKHQDALGECRGCHEKGPGRIDGFDKVMAHG KGCKGCHEEMKIGPVRCGDCHKGGSTH
B2IZS7	MNSKALPRQINNLEVGVYECEIHLKFRLIEEKSLLSDREQLLQVLLDALTEGSDDFLETLQASVKAQEVS EFKASPQMRRQLMRLRNAAENPPT
Q74ED8	MKRLIAAAALTLFCAGLAVAHDKVVVLEAKNGNVTFDHKKHAGVKGECKACHETEAGGKIAGMGK DWAHKTCTGCHKEMGKGPTKCGECHKK
Q5LAZ3	MVKHIVLFKLRDDVPVEEKLVVMNSFKEAIEALPAKISVIRKIEVGLNMNPGETWNIALYSEFDNLDDV KFYATHPEHVAAGKILAETKESRACVDYEF
Q5LGD2	MSNNQQMMLNRIKVVLAEKQRTNRWLAEQMGKSENTISRWCSNKSQPSLDMLVKVAELLNVDPRQLI NGKIKI
B2FQ63	MTLSELPLHTSAVVESVQDLHANDAIARRLRELGFVKGEEVRLVAKGPVGGEPLLVQVGFTRFALRIS EAKRVVVDAASQERRA
Q63K18	MSNPPTPLLADYEWSGYLTGIGRAFDDGVKDLNKQLQDAQANLTKNPSDPTALANYQMIMSEYNLYR NAQSSAVKSMKDIDSSIVSNFR

Table A.6: Xiao et al. (2013) Data Set Training Non-AMP Sequences Continued...

Definition	Sequence
Q0S2K3	MTVSDRELEECIRALLDARADSASICPSDVARAVAPDDWRPLMEPVREAAGRLADAGEVEVTQKGAV VDPRSARGPIRIRWTRTD
Q0S9B8	MVRVPLTAEELERGQRLGELLRSARGDMSMVTVAFDAGISVETLRKIETGRIATPAFFTIAAVARVLD LSLDDVAAVVTFGPVSTS
Q8BG62	MDAVTYDDVHMNFTEEEWDLLDSSQKRLYEEVMLETYQNLTDIGYNWQDHHIEEHCQSSRRHERHE RSHIGEKPYERNQCGDYSYMIG
Q7M036	VTIFVALYDYEAAVQFNSVQDLTAPEAALFSDVWSFGILEVLEQVEPQYTPGANT
Q7M035	FQILNSSELTTGETGYIPSNY
Q7M065	TTEITTTPEPGTPDPYPASPQNLGSSG
Q7M0H2	QTSLTDFYHSKVYLSPGSRSLGLPK
D0VWQ2	MKMRKLVKDFGDDYTLIQDSQEVKAILEYIGSEEEPHALFVKVGDGDYEEVWGIDSFVPYNFLEAYRL K
Q91LD0	MATFQTDADFLLVGDDTSRYEEVMKTFDTVEAVRKSDLDDRVYMVCLKQGSTFVLNGGIEELRLLT GDSTLEIQPMIVPTTE
D0VX05	MQTNVPKFTAVAEQVSAVLSQYGITGPNRAIYQGFGLKVARALNRLGGGPALVNMINGLKAYYISAF NANPTVLDAVTDIITGSPTGYVSGGSGHHHHHH
Q6Q0J9	MKIRKYMRINYYIILKVLVINGSRLEKKRLRSEILKRFDIDISDGVLYPLIDSLIDDKILREEEAPDGKVLF LTEKGMKEFEELHEFFKKIVC
Q6TRU9	MVESKKIAKKKTTLAFDEDVYHTLKLVSVYLNRDMTEIIEEAVVMWLIQNKEKLPNELKPKIDEISKRF FPAK
Q9PRP3	ASGPTQAGIVGRKRQKGEMFVGLM
Q9PRN3	VQESADGYRMQHFRWGQPLP
Q7LZ36	YVDLHVCDKSINTYYPPVQK
E7FH70	GSPNSSPASGPLPEGWEQAITPEGEIYYINHKNKTTSWLDPRLETR
Q7LZJ3	YEPAYPEEVAALKKGYEDDGYISKSADEFLNRVDEFNVSSIQFVGNLGENPL
Q70Q12	MLGAATGLMVLVAVTQGVWAMDPEGPDNDERFTYDYYRLRVVGLIVAAVLCVIGIIILLAGKCRCKF NQNKRTRSNSGTATAQHLLQPGEATEC
Q90248	MKFTTVILIMAIVLPCLFYKEMEANFVCPPGQTFQTCASSCPKTCETRNKLVLCDKKCNQRCGCISGT VLKSKDSSECVHPSKC
Q7SXR4	MSGGTPYIGSKISLISKAEIRYEGILYTIDTENSTVALAKVRSFGTEDRPTDRPIAPRDETFEYIIFRGSDI KDLTVCEPPKPIM
Q7LZS5	SFQEAAAISVNYTAAYVPTLKSDEILVRVQACGLNFGTECAGVVEAIGDLVIDRKVGDLIHMAAGGVGI AATQEVRKIAPKGVDIVL
D0VWW5	YASPKCFRYPNGVLACT
Q9PRN6	PAETPNSLDLTFNRRIMDTI
B5DCK2	MLILDGDLLKDKLKLPVIDNLFGKELLDKFQDDIKDKYGVDTKDLKILKTSEDKRFYYVSVDAGDGEK CKFKIRKDVDVPKMVGRKCRKDDDDDDGY
O89467	MLNLADRRIPGTAEENLQKSSGGVPGQNTGGQEARPNYHCQLCFLRSLGIDYLDASLRKKNKQRLKAI QQGRQPQYLL

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences

Definition	Sequence
AP00002	YVPLPNVPQPGRRPFPTFPGQGPFNPKIKWPQGY
AP00007	GNNRPVYIPQPRPPHPRL
AP00009	RFRPPIRRPPIRPPFYPPFRPPIRPPIFPPIRPPFRPPLGPFP
AP00010	RRIRPRPPRLPRPRPRPLPFPRPGPRPIPRPLPFPRPGPRPIPRPLPFPRPGPRPIPRPL
AP00011	WNPFKELERAGQRVRDAVISAAPAVATVGQAAAIARG
AP00529	WKSESVCTPGCVTGVLQTCFLQTITCNCHISK
AP00032	WNPFKELERAGQRVRDAIISAGPAVATVGQAAAIARG
AP00033	WNPFKELERAGQRVRDAIISAAPAVATVGQAAAIARG
AP00034	WNPFKELERAGQRVRDAVISAAAVATVGQAAAIARG

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00036	DFASCHTNGGICLPNRCPGHMIQIGICFRPRVKCCRSW
AP00037	VRNHVTCRINRGFCVPIRCPGRTRQIGTCFGPRIKCCRSW
AP00038	QGVRNHVTCRINRGFCVPIRCPGRTRQIGTCFGPRIKCCRSW
AP00039	QRVRNPQSCRWNMGVCIPFLCRVGMRQIGTCFGPRVPCCRR
AP00040	QVVRNPQSCRWNMGVCIPISCPGNMRQIGTCFGPRVPCCRRW
AP00041	QGVRNHVTCRIYGGFCVPIRCPGRTRQIGTCFGRPVKCCRRW
AP00042	QGVRNFVTCRINRGFCVPIRCPGHRRQIGTCLGPRIKCCR
AP00043	VRNFVTCRINRGFCVPIRCPGHRRQIGTCLGPQIKCCR
AP00044	QGVRNFVTCRINRGFCVPIRCPGHRRQIGTCLAPQIKCCR
AP00045	QGVRSYLSCWGNRGICLLNRCPGRMRQIGTCLAPRVKCCR
AP00046	GPLSCRRNGGVCIPIRCPGPMRQIGTCFGRPVKCCRSW
AP00047	GPLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
AP00048	SGISGPLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
AP00049	GIGALSAKGALKGLAKGLAEHFAN
AP00051	GIGSAILSAGKSALKGLAKGLAEHFAN
AP00052	GIGAAILSAGKSALKGLAKGLAEHF
AP00053	GIGAAILSAGKSIIKGLANGLAEHF
AP00054	GIGGALLSAGKSALKGLAKGLAEHFAN
AP00055	IIGPVLGMVGSALGGLLKKI
AP00056	LIGPVLGLVGSALGGLLKKI
AP00057	IIGPVLGLVGSALGGLLKKI
AP00059	GIGTKILGGVKTALKGALKELASTYVN
AP00063	ILGPVLSLVGSALGGLIKKI
AP00065	ILGPVLSLVGNALGGLLKNE
AP00067	SKITDILAKLGKVLAHV
AP00068	IKIMDILAKLGKVLAHV
AP00069	INIKDILAKLVKVLGHV
AP00072	VIPFVASVAAEMQHVYCAASRKC
AP00073	FLPLLAGLAANFLPKIFCKITRKC
AP00075	GLLDSLKGFAATAGKGVLQSLLSTASCKLAKTC
AP00076	GILDTLKNLAISAAKGAAQGLVNKASCKLSGQC
AP00077	GILDTLKNLAKTAGKGALQGLVKMASCKLSGQC
AP00079	GILDSLKNLAKNAGQILLNKASCKLSGQC
AP00081	GIFSKLAGKKIKNLLISGLKNVGKEVGMDVVRTGIDIAGCKIKGEC
AP00082	GIFSKLAGKKLKNLLISGLKNVGKEVGMDVVRTGIDIAGCKIKGEC
AP00083	GILSLVKGVAKLAGKGLAKEGGKFGLELIACKIAKQC
AP00084	GIFSLVKGAAKLAGKGLAKEGGKFGLELIACKIAKQC
AP00085	SLFSLIKAGAKFLGKNLLKQGACYAACKASKQC
AP00086	GIMSIVKDVAKNAAKEAAKGALSTLSCKLAKTC
AP00087	GIMSIVKDVAKTAAKEAAKGALSTLSCKLAKTC
AP00089	FLGALFKVASKVLPSVFCAITKKC
AP00092	SLFSLIKAGAKFLGKNLLKQGAQYAACKVSKEC
AP00093	GILDSFKQFAKGVGKDLIKGAAQGVLSTMSCKLAKTC
AP00096	LLPILGNLLNGLL
AP00097	VLPIIGNLLNSLL
AP00098	FLPLIGKVLSGIL
AP00099	FFPVIGRILNGIL
AP00100	LLPNLLKSLL
AP00105	FLPLFASLIGKLL
AP00106	FLPFLASLLTKVL

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00107	FLPFLASLLSKVL
AP00108	FLPFLATLLSKVL
AP00112	FLPIVGKLLSGLL
AP00113	GLLSGLKKVGKHVAKNVAVSLMDSLKCKISGDC
AP00114	SMLSVLKNLGKVGLGFVACKINKQC
AP00117	FLPFIARLAAKVFPSIICSVTKKC
AP00118	GILDSFKGVAKGVAKDLAGKLLDKLKCKITGC
AP00119	GILSSIKGVAKGVAKNVAAQLLDTLKCKITGC
AP00120	GLLDTIKGVAKTVAASMLDKLKCKISGC
AP00122	SMLSVLKNLGKVGLGLVACKINKQC
AP00123	GLFLDTLKGLAGKLLQGLKCIKAGCKP
AP00125	KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAILS
AP00127	RWKVFKKIEKVGRNIRDGVIKAAPAIEVLGQAKAL
AP00129	GWLKKIGKKIERVGQNTRDATVKGLEVAQQAANVAATVR
AP00130	GWLKKLGKRIERIGQHTRDATIQGLGIAQQAANVAATAR
AP00131	WNPFKELERAGQRVRDAIISAGPAVATVAQATALAK
AP00132	GWLKKIGKKIERVGQHTRDATIQTIAVAQQAANVAATAR
AP00133	GGLKKLGKKLEGVGKRVFKASEKALPVLTGYKAIG
AP00138	LRDLVCYCRKRGCKRRERMNGTCRKGHLMYTLCCR
AP00143	KKLLKWLKKLL
AP00147	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKA
AP00158	ALWFTMLKKLGTMALHAGKAALGAAANTISQGTQ
AP00159	ALWKNMLKGIGKLAGKAALGAVKKLVGAES
AP00161	GLWSKIKTAGKSVAKAAAKAAVKAVTNAV
AP00162	GLWNKIKEAASKAAGKAALGFVNEMV
AP00164	ALWKTIIKGAGKMIGSLAKNLLGSQAQPES
AP00166	GWGSFFKKAAHVGKHVGKAALTHYL
AP00169	GRPNPVNTKPTPYPRL
AP00172	GKPRPYSPRPTSHPRPIRV
AP00182	GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCYR
AP00183	RWKIFKKIEKMGRNIRDGIVKAGPAIEVLGSAKAI
AP00185	ICACRRRFCPNSERFSGYCRVNGARYVRCCSRR
AP00189	VSCTCRRFSCGFGERASGSCTVNGVRHTLCCRR
AP00194	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
AP00199	KYYGNGVHCTKSGCSVNWGEAFSAGVHRLANGGNGFW
AP00202	CIGNGGRCNENVGPPYCCSGFCLRQPNQGYGVCRNR
AP00203	QCIGNGGRCNENVGPPYCCSGFCLRQPGQGYGYCKNR
AP00205	ITSISLCTPGCKTGALMGCNMKTATCHCSIHVSK
AP00206	WKSESLCTPGCVTGALQTCFLQTLTCNCKISK
AP00210	GMASKAGAIAGKIAKVALKAL
AP00213	KWCFRVCYRGICYRKCR
AP00215	RWCFRVCYRGICYRKCR
AP00216	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
AP00226	VTCDLLSFKGQVNDSACAANCLSLGKAGGHCEKVGCICRKTSFKDLWDKRF
AP00227	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKAVCVCRN
AP00229	ATCDLLSGIGVQHSACALHCVFRGNRGGYCTGKGICVCRN
AP00230	GWLKKIGKKIERVGQHTRDATIQGLGIAQQAANVAATAR
AP00231	GWLKKIGKKIERVGQHTRDATIQVIGVAQQAANVAATAR
AP00232	GWLRKIGKKIERVGQHTRDATIQVLGIAQQAANVAATAR
AP00235	NPVSCVRNKGICVPIRCPGSMKQIGTCVGRAVKCCRKK

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00241	GLLGVLGSVAKHVLPHVVPVIAEHL
AP00243	GLLSSLSSVAKHVLPHVVPVIAEHL
AP00249	GLVSSIGRALGGLLADVVKSKGQPA
AP00250	GLVSSIGRALGGLLADVVKSKEQPA
AP00251	GLVSSIGKALGGLLADVVKTKEQPA
AP00253	GLWQKIKDKASELVSGIVEGVK
AP00254	GLWEKIKEKASELVSGIVEGVK
AP00255	GLWEKIKEKANELVSGIVEGVK
AP00256	GLWEKIREKANELVSGIVEGVK
AP00258	GLWQKIKSAAGDLASGIVEAIKS
AP00259	GLWQKIKNAAGDLASGIVEGIKS
AP00264	GRKSDCFRKSGFCAFLKCPSLTLISGKCSRFYLCCKRIR
AP00265	GRKSDCFRKNGFCAFLKCPYLTLISGLCSFHLC
AP00267	LFCKRGTCHFGRCPSHLIKVGSCFGFRSCCKWPWDA
AP00269	LSCKRGTCHFGRCPSHLIKGSCSGG
AP00270	QQCGRQASGRLCGNRLCCSQWGYCGSTASYCGAGCQSQCRS
AP00271	RFRPPIRRPPIRPPFNPPFRPPVRPPFRPPFRPPFRPPIGPFP
AP00278	VFHLLGKIIHHVGNFVYGFSHVF
AP00279	AFKLLGRIIHHVGNFVYGFSHVF
AP00280	LFKLLGKIIHHVGNFVHGFSHVF
AP00281	GLLRKGGEKIGEKLKKIGQKIKNFFQKLVPQPEQ
AP00282	SIPCGESCVFIPCTVTALLGCSCKSKVCYKN
AP00285	GLLCYCRKGHCKRGERVRGTCGIRFLYCCPRR
AP00289	GIFSSRKCKTVSKTFRGICTRNANC
AP00290	MFFSSKKCKTVSKTFRGPCVRNA
AP00291	MFFSSKKCKTVSKTFRGPCVRNAN
AP00293	AMWKDVLKKIGTVALHAGKAALGAVADTISQ
AP00296	LLGRCKVKSNRFHGPCLTDTHCSTVCRGEGYKGGDCHGLRRRCMCLC
AP00297	LLGRCKVKSNRFNGPCLTDTHCSTVCRGEGYKGGDCHGLRRRCMCLC
AP00298	LFCKGGSCHFGGCPSHLIKVGSCFGFRSCCKWPWNA
AP00299	GRKSDCFRKSGFCAFLKCPSLTLISGKCSRFYLCCKRIW
AP00300	GRKSDCFRKNGFCAFLKCPYLTLISGKCSRFHLCCKRIW
AP00302	GCRFCCNCCPNMSGCGVCCRF
AP00305	RVCYAIPLPICY
AP00306	RVCYAIPLPIC
AP00309	KSKEKIGKEFKRIVQRIKDFLRNLVPR
AP00312	LRDLVCYCRARGCKGRERMNGTCRKGHLLYMLCCR
AP00314	VFCTCRGFLCGSGERASGSCTINGVRHTLCCRR
AP00317	GFVDLAKKVVGGIRNALGI
AP00318	GFFDLAKKVVGGIRNALGI
AP00319	GILDFAKTVVGGIRNALGI
AP00320	GIVDFAKGVLGKIKNVLGI
AP00321	GIIDIAKKLVGGIRNVLGI
AP00322	GILDVAKTLVGKLRNVLGI
AP00326	GVGSFIHKVVSAIKNVA
AP00329	GFGPAFHSVSNFAKKHKTA
AP00331	GWLRKAAKSVGKFYYKHKYYIKAAWKIGRHAL
AP00335	PKRKSATKGDEPA
AP00340	FFGWLIRGAIHAGKAIHGLIHRRRH
AP00341	FIGLLISAGKAIHDLIRRRH

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00343	VFIDILDKMENAIHKAAQAGIGIAKPIEKMILPK
AP00344	GNNRPIYIPQPRPPHPRL
AP00346	RWKIFKKIERVGQNVRDGIIKAGPAIQVLGTAKAL
AP00347	RWKFFKKIERVGQNVRDGLIKAGPAIQVLGAAKAL
AP00348	RWKVFKKIEKVGRNIRDGVIKAGPAIAVVGQAKAL
AP00349	RWKVFKKIEKVGRHIRDGVIKAGPAITVVGQATAL
AP00350	PWNIFKEIERAVARTRDAVISAGPAVRTVAAATSVAS
AP00360	DLRFLYPRGKLPVPTLPPFNPKPIYIDMGNRY
AP00361	DLRFWNPREKLPLPTLPPFNPKPIYIDMGNRY
AP00362	VDKPDYRPRPRPPNM
AP00363	VDKPDYRPRPWPRPN
AP00365	VDKPDYRPRPWPRPNM
AP00368	GLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
AP00372	GKIPIGAIKKAGKAIGKGLRAVNIASTAHDVYTFFKPKKRH
AP00373	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKH
AP00375	GKVWDWIKSTAKKLWNSEPVKELKNTALNAAKNFVAEKIGATPS
AP00377	GWKDWLKKGKEWLKAKGPGIVKAALQAATQ
AP00378	GWKDWLNKGKEWLKKKGPGIMKAALKAATQ
AP00379	DFKDWMKTAGEWLKKKGPGILKAAMAAAT
AP00380	GLKDWVKIAGGWLKKKGPGILKAAMAAATQ
AP00381	GLVDVLGKVGGLIKKLLP
AP00383	LLKELWTKMKGAGKAVLGKIKGLL
AP00386	WLGSALKIGAKLLPSVVGLFKKKKQ
AP00389	GIWGTALKWGVKLLPKLVGMAQTKKQ
AP00390	FWGALIKGAAKLIPSVVGLFKKKQ
AP00393	YRGGYTGPIPRPPPIGRPPFRPVCNACYRLSVSDARNCCIKFGSCCHLVK
AP00398	AGFVLKGYTKTSQ
AP00406	FLSAIASMLGKFL
AP00407	FISAIASFLGKFL
AP00412	SLQPGAPNVNNKDQPWQVSPHISRDDSGNTRTDINVQRHGENNDFEAGWSKVVRGPNKAKPTWHIG GTHRW
AP00414	SIGSALKKALPVAKKIGKIALPIAKAALP
AP00415	SIGSAFKKALPVAKKIGKAALPIAKAALP
AP00416	SLGGVISGAKKVAKVAIPIGKAVLPVVAKLVG
AP00419	GIGASILSAGKSALKGFAKGLAEHFAN
AP00420	HSSGYTRPLRKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHL
AP00421	YSSGYTRPLPKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHR
AP00422	QGCKGPYTRPILRPYVRPVVSYNACTLSCRGITTTQARSCCTRLGRCCHVAKGYS
AP00423	QGYKGPYTRPILRPYVRPVVSYNACTLSCRGITTTQARSCSTRLGRCCHVAKGYS
AP00426	GVFLDALKKFAKGGMNAVLNPK
AP00434	GLMSVLGHAVGNVLGGLFKS
AP00435	GWFGKAFRSVSNFYKKHKTYIHAGLSAATLL
AP00438	GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLPCTCYRCG
AP00440	VTCFCRRRGCASRERHIGYCRFGNTIYRLCCRR
AP00441	CFCKRPVCDSGETQIGYCRLGNTFYRLCCRQ
AP00442	VTCFCRRRGCASRERLIGYCRFGNTIYGLCCRR
AP00443	ACYCRIPACLAGERRYGTCFYMGRVWAFCC
AP00444	GICACRRRFCLNFEQFSGYCRVNGARYVRCCSRR
AP00448	INLKAIAALAKKLL
AP00450	RICRIIFLRVCR

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00452	FLPVVAGLAAKVLPSIICAVTKKC
AP00453	FLPAIVGAAGQFLPKIFCAISKKC
AP00454	FLPAIVGAAGKFLPKIFCAISKKC
AP00457	GLWETIKNFGKKFTLNILHKLKCKIGGGC
AP00458	GLWETIKNFGKKFTLNILHNLKCKIGGGC
AP00459	FITLLLRKFICSITKKC
AP00460	FLPIIAGIAAKVFPKIFCAISKKC
AP00461	FLPMLAGLAASMVPKLVCLITKKC
AP00462	FLPMLAGLAASMVPKFVCLITKKC
AP00463	FLPFIAGMAAKFLPKIFCAISKKC
AP00464	FLPAIAGMAAKFLPKIFCAISKKC
AP00465	FLPFIAGVAAKFLPKIFCAISKKC
AP00466	FLPAIAGVAAKFLPKIFCAISKKC
AP00467	FLPAIVGAAAKFLPKIFCVISKKC
AP00468	FLPFIAGMAANFLPKIFCAISKKC
AP00469	FLPIIAGVAAKVFPKIFCAISKKC
AP00470	FLPIIASVAAKVFSKIFCAISKKC
AP00471	FLPIIASVAANVFSKIFCAISKKC
AP00472	FLPIIASVAAKVFPKIFCAISKKC
AP00476	GLNALKKVFQGIHEAIKLINNHVQ
AP00477	GINTLKKVIQGLHEVIKLVSNHE
AP00478	GINTLKKVIQGLHEVIKLVSNHA
AP00486	GFGSLFKFLAKKVAKTVAKQAAKQGAKYIANKQME
AP00487	GFGSLFKFLAKKVAKTVAKQAAKQGAKYIANKQTE
AP00488	GFGSLFKFLAKKVAKTVAKQAAKQGAKYVANKHME
AP00923	AISYGNGVYCNKEKCWVNKAENKQAITGIVIGGWASSLAGMGH
AP00498	GLVRKGGEKFGEKLRKIGQKIKEFFQKLALEIEQ
AP00500	GLGSVLGKALKIGANLL
AP00507	GLLSKVLGVGKKVLCGVSGLC
AP00508	GLLDSIKGMAISAGKGALQNLLKVASCKLDKTC
AP00513	FLGGLIKIVPAMICAVTKKC
AP00515	FLGGLMKAFPAIICAVTKKC
AP00519	QWGRRCCGWGPGRRYCRRWC
AP00520	DSHAKRHHGYKRKFHEKHHSHRGYRSNYLYDN
AP00521	ILGTILGLLKGL
AP00523	KFHEKHHSHRGY
AP00525	ILGPVLSMVGSALGGLIKKI
AP00526	ILGPVLGLVGNALGGLIKKI
AP00527	ILGPVISKIGGVLGGLLKNL
AP00530	VLSKSLCTPGCITGPLQTCYLCFPTFAKC
AP00533	GVVDILKGAAKDIAGHLASKVMNKL
AP00534	KSCCRNTTARNCYNVCRIPG
AP00537	AEVAPAPAAAAPAKAPKKKAAAKPKKAGPS
AP00539	GFGCPWNRYQCHSHCRSIGRLGGYCAGSLRLTCTCYRS
AP00540	GLLDTLKGAAKNVVGSLASKVMEKL
AP00545	GVLDILKNAAKNILAHAAEQI
AP00548	RFGRFLRKIRRFRPKVTITIQGSARFG
AP00554	GKIPVKAIKKAGAAIGKGLRAINIASTAHDVYSFFKPKHKKK
AP00555	KGRGKQGGKVRAKAKTRSS
AP00557	RVKRVWPLVIRTVIAGYNLYRAIKKK

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00563	KTCENLADTFRGPCFATSNCDDHCKNKEHLLSGRCRDDFRCWCTRNC
AP00566	GLFDVVKGVLKGVGKNVAGSLLEQLKCKLSGGC
AP00571	SIITMTKEAKLPQSWKQIACRLYNTC
AP00572	GIMDTIKDTAKTVAVGLLNKLKCKITGC
AP00573	GLFSKFNKKKIKSGLIKIIKTAGKEAGLEALRTGIDVIGCKIKGEC
AP00574	GLFSKFAGKGIKNLIFKGVKHIGKEVGMDVIRTGIDVAGCKIKGEC
AP00575	GLLDTFKNLALNAAKSAGVSVLNSLSCKLSKTC
AP00576	GVLGTVKNLLIGAGKSAAQSVLKTLSCKLSNDC
AP00577	GLFTLIKGAAKLIGKTVAKEAGKTGLELMACKITNQC
AP00578	FLPLLAGLAANFLPKIFCKITKKC
AP00579	GLLSGILGAGKHIVCGLSGLC
AP00580	GLFGKILGVGKKVLCGLSGMC
AP00581	GLLSGILGAGKNIVCGLSGLC
AP00582	GFSSLFKAGAKYLLKSVGKAGAQQLACKAANNCA
AP00585	SIWEGIKNAGKGFLVSILDKVRCKVAGGCNP
AP00586	FLPLLFGAISHLL
AP00587	GIMSLFKGVLKTAGKHVAGSLVDQLKCKITGGC
AP00592	GILSSFKGVAKGVAKDLAGKLLETLKCKITGC
AP00593	FLPILAGLAAKIVPKLFCLATKKC
AP00595	FLPIIGKLLSGLL
AP00596	FLPLVTGLLSGLL
AP00601	FLSLALAALPKFLCLVFKKC
AP00602	FLSLALAALPKLFCLIFKKC
AP00603	FLPLLLAGLPKLLCLFFKKC
AP00604	GLFSKFNKKKIKSGLFKIIKTAGKEAGLEALRTGIDVIGCKIKGEC
AP00607	GLFDVVKGVLKGAGKNVAGSLLEQLKCKLSGGC
AP00609	GIFDVVKGVLKGVGKNVAGSLLEQLKCKLSGGC
AP00610	GLFSVVTGVLKAVGKNVAKNVGGSLLEQLKCKISGGC
AP00616	ALSILRGLEKLAKMGIALTNCKATKKC
AP00617	ALSILKGLEKLAKMGIALTNCKATKKC
AP00620	GFLSTVKNLATNVAGTVIDTLKCKVTGGCRS
AP00622	GIFPKIIGKGIKTGIVNGIKSLVKGVGMKVFKAGLSNIGNTGCNEDEC
AP00625	KRIVQRIKDFLRNLVPRTES
AP00626	KSKEKIGKEFKRIVQRIKDFLRNLVPRTES
AP00627	RKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
AP00629	LLGDFFRKSKEKIGKEFKRIVQRIKDFLR
AP00631	KYYGNGVSCNKKGCSVDWGKAIGIIGNNSAANLATGGAAGWSK
AP00632	KYYGNGVSCNKNGCTVDWSKAIGIIGNNAAANLTTGGAAGWNKG
AP00633	KYYGNGVHCGKHSCTVDWGTAIGNIGNNAAANWATGGNAGWNK
AP00634	KYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGNHKC
AP00635	KYYGNGVHCTKSGCSVNWGEAASAGIHRLANGGNGFW
AP00637	ARSYGNGVYCNNKKCWVNRGEATQSIIGGMISGWASGLAGM
AP00638	GLIGSIGKALGGLLVDVLKPKL
AP00642	GFFALIPKIISSPIFKTLLSAVGSALSSSGGQE
AP00643	GFFAFIPKIISSPLFKTLLSAVGSALSSSGEQE
AP00644	GFFALIPKIISSPLFKTLLSAVGSALSSSGGQE
AP00645	GFFAFIPKIISSPLFKTLLSAVGSALSSSGDQE
AP00647	FLPLIAGLAANFLPKIFCAITKKC
AP00648	FLPVIAGVAAKFLPKIFCAITKKC
AP00649	GLFPKINKKKAKTGVFNIIKTVGKEAGMDLIRTGIDTIGCKIKGEC

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00650	GIFTKINKKKAKTGVFNIIKTIGKEAGMDVIRAGIDTISCKIKGEC
AP00651	GLFSILKGVGKIALKGLAKNMGKMGLDLVSCKISKEC
AP00652	GIMDTVKNVAKNLAGQLLDKLKCKITAC
AP00653	GIMDTVKNAAKDLAGQLLDKLKCRITGC
AP00654	GLLDTIKNTAKNLAVGLLDKIKCKMTGC
AP00655	GIMDSVKNVAKNIAGQLLDKLKCKITGC
AP00656	GIMDSVKNAAKNLAGQLLDTIKCKITAC
AP00665	FLPLIAGLIGKLF
AP00666	EGGGPQWAVGHFM
AP00668	EPNPDEFVGLM
AP00669	EPHPNEFVGLM
AP00670	EPNPDEFFGLM
AP00676	RLGNFFRKVKEKIGGGLKKVGQKIKDFLGNLVPRTAS
AP00679	GLFSILKGVGKIAIKGLGKNLGKMGLDLVSCKISKEC
AP00680	GLFGRLRDSLQRGGQKILEKAERIWCKIKDIFR
AP00681	RFRPPIRRPPIRPPFRPPFRPPVRPPIRPPFRPPFRPPIGPFP
AP00682	RRLRPRHQHFPSERPWPKPLPLPLPRPGPRPWPKPLPLPLPRPGLRPWPKPL
AP00683	RRLRPRRPRLPRPRPRPRPRPRSLPLPRPQPRRIPRPILLPWRPPRPIPRPQIQPIPRWL
AP00685	GIMDTVKGVAKTVAASLLDKLKCKITGC
AP00690	AFPPPNVPGPRFPPPNVPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPPFPPPIFPGPWFPPPPPFR PPPFGPPRFP
AP00692	GWFKKAWRKVKNAGRRVLKGVGIHYGVGLI
AP00694	AIGSILGALAKGLPTLISWIKNR
AP00695	FLPILGKLLSGIL
AP00697	GLFDIIKNIFSGL
AP00698	GLWQLIKDKIKDAATGFVTGIQS
AP00699	GLWQFIKDKLKDAATGLVTGIQS
AP00700	GLWQFIKDKFKDAATGLVTGIQS
AP00701	GLLGSIGNAIGAFIANKLKP
AP00702	GLLGSIGNAIGAFIANKLKPK
AP00703	GLLASLGKVLGGYLAEKLKP
AP00704	GLLGSIGKVLGGYLAEKLKPK
AP00705	GLLASLGKVLGGYLAEKLKPK
AP00707	RPDKPRPYLPRPRPPRPVR
AP00709	GFGCPNDYPCHRHCKSIPGRAGGYCGGAHRLRCTCYR
AP00711	GFGCPNNYACHQHCKSIRGYCGGYCAGWFRLRCTCYRCG
AP00712	GFGCPLNQGACHRHCRSIRRRGGYCAGFFKQTCCYRN
AP00713	GFGCPFNQGACHRHCRSIRRRGGYCAGLFKQTCTCYR
AP00715	RTCMKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC
AP00756	ALWKDILKNAGKAALNEINQLVNQ
AP00757	FLSLIPHAINAVSTLVHHF
AP00758	FLSLIPHAINAVSALANHG
AP00759	FLSLIPHAINAVSTLVHHS
AP00760	FLSLIPHAINAVSAIAKHS
AP00761	SLIPHAINAVSAIAKHF
AP00762	FLSLIPHAINAVSAIAKHF
AP00763	GLWSTIKNVGKEAAIAAGKAALGAL
AP00769	GLLGAMFKVASKVLPHVVPAITEHF
AP00771	GIGKFLHSAGKFGKAFVGEIMKS
AP00774	GKFSGFAKILKSIAKFFKGVGKVRKGFKEASDLDKNQ

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00775	GKLSGISKVLRAIAKFFKGVGKARKQFKEASDLDKNQ
AP00776	GKFSVFSKILRSIAKVFKGVGKVRKGFKTASDLDKNQ
AP00777	GKGRWLERIGKAGGIIIGGALDHL
AP00778	WLRRIGKGVKIIGGAALDHL
AP00782	GWGSIFKHGRHAAKHIGHAAVNHYL
AP00783	RWGKWFKKATHVGKHVGKAALTAYL
AP00786	GWKKWFNRAKKVGKTVGGLAVDHYL
AP00787	GWRLLLKKAEVKTVGKLALKHYL
AP00790	GWKKWLRKGAKHLGQAAIK
AP00792	FLGLLFHGVHHVGKWIHGLIHGHH
AP00795	ILGPILGLVSNALGGLL
AP00796	IIGPVLGLVGKPLESLLE
AP00804	RCVCTRGFCRCVCTRGFC
AP00808	CRFCCRCCPRMRGCGLCCRF
AP00728	RWCVYAYVRIRGVLVRYRRCW
AP00732	SFGLCRLRRGSCAHGRCRFPSIPIGRCSRFVQCCRRVW
AP00733	LLGDFFRKAREKIGEEFKRIVQRIKDFLRNLVPRTES
AP00734	SLGNFFRKARKKIGEEFKRIVQRIKDFLQHLIPRTEA
AP00735	RLGNFFRKAKKKIGRGLKKIGQKIKDFLGNLVPRTES
AP00736	RLGDILQKAREKIEGGLKKLVQKIKDFFGKFAPRTES
AP00738	GLVTGLLKTAGKLLGDLFGSLTG
AP00739	GVVTDLLKTAGKLLGNLFGSLSG
AP00740	GVVTDLLKTAGKLLGNLVGSLSG
AP00742	SPIHACRYQRGVCIPGPCRWPYYRVGSCGSGLKSCCVRNRWA
AP00745	MTPFWRGVSLRPVGASCRDNSECITMLCRKNRCFLRTASE
AP00815	GLGSFLKNAIKIAGKVGSTIGKVADAIGNKE
AP00816	GLGSFFKNAIKIAGKVGSTIGKVADAIGNKE
AP00819	FLPLLASLFSRLF
AP00820	FLPLLASLFSGLF
AP00821	GLFNVFKGLKTAGKHVAGSLLNQLKCKVSGGC
AP00822	GIFNVFKGALKTAGKHVAGSLLNQLKCKVSGEC
AP00825	SILPTIVSFLTKFL
AP00826	FILPLIASFLSKFL
AP00827	SMISVLKNLGKVGLGFVACKVNKQC
AP00829	FLGGLMKIIPAAFCAVTKKC
AP00830	GLLLDTLKGAAKDIAGIALEKLKCKITGCKP
AP00831	GLLSGILGAGKHIVCGLTGCAKA
AP00834	KVNANAIKKGGKAIGKGFKVISAASTAHDVYEHIKNRRH
AP00836	KVPIGAIKKGGKIIKKGLGVIGAAGTAHEVYSHVKNRH
AP00837	KVPIGAIKKGGKIIKKGLGVLGAAGTAHEVYNHVRNRQ
AP00838	KVPIGAIKKGGKIIKKGLGVIGAAGTAHEVYSHVKNRQ
AP00839	KVPVGAIKKGGKAIKTGLGVVGAAGTAHEVYSHIRNRH
AP00842	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLGGAIPGKC
AP00843	TKYYGNGVYCNSKKCWVDWGTAQGCIDVVIGQLGGGIPGKGKC
AP00844	KYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGTHKC
AP00845	KSYGNGVHCNKKKCWVDWGSAISTIGNNSAANWATGGAAGWKS
AP00848	KNYGNGVHCTKKGCSVDWGYAWANIANNSVMNGLTGGNAGWHN
AP00850	KYYGNGVSCNSHGCSVNWGQAWTCGVNHLANGGHGVC
AP00851	KYYGNGVTCGKHSCSVNWGQAFSCSVSHLANFGHGKC
AP00852	KYYGNGLSCSKKGCTVNWGQAFSCGVNRVATAGHHKC

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00853	ATRSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH
AP00854	ATYYGNGLYCNKQKCWVDWNKASREIGKIIVNGWVQHGPWAPR
AP00855	RCVCTRGFCRCICLLGIC
AP00856	RCVCTRGFCRCFCRRGVC
AP00859	LSPNLLKSLL
AP00861	FLPLAVSLAANFLPKLFCKITKKC
AP00862	FLPLLAGLAANFFPKIFCKITRKC
AP00864	FLPIVGRLISGLL
AP00865	FLPIIGQLLSGLL
AP00866	FLPIIAKVLSGLL
AP00867	FLPVIAGLLSKLF
AP00869	ILPLVGNLLNDLL
AP00870	AIMDTIKDTAKTVAVGLLNKLKCKITGC
AP00873	ILPILGNLLNGLL
AP00875	FLSSIGKILGNLL
AP00877	FLGSLIGAAIPAIKQLLGLKK
AP00885	FLPILASLAAKLGPKLFCLVTKKC
AP00886	FLPILASLAATLGPKLLCLITKKC
AP00888	GIMDSVKGLAKNLAGKLLDSLKCKITGC
AP00896	KRFKKFFKKLKKSVKKRAKKFFKKPRVIGVSIPF
AP00897	KRFKKFFKKLKNSVKKRAKKFFKKPKVIGVTFPF
AP00899	FLPIVTNLLSGLL
AP00911	FLSLIPHIVSGVAALAKHL
AP00913	QWGRRCCGWGPGRRYCVRWC
AP00914	QYGRRCCNWGPGRRYCKRWC
AP00916	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
AP00917	ACIKNGGRCVASGGPPYCCSNYCLQIAGQSYGVCKKH
AP00924	GYGCPFNQYQCHSHCRGIRGYKGGYCTGRFKQTCKCY
AP00925	GYGCPFNQYQCHSHCSGIRGYKGGYCKGLFKQTCNCY
AP00926	GFGCPFNQYECHAHCSGVPGYKGGYCKGLFKQTCNCY
AP00927	IYFIADKMGIQLAPAWYQDIVNWVSAGGTLTTGFAIIVGVTVPAWIAEAAAAFGIASA
AP00932	IYWIADQFGIHLATGTARKLLDAVASGASLGTAFAAILGVTLPAWALAAAGALGATAA
AP00938	GLVTSLIKGAGKLLGGLFGSVTG
AP00957	ALWKTMLKKLGTMALHAGKAAFGAAADTISQ
AP00958	ALWKTLLKNVGKAAGKAALNAVTDMVNQ
AP00961	ALWKNMLKGIGKLAGQAALGAVKTLVGAES
AP00966	GLWSKIKAAGKEAAKAAAKAAGKAALNAVSEAV
AP00969	GLWSKIKEAAKTAGLMAMGFVNDMV
AP00972	FLSLIPHAINAVGVHAKHF
AP00980	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNRKGVCVCRN
AP00990	RNCESLSHRFKGPCTRDSN
AP00997	ITSISLCTPGCKTGVLMGCNLKTATCNCSVHVSK
AP01000	GSGVIPTISHECHMNSFQFVFTCCS
AP01002	KSWSLCTPGCARTGSFNSYCC
AP01006	YITCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
AP01007	KWKLFKKIGAVLKVL
AP01013	SWASMAKKLKEYMEKLKQRA
AP01015	SLKDKVKSMGEKLKQYIQTWKAKF
AP01128	GSKILHSAGKFGKAFLGEINKS
AP01129	GLGSLVGNALRIGAKLL

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01130	GMASKAGSVLGKVAKVALKAAL
AP01146	DTLACRQSHGSCSFVACRAPSVDIGTCRGGKLKCCKWAPSS
AP01147	DTVACRIQGNFCRAGACPPTFTISGQCHGGLLNCCAKIPAQ
AP01152	SIWGDIGQGVGKAAYWVGKAMGNMSDVNQASRINRKKKH
AP01155	ESVFSKIGNAVGPAAYWILKGLGNMSDVNQADRINRKKH
AP01159	KWKIFKKIEHMGQNIRDGLIKAGPAVQVVGQAATIYKG
AP01170	YSSKDCLKDIGKGIGAGTVAGAAGGGLAAGLGAIPGAFVGAHFGVIGGSAACIGGLLGN
AP01179	NGVYCNKQKCWVDWSRARSEIIDRGVKAYVNGFTKVLGGIGGR
AP01181	AYPGNGVHCGKYSCTVDKQTAIGNIGNNAA
AP01190	MGAIAKLVAKFGWPIVKKYYKQIMQFIGEGWAINKIIEWIKKHI
AP01191	MGAIAKLVAKFGWPIVKKYYKQIMQFIGEGWAINKIIDWIKKHI
AP01194	CSTNTFSLSDYWGNNGAWCTLTHECMAWCK
AP01195	KRGSGWIATITDDCPNSVFVCC
AP01199	KYYGNGVHCGKKTCYVDWGQATASIGKIIVNGWTQHGPWAHR
AP01200	GGGVIQTISHECRMNSWQFLFTCCS
AP01201	KGGSGVIHTISHECNMNSWQFVFTCCS
AP01202	KGGSGVIHTISHEVIYNSWNFVFTCCS
AP01203	KKKSGVIPTVSHDCHMNSFQFVFTCCS
AP01210	PFKLSLHL
AP01212	EPFKLSLHL
AP01214	GNRPVYIPPPRPPHPRL
AP01216	GFRDVLKGAAKAFVKTVAGHIAN
AP01218	GFRDVLKGAAKAFVKTVAGHIANI
AP01220	GIKDWIKGAAKKLIKTVASNIANQ
AP01222	GFKDWIKGAAKKLIKTVASSIANQ
AP01224	VGALAVVVWLFLWLW
AP01225	VGALAVVVWLYLWLW
AP01228	ASGRDIAMAIGTLSGQFVAGGIGAAAGGVAGGAIYDYASTHKPNPAMSPSGLGGTIKQKPEGIPSEAW NYAAGRLCNWSPNNLSDVCL
AP01241	FASLLGKALKALAKQ
AP01236	GLLDFAKHVIGIASKL
AP01244	KGGPCAKKPCCGPLGHYKVDCSTIPDYPCCSKYGFCGSGPQYCG
AP01245	AVTCNTVVSSLAPCVPFFAGSAAQPTAACCNGVRSLNSAARTTPDRRTACNCIKSSASSIGLNYNKAA KLPSRCTVNVTVPISPSVNCAT
AP01246	FLPKTLRKFFCRIRGGRCAVLNCLGKEEQIGRCSNSGRKCCRKKK
AP01249	GILDAIKAIAKAAG
AP01250	GALRGCWTKSYPPKPCK
AP01251	SAPRGCWTKSYPPKPCK
AP01252	LVRGCWTKSYPPKPCFVR
AP01253	GLMSLFKGVLKTAGKHIFKNVGGSLLDQAKCKITGEC
AP01254	GLMSLFRGVLKTAGKHIFKNVGGSLLDQAKCKITGEC
AP01255	GLMSVTKGVLKTAGKHIFKNVGGSLLDQAKCKISGQC
AP01256	GLMSVLKGVLKTAGKHIFKNVGGSLLDQAKCKITGQC
AP01257	GLLSVLKGVLKTTGKHIFKNVGGSLLDQAKCKISGQC
AP01258	GLMDVFKGAAKNLLASALDKIRCKVTKC
AP01263	FLPLVTMLLGKLF
AP01265	HFLGTLVNLAKKIL
AP01266	AVDLAKIANKVLSSLF
AP01267	RRTCHCRSRCLRRESNSGSCNINGRIFSLCCR
AP01270	FVGLAKVAAHVVPAIAEHF

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01272	GIFSKINKKKAKTGLFNIIKTVGKEAGMDVIRAGIDTISCKIKGEC
AP01274	GLMDTVKNAAKNLAGQLLDTIKCKMTGC
AP01275	GILDTIKNAAKTVAVGLLEKIKCKMTGC
AP01276	GFISTVKNLATNVAGTVIDTIKCKVTGGC
AP01285	AGDPLADPNSQIVRQIMSNAAWGPPLVPERFRGMAVGAAGGVTQTVLQGAAAHMPVNVPIPKVPMG PSWNGSKG
AP01287	IRNSLTCRFNFGICLPKRCPGRMRQIGTCF
AP01288	RRLHPQHQRFPRERPWPKPLSLPLPRPGPRPWPKPL
AP01292	GVLGAVKDLLIGAGKSAAQSVLKTLSCKLSNDC
AP01293	GFLDTFKNLALNAAKSAGVSVLNSLSCKLFKTC
AP01295	GFMDTAKNVAKNVAVTLIDNLKCKITKAC
AP01297	GIFGKILGVGKKVLCGLSGWC
AP01298	GLFTLIKCAYQLIAPTVACN
AP01309	GLFGKSSVVGRKYYVDLAGCAKA
AP01311	GLLSGILGAGKQKVCGLSGLC
AP01312	GLLSGILGAGKHIVCGLSGLK
AP01314	GLLSGVLGVGKKVDCGLSGLC
AP01317	GAILCNLCKDTVKLVENLLTVDGAQAVRQYIDNLCGKASGFLGTLCEKILSFGVDELVKLIENHVDPV VVCEKIHAC
AP01318	IPVLCPVCTSLVGKLIDLVLGGAVDKVTDYLETLCAKADGLVETLCTKIVSYGIDKLIEKILEGGSAKLI CGLIHAC
AP01320	LFCRKGTCHFGGCPAHLVKVGSCFGFRACCKWPWDV
AP01322	IPRPLDPCIAQNGRCFTGICRYPYFWIGTCRNGKSCCRRR
AP01327	LFGLIPSLIGGLVSAFK
AP01345	FFGTALKIAANVLPTAICKILKKC
AP01346	FFPLVLGALGSILPKIF
AP01349	FFGTALKIAANILPTAICKILKKC
AP01351	GLWSTIKNVGKEAAIAAGKAALGALGEQ
AP01352	GLWSTIKNVGKEAAIAAGKAVLGSLGEQ
AP01358	VTCDLLSFEAKGFAANHSLCAAHCLAIGRRGGSCERGVCICRR
AP01359	ATCDLLSGFGVGDSACAAHCIARGNRGGYCNSKKVCVCRN
AP01360	ATCDLLSGFGVGDSACAAHCIARGNRGGYCNSQKVCVCRN
AP01361	ATCDLLSGFGVGDSACAAHCIARRNRGGYCNAKKVCVCRN
AP01362	ATCDLLSGFGVGDSACAAHCIARGNRGGYCNSKKVCVCPI
AP01363	ATCDLASGFGVGSSLCAAHCIARRYRGGYCNSKAVCVCRN
AP01366	ATCDLLSMWNVNHSACAAHCLLLGKSGGRCNDDAVCVCRK
AP01368	ATCDLFSFRSKWVTPNHAACAAHCLLRGNRGGRCKGTICHCRK
AP01373	AELRCMCIKTTSGIHPKNIQSLEVIGKGTHCNQVEVIATLKDGRKICLDPDAPRIKKIVQKKLAGDES
AP01389	GLWSTIKQKGKEAAIAAAKAAGQAALGAL
AP01390	ILPFVAGVAAEMMQHVYCAASKKC
AP01391	SFLDTLKNLAISAAKGAGQSVLSTLSCKLSKTC
AP01392	SVLGTVKDLLIGAGKSAAQSVLTALSCKLSNSC
AP01393	GVFTLIKGATQLIGKTLGKELGKTGLELMACKITNQC
AP01398	YQEPVLGPVRGPFPIIV
AP01399	YQEPVLGPVRGPFPI
AP01401	GVLDILKGAAKDLAGHVATKVINKI
AP01402	GVVDILKGAGKDLLAHALSKLSEKV
AP01403	GVLDILKGAGKDLLAHALSKISEKV
AP01404	GVLDILTGAGKDLLAHALSKLSEKV
AP01405	GLLGGLLGPLLGGGGGGGGLL
AP01408	KQEGRDHDKSKGHFHMIVIHHKGGQAHHG

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01409	GVVDILKGAAKDLAGHLATKVMNKL
AP01419	GIFSKFGGKAIKNLFIKGAKNIGKEVGMDVIRTGIDVAGCKIKGEC
AP01420	GIFSLIKGAAQLIGKTVAKEAGKTGLELMACKVTKQC
AP01421	GLLDSLKNLAINAAKGAGQSVLNTLSCKLSKTC
AP01423	FLPAVLRVAAKIVPTVFCAISKKC
AP01424	FLPAVLRVAAQVVPTVFCAISKKC
AP01425	GLLGSLFGAGKKVACALSGLC
AP01426	GLLGSIFGAGKKIACALSGLC
AP01427	GAIKDALKGAAKTVAVELLKKAQCKLEKTC
AP01428	GFKGAFKNVMFGIAKSAGKSALNALACKIDKSC
AP01429	GLLDSFKNAMIGIAKSAGKTALNKIACKIDKTC
AP01431	GFLDSFKNAMIGVAKSVGKTALSTLACKIDKSC
AP01432	FMGGLIKAATKIVPAAYCAITKKC
AP01435	GILSSFKGVAKGVAKNLAGKLLDELKCKITGC
AP01436	FLPILAGLAAKLVPKVFCSITKKC
AP01437	FLPILAGLAANILPKVFCSITKKC
AP01438	FFPIIAGMAAKLIPSLFCKITKKC
AP01440	FFPIIAGMAAKVICAITKKC
AP01445	FMGSALRIAAKVLPAALCQIFKKC
AP01446	FFGSVLKVAAKVLPAALCQIFKKC
AP01448	FLPIALKALGSIFPKIL
AP01450	FFGAIAAALPHVISAIKNAL
AP01451	FVGAIAAALPHVISAIKNAL
AP01452	IIGAIAAALPHVINAIKNTF
AP01453	LLSLALAALPKLFCLIFKKC
AP01458	GLKEIFKAGLGSLVKGIAAHVAN
AP01459	GLKEIFKAGLGSLVKGIAAHVAS
AP01461	ILGKLLSTAAGLLSNL
AP01469	AQRCGDQARGAKCPNCLCCGKYGFCGSGDAYCGAGSCQSQCRGC
AP01470	AQRCGDQARGAKCPNCLCCGKYGFCGSGDAYCGAGSCQSQCRGCR
AP01474	YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY
AP01493	ASIIKTTIKVSKAVCKTLTCICTGSCSNCK
AP01495	IASKFICTPGCAKTGSFNSYCC
AP01504	GFFSLIKGVAKIATKGLAKNLGKMGLDLVGCKISKEC
AP01518	AMVSS
AP01524	DIDFSTCARMDVPILKKAAQGLCITSCSMQNCGTGSCKKRSGRPTCVCYRCANGGGDIPLGAL
AP01526	SWLSKTAKKLENSAKKRISEGIAIAIKGGSR
AP01527	SWLSKTYKKLENSAKKRISEGVAIAILGGLR
AP01532	ATCDLLSGTGVKHSACAAHCLLRGNRGGYCNGRAICVCRN
AP01535	GILSGILGMGKKLVCGLSGLC
AP01536	GILSGILGAGKSLVCGLSGLC
AP01537	GILSGVLGMGKKIVCGLSGLC
AP01538	GILSNVLGMGKKIVCGLSGLC
AP01539	SILSGNFGVGKKIVCGLSGLC
AP01541	AVLDILKDVGKGLLSHFMEKV
AP01542	AVLDFIKAAGKGLVTNIMEKVG
AP01546	GMWSKIKNAGKAAKAAAKAAGKAALGAVSEAM
AP01547	GVIKSVLKGVAKTVALGML
AP01554	DFGCGQGMIFMCQRRCMRLYPGSTGFCRGFRCMCDTHIPLRPPFMVG
AP01558	LGAWLAGKVAGTVATYAWNRYV

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01564	ATCDLLSKWNWNHTACAGHCIAKGFKGGYCNDKAVCVCRN
AP01574	SWFSRTVHNVGNAVRKGIHAGQGVCSGLGL
AP01582	CIAKGNGCQPSGVQGNCCSGHCHKEPGWVAGYCK
AP01584	VGECVRGRCPSGMCCSQFGYCGKGPKYCG
AP01586	KSCCRSTTARNIYNGCRVPGTARPVCAKKSGCKIQEAKKCEPPYD
AP01587	AQCGAQGGGATCPGGLCCSQWGWCGSTPKYCGAGCQSNCK
AP01588	AQCGAQGGGATCPGGLCCSQWGWCGSTPKYCGAGCQSNCR
AP01590	DHYICAKKGGTCNFSPCPLFNRIEGTCYSGKAKCCIR
AP01594	CRQSCSFGPLTFVCDGNTK
AP01595	CANSCSYGPLTWSCDGNTK
AP01597	SVSCLRNKGVCMPGKCAPKMKQIGTCGMPQVKCCKRK
AP01598	IFNSIYHRKCVVKNRCETVSGHKTCKDLTCCRAVIFRHERPEVCRPST
AP01605	SSGWVCTLTIECGTVICAC
AP01611	IASKFLCTPGCAKTGSFNSYCC
AP01612	ASIVKTTIKASKKLCRGFTLTCGCHFTGKK
AP01617	VTSWSLCTPGCTSPGGGSNCSFCC
AP01620	VDKPPYLPRPPPPRRIYNNR
AP01621	CAWYNISCRLGNKGAYCTLTVECMPSCN
AP01626	LQDAAVGWGRRCPQCPRCPSCPSCPRCPRCPRCKCNPK
AP01627	LQDAALGWGRRCPRCPRCPRCSWCPRCPTCPRCNCNPK
AP01628	LQDAALGWGRRCPRCPPCPRCSWCPRCPTCPGCNCNPK
AP01629	LQDAALGWGRRCPRCPRCPNCRRCPRCPTCPSCNCNPK
AP01630	LQDAALGWSRRCPRCPPCPNCRRCPRCPTCPSCNCNPK
AP01631	LQDAALGWGRRCPRCPRCPNCKRCPRCPTCPRCNCNPK
AP01634	INWKKIFEKVKNLV
AP01635	INWLKLGKKILGAL
AP01636	INWLKLGKKMMSAL
AP01638	INWKKIAEVGGKILSSL
AP01639	INWKKIAEIGKQVLSAL
AP01640	IDWLKLGKMVIDAL
AP01641	IDWLKLGKMVMDVL
AP01643	RIKRFWPVVIRTVVAGYNLYRAIKKK
AP01645	GAFGNFLKGVAKKAGLKILSIAQCKLFGTC
AP01653	GLFSVLGSVAKHLLPHVVPVIAEKL
AP01654	GLFKVLGSVAKHLLPHVAPIIAEKL
AP01655	IIGHLIKTALGFLGL
AP01656	GIFSKLAGKKIKNLISGLKNIGKEVGMDVVRTGIDIAGCKIKGEC
AP01659	NALSSPRNKCDRASSCFG
AP01660	WNSNRRFRVGRPPVVGRPGCVCFRAPCPCSNY
AP01661	ACYCRIPACLAGERRYGTCFYLGRVWAFCC
AP01662	ACYCRIPACLAGERRYGTCFYRRRVWAFCC
AP01664	RTCRCRFGRCFRRESYSGSCNINGRIFSLCCR
AP01665	RRTCRCRFGRCFRRESYSGSCNINGRISSLCCR
AP01666	RTCRCRFGRCFRRESYSGSCNINGRISSLCCR
AP01667	RRICRCRIGRCLGLEVYFGVCFLHGRLARRCCR
AP01668	RICRCRIGRCLGLEVYFGVCFLHGRLARRCCR
AP01669	RTCRCRLGRCSRRESYSGSCNINGRIYSLCCR
AP01670	ACYCRIPACFAGERRYGTCFYLGRVWAFCC
AP01671	RCVCRRGVCRCVCTRGFC
AP01673	GICRCICTRGFCRCICVL

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01674	GICRCLCRRGVCRCICVL
AP01675	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
AP01694	KFCEKPSGTWSGVCGNSGACKDQCIRLEGAKHGSCNYKLPAHRCICYYEC
AP01695	GFGSLLGKALRLGANVL
AP01697	GLFSVLGSVAKHLLPHVAPIIAEKL
AP01698	GLWEKVKEKANELVSGIVEGVK
AP01702	MKTFSVAVAVAVVLAFICTQESSALPVTGIEELVEPVSSDNNDNHQGLPVELRERLVNIRKKRAPTDCI PYCYPTGDGFHCGVTCRF
AP01703	GIFSKFAGKGIKNLLVKGVKNIGKEVGMDVIRTGIDIAGCKIKGEC
AP01704	RIFSKIGGKAIKNLILKGIKNIGKEVGMDVIRTGIDVAGCKIKGEC
AP01705	GIFSLIKGAAKLITKTVAKEAGKTGLELMACKVTNQC
AP01706	GFMDTAKNVAKNVAVTLLDKLKCKITGGC
AP01707	SLLDTFKNLAVNAAKSAGVSVLNALSCKISRTC
AP01708	SFLTTFKDLAIKAAKSAGQSVLSTLSCKLSNTC
AP01709	SVLGTVKDLLIGAGKSAAQSVLTTLSCKLSNSC
AP01710	GIFSTVFKAGKGIVCGLTGLC
AP01711	GILGTVFKAGKGIVCGLTGLC
AP01712	WKSESVCTPGCVTGLLQTCFLQTITCNCKISK
AP01716	GVFSFLKTGAKLLGSTLLKMAGKAGAEHLACKATNQC
AP01717	GIFSALAAGVKLLGNTLFKMAGKAGAEHLACKATNQC
AP01719	GILDTFKGVAKGVAKDLAVHMLENLKCKMTGC
AP01720	RPRPNYRPRPIYRP
AP01722	FLPVLAGLTPSIVPKLVCLLTKKC
AP01723	FFPMLAGVAARVVPKVICLITKKC
AP01725	FLPILGNLLSGLL
AP01726	FLPIITNLLGKLL
AP01727	NFLDTLINLAKKFI
AP01728	GSQLVYREWVGHSNVIKP
AP01730	GIGTKILGGVKAALKGALKELASTYVN
AP01731	GIGTKFLGGVKTALKGALKELASTYVN
AP01732	GIGGALLSAGKSALKGLAKGLAEHF
AP01733	GIGGKILGGLKTALKGAAKELASTYLH
AP01734	GIGTKFLGGVKTALKGALKELAFTYVN
AP01735	GIGGALLSVGKSALKGLTKGLAEHF
AP01736	GIGGKILGGLRTALKGAAKELAATYLH
AP01738	GIGGVLLSAGKAALKGLTKVLAEKYAN
AP01739	GIGGVLLGAGKATLKGLAKVLAEKYAN
AP01740	GIGGALLSAGKAALKGLAKVLV
AP01741	SIGAKILGGVKTFFKGALKELAFTYLQ
AP01742	GIGGALLSAGKSALKGLAKGLAEHL
AP01743	GIGGALLSVGKLALKGLANVLADKFAN
AP01744	ILGPVIKTIGGVIGGLLKNL
AP01737	ILGPVLGLVGNALGGLIKKL
AP01748	SFHVFPPWMCKSLKKC
AP01751	RWKIFKKIEKVGRNVRDGIIKAGPAVAVVGQAATVVK
AP01755	GGYYCPFRQDKCHRHCRSFGRKAGYCGGFLKKTCICV
AP01759	RMRRSKSGKGSGGSKGSGSKGSKGSKGSGSKGSGSKGGSRPGGGSSIAGGGSKGKGGTQTA
AP01761	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGIKHHSSGSSSYHC
AP01767	LVNQLGISKSLANTILGAIAVGNLASWVLALVPGPGWATKAALATAETIVKHEGKAAAIAW
AP01768	QLGELIQQGGQKIVEKIQKIGQRIRDFFSNLRPRQEA

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01770	RPRCWIKIKFRCKSLKF
AP01771	ILPLLLGKVVCAITKKC
AP01779	GFMDTAKNVAKNVAVTLIDKLRCKVTGGC
AP01780	GFMDTAKQVAKNVAVTLIDKLRCKVTGGC
AP01781	GFMDTAKNVAKNVAATLLDKLKCKITGGC
AP01782	GILSTVFKAGKGIVCGLSGLC
AP01787	GRLQAFLAKMKEIAAQTL
AP01792	GWINEEKIQKKIDERMGNTVLGGMAKAIVHKMAKNEFQCMANMDMLGNCEKHCQTSGEKGYCHGT KCKCGTPLSY
AP01814	KIAKVALKAL
AP01820	VIVFVASVAAEMMQHVYCAASKKC
AP01822	FLPAVIRVAANVLPTAFCAISKKC
AP01823	LPFVAGVAAEMMQHVYCAASKKC
AP01825	IDPFVAGVAAEMMQHVYCAASKKC
AP01826	INPFVAGVAAEMMQHVYCAASKKC
AP01827	ILPFVAGVAAEMMKHVYCAASKKC
AP01828	IIPFVAGVAAEMMEHVYCAASKKC
AP01829	QLPFVAGVACEMCQCVYCAASKKC
AP01830	ILPFVAGVAAEMMEHVYCAASKKC
AP01831	ILPFVAGVAAMEMEHVYCAASKKC
AP01832	FLPAVLLVATHVLPTVFCAITRKC
AP01833	LAFVAGVAAEMMQHVYCAASKKC
AP01834	GILDTFKNMALNAAKSAGVSVLNALSCKLSKTC
AP01835	GLLDTFKNMALNAAKSAGVSVLNALSCKLSKTC
AP01836	GLLDTFKNMAINAAHGAGVSVLNALSCKLKKTC
AP01837	GLLDTFKNLAINAAESAGVSVLNSLSCKLSKTC
AP01838	GLLDGILNANFNAAKSAGTSVLNALSCKLSKTC
AP01839	GVLATVKNLLIGTGDGAAQSVLKTLSCKLSNDC
AP01840	GVLGTVKDLLIGAGKSAAQSTLKTLSCKISNDC
AP01841	GVLATVKNLLNGTGDGAAQSVLKTLSCKLSNDC
AP01843	GLLDTIKNMALNAAKSAGVSVLNSLSCKLSKTC
AP01844	GLIDTIKNMALNAAKSAGVSVLNTLSCKLSKTC
AP01845	SVLGTVKDLLIGAGKSAAQSVLTANSCKLSNSC
AP01846	SFLDTLKNLAISAAKGAGQSVLSTLSCKLSETC
AP01847	GLLDTIKNMALNAAKSAGVSVLNSLSCKDSKTC
AP01848	GLLDTIKNMALNAAKSAGVSVLNTLSCKLSKTC
AP01851	SFLSTFKELAINAAKNAGQSLLHTLSCKLDKTC
AP01853	SVMGTVKDLLIGAGKSAAQSVLKSLSCKLSNDC
AP01854	SVMGTVKDLLIGAGKSAAQSVLKALSCKLSKDC
AP01856	GLFSKFSGKGIKNFLIKGVKHIGKEVGMDVIRTGIDVAGCKIKGEC
AP01857	GLFSKFAGKGIKDLIFKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01858	GLFTKFAGKGIKDLIFKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01859	GLFSKFAGKGIKNFLIKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01861	GIFSKISGKAIKNLFIKGAKNVGKEVGMDVVRTGIDVVGCKIKGEC
AP01862	AVNIPFKVHFRCKAAFC
AP01863	GIFSKISGKAIKNLFIKGAKNVGKRVGMDVVRTGMDVVGCKIKGEC
AP01864	GIFSKISGKAIKNLFIKGAENVGKHVGIDVVRTGIDVVGCKIKGEC
AP01865	GIFSKISGKAIKNLFIKGAKNVGKEVGIDVVRTGMDVVGCKIKGEC
AP01866	GIFTLIKGAAKLIGKTVAKEAGKTGLELMACKITNQC
AP01867	GLFTLIKGAAKLIGKTTAKEAGKTGKLEMACKITNQC

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01868	GFMNTAKNVAKNVAVTLLDNLKCKITGGC
AP01869	GLFTLIKGAYKNDAPTVACN
AP01870	GVFTLIKGATQLIGKTLGKEVGKTGLELMACKITKQC
AP01871	GVFTLLKGATQLIGKTLGKELGKTGLELMACKITNQC
AP01872	GVFTLIKGATQLIGKTLGKELGKTGLELMACKITEQC
AP01873	GVFTLIKGATQLIGKTLGKELGKTGLEIMACKITKQC
AP01875	GLLSGVLGVGKKIVCGLSGLC
AP01876	GIFGKILGVGKKTLCELSGMC
AP01878	GLISGILGVGKKLVCGLSGLC
AP01879	GLISGILGVGKMLVCGLSGLC
AP01880	GLISGLLGVGKMLVCGLSGLC
AP01881	GLFTLIKGAYKLDAPTVACN
AP01882	GILGNIVGMGKKIVCGLSGLC
AP01883	GILGNIVGMGKKVVCGLSGLC
AP01884	GILSGVLGMGKKIVCGLRGLC
AP01885	GILGNIVGMGKQVVCGLSGLC
AP01887	GFMDTAKNVAKNMAGNLLDNLKCKITKAC
AP01888	GFMDTAKNVAKNMAVTLLDNLKCKITKAC
AP01889	GFMDTAKNVAKNEAGNLLDNLKCKITKAC
AP01890	GFMATAKNVAKNMDVTLLDNLKCKITKAC
AP01894	FLPLLAGVVANFLPQIICKIARKC
AP01896	GLMSTLKDFGKTAAKEIAQSLLSTASCKLAKTC
AP01897	GILDTLKEFGKTAAKGIAQSLLSTASCKLAKTC
AP01898	RRSRRGRGGGRRGGSGGRGGRGGGGRSGAGSSIAGVGSRGGGGGRHYA
AP01901	GLASFLGKALKAGLKIGSHLLGGAPQQ
AP01902	GFGSFLGKALKAALKIGANVLGGAPQQ
AP01903	GFGSFLGKALKAALKIGANVLGGAPEQ
AP01904	GFGSFLGKALKAALKIGADVLGGAPQQ
AP01905	GFGSLLGKALKIGTNLL
AP01906	GIGSLLAKAAKLGANLL
AP01907	GIGSALAKAAKLVAGIV
AP00268	NRLSCHRNKGVCVPSRCPRHMRQIGTCRGPPVKCCRKK
AP00490	AALKGCWTKSIPPKPCSGKR
AP00491	AALRGCWTKSIPPKPCSGKR
AP01908	AALRGCWTKSIPPKPCPGKR
AP01909	SALVGCWTKSYPPKPCFGR
AP01910	SALVGCGTKSYPPKPCFGR
AP01911	SALVGCWTKSYPPNPCFGRG
AP01912	SALVGCWTKSWPPKPCFGRG
AP01913	SALVGCWTKSWPPKPCFGR
AP01914	AAFRGCWTKNYSPKPCL
AP01916	GTRCGETCFVLPCWSAKFGCYCQKGFCYRN
AP01917	GLWDSIKNFGKTIALNVMDKIKCKIGGGCPP
AP01918	IGVIKLSLCEEERNADEEKRRDDPDEMDVEVEKR
AP01920	FTSKKSMLLFFFLGTISLSLCQ
AP01921	ILPIIGKILSTIF
AP01924	FLPVILPVIGKLLNGIL
AP01925	GLLDAIKDTAQNLFANVLDKIKCKFTKC
AP01926	VIGSILGALASGLPTLISWIKNR
AP01927	GLFNVFKKVGKNVLKNVAGSLMDNLKCKVSGEC

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01928	GLFSAFKKVGKNVLKNVAGSLMDNLKCKVSGEC
AP01929	GIFALIKTAAKFVGKNLLKQAGKAGLEHLACKANNQC
AP01930	GIFSLIKTAAKFVGKNLLKQAGKAGVEHLACKANNQC
AP01931	FLPIAGKLLSGLSGLL
AP01932	FFPIVGKLLFGLSGLL
AP01933	FFPIVGKLLSGLSGLL
AP01934	FFPIVGKLLFGLFGLL
AP01935	FLPIVGKLLSGLSGLS
AP01936	FFPIVGKLLSGLL
AP01937	FFPIVGKLLFGLL
AP01938	FFPIVGKLLS
AP01942	DCTRWIIGINGRICRD
AP01945	FIGPVLKIAAGILPTAICKIFKKC
AP01946	FVGPVLKIAAGILPTAICKIYKKC
AP01947	FLGPIIKIATGILPTAICKFLKKC
AP01948	SIRDKIKTIAIDLAKSAGTGVLKTLICKLDKSC
AP01949	SIRDKIKTIAIDLAKSAGMGILKTLICKLDKSC
AP01950	SIRDKIKTIAIDLAKSAGTGVLKTLICKLNKSC
AP01951	FLPLVLGALSGILPKIL
AP01954	FALGAVTKRLPSLFCLITRKC
AP01960	NILNTIINLAKKIL
AP01961	FLPLIASLAANFVPKIFCKITKKC
AP01962	FLPLIASVAANLVPKIFCKITKKC
AP01966	IKIPAVVKDTLKKVAKGVLSAVAGALTQ
AP01967	IKIPAFVKDTLKKVAKGVISAVAGALTQ
AP01968	IKIPPIVKDTLKKVAKGVLSTIAGALST
AP01973	MLCKLSMFGAVLGVPACAIDCLPMGKTGGSCEGGVCGCRKLTFKILWDKKFG
AP01980	IFGAIWNGIKSLF
AP01998	GFLGPLLKLGLKGVAKVIPHLIPSRQQ
AP01999	GFLGPLLKLGLKGAAKLLPQLLPSRQQ
AP02000	GFLGSLLKTGLKVGSNLL
AP02002	GLFLDTLKKFAKAGMEAVINPK
AP02003	GFWTTAAEGLKKFAKAGLASILNPK
AP02004	GVWTTILGGLKKFAKGGLEALTNPK
AP02005	GLMSSIGKALGGLIVDVLKPKTPAS
AP02006	GLLDALSGILGL
AP02008	GLVSSIGKVLGGLLADVVKSKGQPA
AP02009	GLVSSIGKALGGLLVDVVKSKGQPA
AP02010	GLFGILGSVAKHVLPHVIPVVAEHL
AP02015	NANSNFEGGPRNDRSSGARNSLGRNAPTHIYSDPSTVKCANAVFSGMIGGAIKGGPIGMARGTIGGAV VGQCLSDHGSGNGSGNRGSSSSCSGNNVGGTCNR
AP02016	GKIPVKAIKQAGKVIGKGLRAINIAGTTHDVVSFFRPKKKKH
AP02019	ILGAIIPLVSGLLSHL
AP02020	FLSTLLKVAFKVVPTLFCPITKKC
AP02022	FLPLFLPKIICVITKKC
AP02025	DCYEDWSRCTPGTSFLTGILWKDCHSRCKELGHRGGRCVDSPSKHCPGVLKNNKQCHCY
AP02026	GFWGKLWEGVKSAI
AP02031	KPFKKLEKVGRNIRDGIIKAGPAVAVIGQATSIARPTGK
AP02032	KPFKKLEKVGRNIRNGIIRYNGPAVAVIGQA
AP02038	FLGLIFHGLVHAGKLIHGLIHRNRG

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP00192	ICIFCCGCCHRSKCGMCCKT
AP00436	LCNERPSQTWSGNCGNTAHCDKQCQDWEKASHGACHKRENHWKCFCYFNC
AP00287	QKLCQRPSGTWSGVCGNNNACKNQCIRLEKARHGSCNYVFPAHKCICYFPC
AP00337	PDPAKTAPKKKSKKAVT
AP00483	KTCEHLADTYRGVCFTNASCDDHCKNKAHLISGTCHNWKCFCTQNC
AP00751	DTLIGSCVWGATNYTSDCNAECKRRGYKGGHCGSFLNVNCWCE
AP00798	DSHEKRHHGYRRKFHEKHHSHREFPFYGDYGSNYLYDN
AP00799	RKFHEKHHSHREFPFYGDYGSNYLYDN
AP00800	RKFHEKHHSHRGYRSNYLYDN
AP00801	DSHAKRHHGYKRKFHEKHHSHRGYR
AP00802	RKFHEKHHSHRGY
AP00803	RKFHEKHHSHRGYR
AP00746	CIKNGNGCQPDGSQGNCCSRYCHKEPGWVAGYCR
AP00747	CIANRNGCQPDGSQGNCCSGYCHKEPGWVAGYCR
AP00813	CIKNGNGCQPNGSQNGCCSGYCHKQPGWVAGYCRRK
AP00912	AGECVQGRCPSGMCCSQFGYCGRGPKYCGR
AP00918	ELCEKASKTWSGNCGNTGHCDNQCKSWEGAAHGACHVRNGKHMCFCYFNC
AP00919	KTCENLVDTYRGPCFTTGSCDDHCKNKEHLLSGRCRDDVRCWCTRNC
AP00920	NTCENLAGSYKGVCFGGCDRHCRTQEGAISGRCRDDFRCWCTKNC
AP00921	RVCMKGSAGFKGLCMRDQNCAQVCLQEGWGGGNCDGVMRQCKCIRQC
AP00922	RECKTESNTFPGICITKPPCRKACISEKFTDGHCSKLLRRCLCTKPC
AP00978	RTCENLADKYRGPCFSGCDTHCTTKENAVSGRCRDDFRCWCTKRC
AP00979	RECKTESNTFPGICITKPPCRKACISEKFTDGHCSKILRRCLCTKPC
AP00981	ATCKAECPTWDSVCINKKPCVACCKKAKFSDGHCSKILRRCLCTKEC
AP00984	QICKAPSQTFPGLCFMDSSCRKYCIKEKFTGGHCSKLQRKCLCTKPC
AP00985	KDCKTESNTFPGICITKPPCRKACIKEKFTDGHCSKILRRCLCTKPC
AP00986	KSTCKAESNTFPGLCITKPPCRKACLSEKFTDGKCSKILRRCICYKPC
AP01052	QQCGRQASGRLCGNRLCCSQWGYCGSTASYCGAGCQSQCR
AP01163	DKLIGSCVWGAVNYTSNCNAECKRRGYKGGHCGSFANVNCW
AP01164	ETCASRCPRPCNAGLCCSIYGYCGSGNAYCGAGNCRCQCRG
AP01310	GLLSGILGAGKHIICGLSGLC
AP01330	RTCESQSHRFKGTCVRQSNCAAVCQTEGFHGGNCRGFRRRCFCTKHC
AP01410	GLLSSFKGVAKGVAKDLAGKLLEKLKCKITGC
AP01411	GIMDSVKGVAKNLAAKLLEKLKCKITGC
AP01508	ELCEKASQTWSGTCGKTKHCDDQCKSWEGAAHGACHVRDGKHMCFCYFNC
AP01551	DCLSGKYKGPCAVWDNEMCRRICKEEGHISGHCSPSLKCWCEGC
AP01553	RMCKTPSGKFKGYCVNNTNCKNVCRTEGFPTGSCDFHVAGRKCYCYKPCP
AP01561	SKYGGECSVEHNTCTYLKGGKDHIVSCPSAANLRCKTERHHCEYDEHHKTVDCQTPV
AP01563	GWLRKLGKKIERIGQHTRDASIQVLGIAQQAANVAATAR
AP01567	DSHEERRQGRHGHHEYGRKFHEKHHSHRGY
AP01568	DSHEKRHHEHRRKFHEKHHSHRGY
AP01677	KTCENLADTYKGPCFTTGSCD
AP01684	KLCERSSGTWSGVCGNNNACKNQCIRLEGAQHGSCNYVFPAHKCICYFPC
AP01685	QKLCERSSGTWSGVCGNNNACKNQCINLEGARHGSCNYIFPYHRCICYFPC
AP01686	KTCENLSGTFKGPCIPDGNCNKHCRNNEHLLSGRCRDDFRCWCTNRC
AP01687	KLCERSSGTWSGVCGNNNACKNQCINLEGARHGSCNYVFPYHRCICYFPC
AP01689	KLCERPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
AP01690	KLCQRPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
AP01692	KLCERPSGTWSGVCGNSNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
AP01772	KICERASGTWKGICIHSNDCNNQCVKWENAGSGSCHYQFPNYMCFCYFDC

Table A.7: Xiao et al. (2013) Data Set Testing AMP Sequences Continued...

Definition	Sequence
AP01773	KICERASGTWKGICIHSNDCNNQCVKWENAGSGSCHYQFPNYMCFCYFNC
AP01816	RYCERSSGTWSGVCGNSGKCSNQCQRLEGAAHGSCNYVFPAHKCICYYPC
AP01817	QKLCERPSGTWSGVCGNNNACRNQCINLEKARHGSCNYVFPAHKCICYFPC
AP01818	RYCERSSGTWSGVCGNTDKCSSQCQRLEGAAHGSCNYVFPAHKCICYYPC
AP01819	QKLCERPSGTWSGVCGNNGACRNQCIRLERARHGSCNYVFPAHKCICYFPC
AP02023	RYCLSQSHRFKGLCMSSSNCANVCQTENFPGGECKADGATRKCFCKKIC
AP02035	DPQTECQQCQRRCRQQESGPRQQQYCQRRCKEICEEEEEYN
AP00989	RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGDCKGMTRTCYCLVNC
AP01035	GIPCAESCVYIPCTVTALLGCSCSNRVCYN
AP01915	GGSVPCGESCVFIPCITSLAGCSCKNKVCYYD
AP01037	GIPCGESCVWIPCLTSAIGCSCKSKVCYRN
AP01042	GTLPCESCVWIPCISSVVGCSCKSKVCYKN
AP01047	TPCGESCVYIPCISGVIGCSCTDKVCYLN
AP01053	GLPVCGETCFGGTCNTPGCSCSSWPICTRN
AP01054	GLPVCGETCTLGTCYTQGCTCSWPICKRN
AP01067	GLVPCGETCFTGKCYTPGCSCSYPICKKN
AP01068	GLPCGETTCFTGKCYTPGCSCSYPICKKIN
AP01464	VNPSYRLDPESRPQCEAHCGQLGMRLGAIVIMGTATGCVCEPKEAATPESR
AP01699	GRGREFMSNLKEKLSGVKEKMKNS
AP01700	VKLIQIRIWIQYVTVLQMFSMKTKQ

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences

Definition	Sequence
Q01468	MPIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASKVRR
D3Z9R8	MLQSFIKKVWVPMKPYYTQVYQEIWVGVGLMSLIVYKIRSADKRSKALKGCSPAHAHGHH
Q89GR1	MHNTAINVQNRVLSVVRSVLQQNAISADVHPESRLVDIGLSSMGMVELMLKVEAEFDLILPQFEITPEN FRSVKAMERMILNQLGSGSG
P39758	MKSIGVVRKVDELGRIVMPIELRRALDIAIKDSIEFFVDGDKIILKKYKPHGVCLMTGEITSENKEYGNG KITLSPEGAQLLLEEIQAALKE
P0C2W9	MGNLTSTCLFSSRENTAAKINDSSTWYPQQGQHISIQTFRELNRLPTSRRTSTKTEILLYGENSRSTVEV LEEVAKHLTTLQQRR
Q8N6N7	MALQADFDRAAEDVRKLKARPDDGELKELYGLYKQAIVGDINIACPGMLDLKGKAKWEAWNLKKGL STEDATSAYISKAKELIEKYGI
P61867	MVSQLFEEKAKAVNELPTKPSTDELLELYGLYKQATVGDNDKEKPGIFNMKDRYKWEAWEDLKGKS QEDAEKEYIAYVDNLIAKYSS
Q5FXM5	MTTFEEAAQKVKEFTKKPSNDELLSLYGLYKQGTDGDCNISEPWAVQVEAKAKYNAWNALKGTSKE DAKAKYVALYEQLATKYA
P31786	MSQAEFDKAAEEVKRLKTQPTDEEMLFIYSHFKQATVGDVNTDRPGLLDLKGKAKWDSWNKLKGTS KESAMKTYVEKVDELKKKYGI
P31787	MVSQLFEEKAKAVNELPTKPSTDELLELYALYKQATVGDNDKEKPGIFNMKDRYKWEAWENLKGKS QEDAEKEYIALVDQLIAKYSS
P94123	MSDVAERVKKIVVDHLGVEESKVTENASFIDDLGADSLDTVELVMAFEEEFGCEIPDDAAEKILTVKD AIDFIKANAAA
P0A6A8	MSTIEERVKKIIGEQLGVKQEEVTNNASFVEDLGADSLDTVELVMALEEEFDTEIPDEEAEKITTVQAA IDYINGHQA
P80922	MSTIEERVKKIVSEQLGVKEEEITNASSFVDDLGADSLDTVELVMALEEEFETEIPDEEAEKITTVQEAI DYVVSHQ
P80923	MSTIEERVKKIVAEQLGVKSEEVVNTASFVEDLGADSLDTVELVMALEEEFETEIPDEEAEKITTVQAA IDYVNSHQA
P19372	MSDIAERVKKIVIDHLGVDAEKVSEGASFIDDLGADSLDTVELVMAFEEEFGVEIPDDAADSILTVGDA VKFIEKAQA

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P12784	MSDIADRVKKIVVEHLGVEEEKVTETTSFIDDLGADSLDTVELVMAFEEEFGIEIPDDAAETIQTFGDA P
P07311	MAEGNTLISVDYEIFGKVQGVFFRKHTQAEGKKLGLVGWVQNTDRGTVQGQLQGPISKVRHMQEWL ETRGSPKSHIDKANFNNEKVILKLDYSDFQIVK
P35744	MSAAAQLKSVDYEVFGRVQGVCFRMYTEGEAKKIGVVGWVKNTSKGTVTGQVQGPEEKVNSMKSW LSKVGSPSSRIDRTNFSNEKSISKLEYSNFSIRY
P00818	MSTARPLKSVDYEVFGRVQGVCFRMYAEDEARKIGVVGWVKNTSKGTVTGQVQGPEEKVNSMKSW LSKVGSPSSRIDRTNFSNEKTISKLEYSNFSVRY
P14621	MSTAQSLKSVDYEVFGRVQGVCFRMYTEDEARKIGVVGWVKNTSKGTVTGQVQGPEDKVNSMKSW LSKVGSPSSRIDRTNFSNEKTISKLEYSNFSIRY
P00819	MSTARPLKSVDYEVFGRVQGVCFRMYTEDEARKIGVVGWVKNTSKGTVTGQVQGPEEKVNSMKSW LSKIGSPSSRIDRTNFSNEKTISKLEYSNFSIRY
P00820	MSTAGPLKSVDYEVFGRVQGVCFRMYTEGEAKKIGVVGWVKNTSKGTVTGQVQGPEDKVNSMKSW LSKVGSPSSRIDRTNFSNEKTISKLEYSNFSIRY
P35745	MAEPLKSVDYEVFGTVQGVCFRMYTEGEAKKRGLVGWVKNTSKGTVTGQVQGPEEKVNSMKSWLS KVGSPSSRIDRADFSNEKTISKLEYSNFSIRY
P84142	MAIVRAHLKIYGRVQGVGFRWSMQREARKLGVNGWVRNLPDGSVEAVLEGDEERVEALIGWAHQGP PLARVTRVEVKWEQPKGEKGFRIVG
Q5SKS6	MPRLVALVKGRVQGVGYRAFAQKKALELGLSGYAENLPDGRVEVVAEGPKEALELFLHHLKQGPRL ARVEAVEVQWGEEAGLKGFHVY
P55319	MVQRCALVVLLVVAVAAALCSAQLNFTPNWGTGKRDAADFADPYSFLYRLIQAEARKMSGCSN
P08379	MTQSCTLTLVLVVAVLAALATAQLNFSAGWGRRYADPNADPMAFLYRLIQIEARKLAGCSD
P35807	MRQGCALTLMLLVVVCAALSAAQLNFSTGWGRRYADPNADPMAFLYKLIQIEARKLAGCSN
P67787	QLTFTSSWG
P14595	QLTFTPGW
O28323	MAEHVVYVGNKPVMNYVLATLTQLNEGADEVVIKARGRAISRAVDVAEIVRNRFMPGVKVKEIKIDT EELESEQGRRSNVSTIEIVLAK
O27527	MSEENVVYIGNKPVMNYVLAVVTQMNGGTSEVILKARGIAISRAVDVAEIVRNRFIPDIQIENIDICTEEI IGNEGTATNVSAIEIQLRKD
P85421	IWGIGCNP
P19614	NQASVVANQLIPINTALTLVMMRSEVVTPVGIPAEDIPRLVSMQVNRAVPLGTTLMPDMVKGYPPA
P12101	TKSVVANQLIPINTALTLVMMKAEEVSPKGIPAEEIPRLVGMQVNRAVYLDETLMPDMVKNYE
Q9R0R3	MNLSFCVQALLLLWLSLTAVCGVPLMLPPDGKGLEEGNMRYLVKPRTSRTGPGAWQGGRRKFRRQR PRLSHKGPMPF
P83704	QAEESCLQNLASRYLQTVTDYGKDLVEKALAPELQAQAKAYFEKTQEQLTPLVKKIGNDLLNFFSHFI ELKTQPAT
P18656	MKLLAATVLLLTICSLEGALVRRQAEEPSVESLVSQYFQTVTDYGKDLMEKVKSPELQAQAKAYFEKS KEQLTPLVKKAGTDLVNFLSYFVELRTQPATQ
P02653	QAEEPSVESLVSQYFQTVTDYGKDLMEKVKSPELQAQAKAYFEKSKEQLTPLVKKAGTDLVNFLSYF VELRTQPATQ
Q8MIQ5	MKLLAATVLLLTICSLEGALVRRQAKEPCVDNLVSQYFQTVTDYGKDLMEKVKSPELQAEAKSYFEK SKEQLTPLIKKAGTELVNFLSYFMELGTQPATQ
P19035	MQPRLLLLAAFLALLVLPEATKAEEGSLLDKMQGYVKEATKTAKDALSSVQESQVAQQARDWMTES FSSLKDYWSSFKGKFTDFWESATSPTQSPP
P18659	MQPRVLLVAALLSLLASARASEAEDTSLLGFMQGYMQHATKTAKDALTSVQESQVAQQARGWVTDG FSSLKDYWSTVKDKLSGFWDLNPEAKPTLAEAA
P27917	MQPRVLLVAGLLVLLACAQAIEAEDTSLLDKMQDYVKQATRTAQDALTSVKESEVAQQARGWVTDSI SSLKDYWSTFKGKFTDFWDYTPKPEPSSS
P02658	KSIFERDRRDWLVIPDAIAAYIYETVNKMSPRVGQFLADAAQTPVVVGTRTFLIRETSKLTLLAEQLM EKIKNLWYTKVLGY
P02660	KSIFERDRRDWLVIPDAVAAYIYEAVNKMSPRAGQFLVDISQTTVVSGTRNFLIRETARLTILAEQLME KIKNLWYTKVQGY
Q9WUC4	MPKHEFSVDMTCGGCAEAVSRVLNKLGGVEFNIDLPNKKVCIESEHSSDILLATLNKTGKAVSYLGPK
Q06450	MASNAAVPFWRAAGMTYITYSNLCANMVRNCLKEPYRAEALSREKVHFSFSKWVDGKPQKPAIRSD TGEE
P56385	MVPPVQVSPLIKLGRYSALFLGVAYGATRYNYLKPRAEEERRIAAEEKKKQDELKRIARELAEDDSILK
P03929	MPQLDTSTWLTMILSMFLTLFIIFQLKVSKHNFYHNPELTPTKMLKQNTPWETKWTKIYLPLLLPL
P03930	MPQLDTSTWFITIISSMITLFILFQLKVSSQTFPLAPSPKSLTTMKVKTPWELKWTKIYLPHSLPQQ

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P11608	MPQLDTSTWFITIISSMATLFILFQLKISSQTFPAPPSPKTMATEKTNNPWESKWTKIYLPLSLPPQ
Q28851	MASVVPLKEKKLLEVKLGELPSWILMRDFTPSGIAGAFQRGYYRYYNKYVNVKKGSIAGLSMVLAAY VFLNYCRSYKELKHERLRKYH
P56135	MASLVPLKEKKLMEVKLGELPSWIMMRDFTPSGIAGAFRRGYDRYYNKYINVRKGSISGISMVLAAYV VFSYCISYKELKHERRRKYH
D3ZAF6	MASIVPLKEKKLMEVKLRELPSWILMRDFTPSGIAGAFRRGYDRYYNKYINVRKGSISGINMVLAAYV VFSYCISYKELKHERRRKYH
P00845	MSLGVLAAAIAVGLGALGAGIGNGLIVSRTIEGIARQPELRPVLQTTMFIGVALVEALPIIGVVFSFIYLG R
Q03565	MSTSVKHREFVGEPMGDKEVTCIAGIGPTYGTKLTDAGFDKAYVLFGQYLLLKKDEDLFIEWLKETA GVTANHAKTAFNCLNEWADQFM
Q9H503	MDNMSPRLRAFLSEPIGEKDVCWVDGISHELAINLVTKGINKAYILLGQFLLMHKNEAEFQRWLICCFG ATECEAQQTSHCLKEWCACFL
Q6P026	MSSTSQKHKDFVAEPMGEKSVMALAGIGEVLGKRLEEKGFDKAYVVLGQFLVLRKDEELFREWLKD TCGANTKQQGDCYSCLREWCDSFL
O75531	MTTSQKHRDFVAEPMGEKPVGSLAGIGEVLGKKLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDT CGANAKQSRDCFGCLREWCDAFL
O54962	MTTSQKHRDFVAEPMGEKPVGSLAGIGDVLSKRLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDT CGANAKQSRDCFGCLREWCDAFL
Q9R1T1	MTTSQKHRDFVAEPMGEKPVGSLAGIGDALGKRLEERGFDKAYVVLGQFLVLKKDEDLFREWLKDT CGANAKQSRDCFGCLREWCDAFL
Q8BGS2	MELSADYLREKLRQDLEAEHVEVEDTTLNRCATSFRVLVVSAKFEGKPLLQRHRLVNECLAEELPHIH AFEQKTLTPEQWTRQRRE
P86483	MSLLPAVKVLPLGYLGIVLVFSLILRSAMVDFIQDAGKLERIDTYKREAQMIFGAPMWALGHLMGRK
P85160	QDGPIPP
P15411	MKILLAIALMLSTVMWVSTQQPQEVHTYCGRHLARTMADLCWEEGVDKRSDAQFASYGSAWLMPYS AGRGIVDECCLRPCSVDVLLSYC
P26729	MKILLAIALMLSTVMWVSTQQPQAVHTYCGRHLARTLADLCWEAGVDKRSGAQFASYGSAWLMPYS EGRGKRGIVDECCLRPCSVDVLLSYC
P26730	MKLLLAIALMLTIVMWVSTQQPQAVHTYCGRHLARTLADLCWEAGVDKRSDAQYASYGSAWLMPYS EGRGKRGIVDECCLRPCSVDVLLSYC
P26732	MKLLLAIALMLTTVMWASTQQPQAVHTYCGRHLARTLADLCWEAGVDKRSDAQFASYGSAWLMPY SEGRDQRGIVDECCLRPCSVDVLLSYC
Q9NRR3	MSEFWLCFNCCIAEQPQPKRRRRIDRSMIGEPTNFVHTAHVGSGDLFSGMNSVSSIQNQMQSKGGYGG GMPANVQMQLVDTKAG
Q8BGH7	MSEFWLCFNCCIAEQPQPKRRRRIDRSMIGEPTNFVHTAHVGSGDLFSGMNSVSSIQNQMQSKGGYGG GMPANVQMQLVDTKAG
C6ZJQ2	MGMRMMFAVFLLVVLATTVVSFNSDRASDGRNAAANVKASDLMARVLEKDCPPHPVPGMHKCVCL KTCR
P56639	MFTVFLLVVLATTVVSFTSDRASDGRKDAASGLIALTIKGCCSYPPCFATNSDYCG
P69657	MFTVFLLVVLTTTVVSFPSDRASDGRNAAANDKASDVVTLVLKGCCSTPPCAVLYCGRRR
P58811	MFTVFLLVVLATTGVSFTLDRASDGGNAVAKKSDVTARFNWRCCLIPACRRNHKKFCG
P56640	MFTVFLLVVLATTVVSFTSDRASDGRKDAASGLIALTMKGCCSYPPCFATNPDCGRRR
P56638	MFTVFLLVVLATTVVSFTSDRASDSRKDAASGLIALTIKGCCSDPRCNMNNPDYCG
P56636	MGMRMMFTVFLLVVLATTVVSFPSDRASDGRNAAANDKASDVITLALKGCCSNPVCHLEHSNLCGRR R
P0C1X1	MGMRMMFTVFLLVVLATTVVSIPSDRASDGRNAVVHERAPELVVTATTNCCGYNPMTICPPCMCTY SCPPKRKPGRRND
P86506	QQDYTGWMDF
P62540	QQDYTGWFDF
P86507	QQDYTGWFDF
P62541	QQDYTGWFDF
P86486	QQDYTGWMDF
P56264	QQDYTGWMDF
P69540	MKKSLVLKASVAVATLVPMLSFAAEGDDPAKAAFDSLQASATEYIGYAWAMVVVIVGATIGIKLFKKF TSKAS
P69539	MKKSLVLKASVAVATLVPMLSFAAEGDDPAKAAFDSLQASATEYIGYAWAMVVVIVGATIGIKLFKKF TSKAS

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P69541	MKKSLVLKASVAVATLVPMLSFAAEGDDPAKAAFNSLQASATEYIGYAWAMVVVIVGATIGIKLFKKF TSKAS
P10420	AMPMLRL
Q97Y85	MKLLVVYDVSDDSKRNKLANNLKKLGLERIQRSAFEGDIDSQRVKDLVRVVKLIVDTNTDIVHIIPLGIR DWERRIVIGREGLEEWLV
Q8U1T8	MYIVVVYDVGVERVNKVKKFLRMHLNWVQNSVFEGEVTLAEFERIKEGLKKIIDENSDSVIIYKLRSM PPRETLGIEKNPIEEII
P80573	MLYLVRMDVNLPHDMPAAQADDIKAREKAYAQQLQHEGKWQQLYRVVGEYANYSIFDVGSHDELHT LLSGLPLFPYMKIHVTPLAKHPSSIR
P95609	MALFHVRMDVAIPRDLDPKVRDETIAKEKAYSQELQRSGKWPEIWRIVGQYSNISIFDVESADELHEIL WNLPLFPYMNIEIMPLTKHGSDVK
P20104	LVTLVFV
P24807	MGRAMVARLGLGLLLLALLLPTQIYCNQTSVAPFPGNQNISASPNPSNATTRGGGSSLQSTAGLLALSL SLLHLYC
C3VVN5	MPKLAVVLLVLLILPLSYFDAAGGQVVQGDRRGNGLARYLQRGDRDVRECQVNTPGSSWGKCCMTR MCGTMCCARSGCTCVYHWRRGHGCSCPG
P0DJH7	MEEREGGFRKETVERLLRLHFRDGRTRVNGDALLLMAELLKVFVREAAARAARQAQAEDLEKVDIE HVEKVLPQLLLDFV
P25969	MAFKPLHDRVLVRRVQSDEKTKGGLIIPDTAKEKPAEGEVVSCGEGARKDSGELIAMSVKAGDRVLF GKWSGTEVTIDGAELLIMKESDILGILS
P28599	MLKPLGDRVVIELVESEEKTASGIVLPDSAKEKPQEGKIVAAGSGRVLESGERVALEVKEGDRIIFSKY AGTEVKYEGTEYLILRESDILAVIG
P31295	MNIRPLHDRVVVRRMEEERLSAGGIVIPDSATEKPIQGEIIAVGHGKILDNGSVRALDVKVGDSVLFGK YSGTEVKLDGKEFLVMREEDIMAVVEG
Q8NSS1	MANVNIKPLEDKILVQINEAETTTASGLVIPDSAKEKPQEATVIAVGPGRFDDKGNRIPLDIKEDDVVIF SRYGGTEIKFGGVEYLLLSARDILAIVEK
P0A6F9	MNIRPLHDRVIVKRKEVETKSAGGIVLTGSAAAKSTRGEVLAVGNGRILENGEVKPLDVKVGDIVIFND GYGVKSEKIDNEEVLIMSESDILAIVEA
P94797	MNIRPLQDRVLVRRAEEEKKSAGGIILTGNAQEKPSQGEVVAVGNGKKLDNGTTLPMDVKVGDKVLF GKYSGSEVKVGDETLLMMREEDIMGIIA
P26879	MKIRPLHDRVVVRRMEEERTTAGGIVIPDSATEKPMRGEIIAVGAGKVLENGDVRALAVKVGDVVLF GKYSGTEVKVDGKELVVMREDDIMGVIEK
P24301	MAKVKIKPLEDKILVQAGEAETMTPSGLVIPENAKEKPQEGTVVAVGPGRWDEDGAKRIPVDVSEGD IVIYSKYGGTEIKYNGEEYLILSARDVLAVVSK
A0QSS3	MASVNIKPLEDKILVQANEAETTTASGLVIPDTAKEKPQEGTVVAVGPGRWDEDGEKRIPLDVAEGD TVIYSKYGGTEIKYNGEEYLILSARDVLAVVSK
Q3K7L5	MKLRPLHDRVVIRRSEEEKKTAGGIVLPGSAAEKANHGEILAVGPGKALESGEVRALSVKVGDKVVFG PYSGSNTVKVDGEDLLVMSENEILAVIEG
P80469	MSFKPLHDRIAIKPIENEEKTKGGIIIPDTAKEKPMQGEIVAVGNGVLNKNGEIYPLELKVGDKVLYGK WAGTEIEIKGEKLIVMKESDVFGIIN
P0A012	MLKPIGNRVIIEKKEQEQTTKSGIVLTDSAKEKSNEGVIVAVGTGRLLNDGTRVTPEVKEGDRVVFQQ YAGTEVKRDNETYLVLNEEDILAVIE
P99104	MLKPIGNRVIIEKKEQEQTTKSGIVLTDSAKEKSNEGVIVAVGTGRLLNDGTRVTPEVKEGDRVVFQQ YAGTEVKRDNETYLVLNEEDILAVIE
P0A014	MLKPIGNRVIIEKKEQEQTTKSGIVLTDSAKEKSNEGVIVAVGTGRLLNDGTRVTPEVKEGDRVVFQQ YAGTEVKRDNETYLVLNEEDILAVIE
Q97NV3	MLKPLGDRVVLKIEEKEQTVGGFVLAGSAQEKTKTAQVVATGQGVRTLNGDLVAPSVKTGDRVLVE AHAGLDVKDGDEKYIIVGEANILAIIEE
Q7YZS9	MMFRLTSVSCFLLVIACLNLFQVVLTSRCFPPGIYCTPYLPCCWGICCGTCRNVCHLRIGKRATFQE
D2DGD4	MKLVLAIVVILMLLSLSTGAEMSDNHASMSANALRDRLLGPKALLCGGTHARCNRDNDCCGSLCCFGT CISAFVPC
P13736	GSPMFV
P13737	GAPMFV
P02903	MEMKIDALAGTLESSDVMVRIGPAAQPGIQLEIDSIVKQQFGAAIEQVVRETLAQLGVKQANVVVDDK GALECVLRARVQAAALRAAQQTQLQWSQL
Q6QWF9	MSEVLPYGDEKLSPYGDGGDVGQIFSCRLQDTNNFFGAGQSKRPPKLGQIGRSKRVVIEDDRIDDVLK TMTDKAPPGV
Q9JI15	MSEVLPYGDEKLSPYGDGGDVGQIFSCRLQDTNNFFGAGQSKRPPKLGQIGRSKRVVIEDDRIDDVLK TMTDKAPPGV

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
O23249	MGQIQYSEKYFDDTFEYRHVVLPPEVAKLLPKNRLLSENEWRAIGVQQSRGWVHYAVHRPEPHIMLF RRPLNYQQQQENQAQNMLVK
P61024	MSHKQIYYSDKYDDEEFEYRHVMLPKDIAKLVPKTHLMSESEWRNLGVQQSQGWVHYMIHEPEPHIL LFRRPLPKKPKK
P55933	MPRDTIQYSEKYYDDKFEYRHVILPPDVAKEIPKNRLLSEGEWRGLGVQQSQGWVHYALHRPEPHILL FRREVPMPAASLSHNP
P33552	MAHKQIYYSDKYFDEHYEYRHVMLPRELSKQVPKTHLMSEEEWRRLGVQQSLGWVHYMIHEPEPHIL LFRRPLPKDQQK
Q2I2Q5	MLKMGVLLFTFLVLFPLTTLELDTDRPVERHAAIKQDLKPQERRGIRLHAPRDECCEPQWCDGACDC CS
P0C1N6	MLKMGVVLFIFLVLFPLATLQLDADQPVERYAENKQLLNPDERREILLPALRKFCCDSNWCHISDCEC CYG
P0C1N7	MSKLGVLLTICLLLFPLTALPLDGDQPADQAAERMQAEQHPLFDQKRRCCKFPCPDSCRYLCCG
Q86DU6	MMSKLGVLLTVCPLLFPLTALPPDGDQPADRPAERMQDDISSDEHPLFDKRQNCCNGGCSSKWCRDH ARCCGR
Q9BH73	MLKMGVVLFIFLVLFPLATLQLDADQPVERYAENKQLLSPDERREIILHALGTRCCSWDVCDHPSCTC CG
P0C1N8	MSKLGALLTICLLLFSLTAVPLDGDQHADQPAQRLQDRIPTEDHPLFDPNKRCCPPVACNMGCKPCC G
P58927	MMSKLGVLLIICLLLCPLTAVPQDGDQPADQPAERMQDDISSEHHPFFDPVKRCCKYGWTCWLGCSP CGC
D5L5Q7	MLKMGVVLFIFLVLFPLATLQLNADQPVERNAENIQDLNPDKRVIKIPVPRRRGPYRRYGNCYCPIG
Q9JJN6	MNREGAPGKSPEEMYIQQKVRVLLMLRKMGSNLTASEEEFLRTYAGVVSSQLSQLPQHSIDQGAEDV VMAFSRSETEDRRQ
P18545	MNLEPPKAEFRSATRVAGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDITVI CPWEAFNHLELHELAQYGII
P58917	MKLTCVVIVAMLLLTACQLITANDSRGTQKHRALRSDTKLSMSTRCKSTGASCRRTSYDCCTGSCRSG RCG
P58913	MKLTCVMIVAVLFLTAWTFVTADDSKNGLENHFWKARDEMKNREASKLDKKEACYAPGTFCGIKPG LCCSEFCLPGVCFGG
Q9XZK5	MKLTCVMIVAVLFLTTWTFVTADDSRYGLKNLFPKARHEMKNPEASKLNKRDGCSSGGTFCGIHPGL CCSEFCFLWCITFID
P05484	MKLTCVVIVAVLLLTACQLITADDSRGTQKHRALRSTTKLSTSTRCKGKGAKCSRLMYDCCTGSCRSG KCG
Q9BPA9	MEKLTILLLVAAVLTSTQALIQGGGDERQKAKINFLSRSDRDCRGYDAPCSSGAPCCDWWTCSARTN RCF
P56712	MKLTCMMIIAVLFLTAWTFVMADDPRDEPEARDEMNPAASKLNERGCLEVDYFCGIPFANNGLCCSG NCVFVCTPQGK
P85141	CLIQDCPEG
P84700	APANSVWS
P05486	CFIRNCPKG
P86255	CFIRNCPP
Q9NDA6	MGKLTILVLVAAVLLSAQVMVQGDGDQPADRKAVPREDNPGGASGKLMDVLRPKKCVLYPWCG
Q01519	MADQENSPLHTVGFDARFPQQNQTKHCWQSYVDYHKCVNMKGEDFAPCKVFWKTYNALCPLDWIE KWDDQREKGIFAGDINSD
P77921	MASHHEITDHKHGEMDIRHQQATFAGFIKGATWVSILSIAVLVFLALANS
P04038	MSTALAKPQMRGLLARRLRFHIVGAFMVSLGFATFYKFAVAEKRKKAYADFYRNYDSMKDFEEMRK AGIFQSAK
P09669	MAPEVLPKPRMRGLLARRLRNHMAVAFVLSLGVAALYKFRVADQRKKAYADFYRNYDVMKDFEEM RKAGIFQSVK
P15954	MLGQSIRRFTTSVVRRSHYEEGPGKNLPFSVENKWSLLAKMCLYFGSAFATPFLVVRHQLLKT
P17665	MLGQSIRRFTTSVVRRSHYEEGPGKNLPFSVENKWRLLAMMTVYFGSGFAAPFFIVRHQLLKK
Q1W0Y2	MLGQSIRRFTTSVVRRSHYEEGPGKNLPFSVENKWRLLAMMTLYFGSGFAAPFFIVRHQLLKK
P80432	MLGQSIRRFTTSVVRRSHYEEGPGKNLPFSVENKWRLLLMMTVYFGSGFAAPFFIVRHQLLKK
P80433	MSSLTPLLLRSLTGPARRLMVPRAQVHSKPPREQLGVLDITIGLTSCFVCCLLPAGWVLSHLESYKKRE
P0A314	MSGGGVFTDILAAAGRIFEVMVEGHWETVGMLFDSLGKGTMRINRNAYGSMGGGSLRGSSPEVSGYA VPTKEVESKFAK

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P45688	MAAVTGIALGMIETRGLVPAIEAADAMTKAAEVRLVGRQFVGGGYVTVLVRGETGAVNAAVRAGAD ACERVGDGLVAAHIIARVHSEVENILPKAPEA
P39158	MEQGTVKWFNAEKGFGFIERENGDDVFVHFSAIQSDGFKSLDEGQKVSFDVEQGARGAQAANVQKA
P0A978	MSNKMTGLVKWFNADKGFGFITPDDGSKDVFVHFTAIQSNEFRTLNENQKVEFSIEQGQRGPAAANV VTL
O54310	MRGKVKWFDSKKGYGFITKDEGGDVFVHWSAIEMEGFKTLKEGQVVEFEIQEGKKGPQAAHVKVV E
Q6D1T1	MLILTRRVGETLMIGDEVTVTVLGVKGNQVRIGVNAPKEVSVHREEIYQRIQAEKSQPTSY
Q3YEH5	MRCVPVFIILLLLSPSAPSVDAHPKTKDDVPLASFHDDAKRTLQRLWQNTWCCRDHLRCCG
P80232	GYLGGYAAPAVAVAPAPALAVAHAPAVVVATSYARISQVTNSVPIAVAAPAVPKAAVPVAAPVVAAA PVIAAHAPLALGHGFGYGGYH
P56562	QSGKDATIVELTNDNDGLGQYNFAYRTSDGIARQEQGALKNAGSENEAIEVQGSYTYKGVDGKDYTV TFVANENGYQPRVQS
O22912	MAGHKVAHATLKGPSVVKELFIGLALGLAAGGLWKMHHWNEQRKTRTFYDLLERGEISVVAAEE
P43023	MALPLKVLSRSMASAAKGDHGGAGANTWRLLTFVLALPGVALCSLNCWMHAGHHERPEFIPYHHLRI RTKPFAWGDGNHTLFHNPHVNPLPTGYEHP
P14854	MAEDMETKIKNYKTAPFDSRFPNQNQTRNCWQNYLDFHRCQKAMTAKGGDISVCEWYQRVYQSLC PTSWVTDWDEQRAEGTFPGKI
P56391	MAEDIKTKIKNYKTAPFDSRFPNQNQTKNCWQNYLDFHRCEKAMTAKGGDVSVCEWYRRVYKSLC PVSWVSAWDDRIAEGTFPGKI
P11951	MSSGALLPKPQMRGLLAKRLRVHIVGAFVVALGVAAAYKFGVAEPRKKAYADFYRNYDSMKDFEEM RQAGVFQSAK
P07470	MRALRVSQALVRSFSSTARNRFENRVAEKQKLFQEDNGLPVHLKGGATDNILYRVTMTLCLGGTLYS LYCLGWASFPHKK
P24310	MQALRVSQALIRSFSSTARNRFQNRVREKQKLFQEDNDIPLYLKGGIVDNILYRVTMTLCLGGTVYSLY SLGWASFPRN
P14406	MLRNLLALRQIGQRTISTASRRHFKNKVPEKQKLFQEDDEIPLYLKGGVADALLYRATMILTVGGTAY AIYELAVASFPKKQE
P48771	MLRNLLALRQIAQRTISTTSRRHFENKVPEKQKLFQEDNGMPVHLKGGASDALLYRATMALTLGGTA YAIYLLAMAAFPKKQN
P35171	MLRNVLALRQIAQRTISTTSRRHFENKVPEKQKLFQEDNGMPVHLKGGTSDALLYRATMLLTVGGTA YAIYMLAMAAFPKKQN
C4PWC3	MLKLEMMLVVLLILPLFYFDAGGQVVQRDWRSDGLARYLQRGDRDVRECNINTPGSSWGKCCLTRM CGPMCCARSGCTCVYHWRRGHGCSCPG
P0CE25	MPKLAVVLLVLLILPLSYFDAAGGQAVQWDRRGNGLARYLQRGDRDVRECQVDTPGSSWGKCCMTR MCGTMCCSRSVCTCVYHWRRGHGCSCPG
P0CE29	MPKLAVVLLVLLILPLSYFDVAGGQAAEGDRRGNGLARYPQRGGRDNEAECQINTPGSSWGKCCMT RMCGTMCCARSGCTCVYHWRRGHGCSCPG
P0CE27	MPKLAVVLLVLLILPLSYFDAAGGQAVQGDRRGNGLARYLQRNGRDNEAECQIDTPGSSWGKCCMT RMCGTMCCSRSVCTCVYHWRRGHGCSCPG
P0CE30	MPKLAVVLLVLLILPLSYFDAAGGQAAEGDRRGNGLARYLQRGGRDNEAECQINTPGSSWGKCCLTR MCGPMCCARSGCTCVYHWRRGHGCSCPG
P00122	DGESIYINGTAPTCSSCHDRGVAGAPELNAPEDWADRPSSVDELVESTLAGKGAMPAYDGRADREDL VKAIEYMLSTL
P00100	EDGAALFKSKPCAACHTIDSKMVGPALKEVAAKNAGVKDADKTLAGHIKNGTQGNWGPIPMPPNQV TDAEALTLAQWVLSLK
P00102	ASGEELFKSKPCGACHSVQAKLVGPALKDVAAKNAGVDGAADVLAGHIKNGSTGVWGAMPMPPNP VTEEEAKTLAEWVLTLK
P00121	AASAGGGARSADDIIAKHCNACHGAGVLGAPKIGDTAAWKERADHQGGLDGILAKAISGINAMPPKG TCADCSDDELREAIQKMSGL
P00109	IDINNGENIFTANCSACHAGGNNVIMPEKTLKKDALADNKMVSVNAITYQVTNGKNAMPAFGSRLAE TDIEDVANFVLTQSDKGWD
P00116	ADTVSGAALFKANCAQCHVGGGNLVNRAKTLKKEALEKYNMYSAKAIIAQVTHGKGAMPAFGIRLK AEQIENVAAYVLEQADNGWKK
P11448	GGDLEIGADVFTGNCAACHAGGANSVEPLKTLNKEDVTKYLDGGLSIEAITSQVRNGKGAMPAWSDR LDDEEIDGVVAYVFKNINEGW
Q09099	EADLALGKAVFDGNCAACHAGGGNNVIPDHTLQKAAIEQFLDGGFNIEAIVYQIENGKGAMPAWDGR LDEDEIAGVAAYVYDQAAGNKW
P00108	VDINNGESVFTANCSACHAGGNNVIMPEKTLKKDALEENEMNNIKSITYQVTNGKNAMPAFGGRLSE TDIEDVANFVISQSQKGW

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P00117	ADAAAGGKVFNANCAACHASGGGQINGAKTLKKNALTANGKDTVEAIVAQVTNGKGAMPAFKGRL SDDQIQSVALYVLDKAEKGW
P57736	SADLALGKQTFEANCAACHAGGNNSVIPDHTLRKAAMEQFLQGGFNLEAITYQVENGKGAMPAWSG TLDDDEIAAVAAYVYDQASGDKW
P00118	GDVAAGASVFSANCAACHMGGRNVIVANKTLSKSDLAKYLKGFDDDAVAAVAYQVTNGKNAMPGF NGRLSPKQIEDVAAYVVDQAEKGW
P0A3Y0	ADLANGAKVFSGNCAACHMGGGNVVMANKTLKKEALEQFGMYSEDAIIYQVQHGKNAMPAFAGRL TDEQIQDVAAYVLDQAAKGWAG
P00114	ADLANGAKVFSGNCAACHMGGGNVVMANKTLKKEALEQFGMNSEDAIIYQVQHGKNAMPAFAGRL TDEQIQDVAAYVLDQAAKGWAG
P00115	ADIADGAKVFSANCAACHMGGGNVVMANKTLKKEALEQFGMNSADAIMYQVQNGKNAMPAFGGRL SEAQIENVAAYVLDQSSNKWAG
P85003	PGLGFY
Q8WNR9	MIPGGLSEAKPATPEIQEIANEVKPQLEEKTNETYQKFEAIEYKTQVVAGINYYIKVQVDDNRYIHIKV FKGLPVQDSSLTLTGYQTGKSEDDELTGF
P01040	MIPGGLSEAKPATPEIQEIVDKVKPQLEEKTNETYGKLEAVQYKTQVVAGTNYYIKVRAGDNKYMHL KVFKSLPGQNEDLVLTGYQVDKNKDDELTGF
P25417	MMCGGTSATQPATAETQAIADKVKSQLEEKENKKFPVFKALEFKSQLVAGKNYFIKVQVDEDDFVHI RVFESLPHENKPVALTSYQTNKGRHDELTYF
P04080	MMCGAPSATQPATAETQHIADQVRSQLEEKENKKFPVFKAVSFKSQVVAGTNYFIKVHVGDEDFVHL RVFQSLPHENKPLTLSNYQTNKAKHDELTYF
Q62426	MMCGAPSATMPATAETQEVADQVKSQLESKENQKFDVFKAISFKRQIVAGTNLFIKVDVGGDKCVHL RVFQPLPHENKPLTLSSYQTNKERHDELSYF
Q29290	MMCGAPSATQPATAEIQAIADKVKSQLEEKENKTFPVFKAVEFKSQVVAGRNLFIKVQVDDDDFVHL RVFESLPHENKPLTLSSYQTNKSRHDELTYF
P01041	MMCGAPSATMPATTETQEIADKVKSQLEEKANQKFDVFKAISFRRQVVAGTNFFIKVDVGEEKCVHL RVFEPLPHENKPLTLSSYQTDKEKHDELTYF
Q10994	MMCGAPSATQPATAETQAIADKVKSQLEEKENKKFPVFKALEFKSQLVAGKNYFIKVQVDEDDFVHI RVFESLPHENKPVALTSYQTNKGRHDELTYF
P08821	MNKTELINAVAEASELSKKDATKAVDSVFDTILDALKNGDKIQLIGFGNFEVRERSARKGRNPQTGEEI EIPASKVPAFKPGKALKDAVAGK
P0ACF0	MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTFKVNHRAERTGRNPQTGK EIKIAAANVPAFVSGKALKDAVK
E0J6W8	MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTFKVNHRAERTGRNPQTGK EIKIAAANVPAFVSGKALKDAVK
P0ACF4	MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTGRNPQTGKEI TIAAAKVPSFRAGKALKDAVN
P0A3H1	MNKTELINAVAETSGLSKKDATKAVDAVFDSITEALRKGDKVQLIGFGNFEVRERAARKGRNPQTGE EMEIPASKVPAFKPGKALKDAVK
P05385	MNKAELITSMAEKSKLTKKDAELALKALIESVEEALEKGEKVQLVGFGTFETRERAAREGRNPRTKE VINIPATTVPVFKAGKEFKDKVNK
P0A3H0	MNKTELINAVAETSGLSKKDATKAVDAVFDSITEALRKGDKVQLIGFGNFEVRERAARKGRNPQTGE EMEIPASKVPAFKPGKALKDAVK
P36206	MTKKELIDRVAKKAGAKKKDVKLILDTILETITEALAKGEKVQIVGFGSFEVRKAAARKGVNPQTRKP ITIPERKVPKFKPGKALKEKVK
P19436	AAKKTVTKADLVDQVAQATGLKKKDVKAMVDALLAKVEEALANGSKVQLTGFGTFEVRKRKARTG VKPGTKEKIKIPATQYPAFKPGKALKDKVKK
D2KX92	MAYLKIVLVALMLVLAVSAMRRPDQQDQDISVAKRVACKCDDDGPDIRSATLTGTVDLGSCDEGWEK CASYYTVIADCCRRPRS
D2KX91	MAYQKIVFVALMLVLAVSAMRLPDQQDQDISVAKRVACKCDDDGPDIRSATLTGTVDLGSCNEGWE KCASYYTVVADCCRRPRS
P82373	TGTGPSLSIVNPLDVLRQRLLLEIARRRMRQTQNMIQANRDFLESI
P23465	MGMGPSLSIVNPMDVLRQRLLLEIARRRLRDAEEQIKANKDFLQQI
P67800	NKPSLSIVNPLDVLRQRLLLEIARRQMKENTRQVELNRAILKNV
P41538	TGSGPSLSIVNPLDVLRQRLLLEIARRRMRQSQDQIQANREILQTI
P67801	NKPSLSIVNPLDVLRQRLLLEIARRQMKENTRQVELNRAILKNV
P82707	TGAVPSLSIVNPLDVLRQRLLLEIARRRMRQSQDQIQANREMLQTI
Q9NP97	MAEVEETLKRLQSQKGVQGIIVVNTEGIPIKSTMDNPTTTQYASLMHSFILKARSTVRDIDPQNDLTFL RIRSKKNEIMVAPDKDYFLIVIQNPTE

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P62627	MAEVEETLKRLQSQKGVQGIIVVNTEGIPIKSTMDNPTTTQYANLMHNFILKARSTVREIDPQNDLTFL RIRSKKNEIMVAPDKDYFLIVIQNPTE
P62628	MAEVEETLKRLQSQKGVQGIIVVNTEGIPIKSTMDNPTTTQYANLMHNFILKARSTVREIDPQNDLTFL RIRSKKNEIMVAPDKDYFLIVIQNPTE
Q8TF09	MAEVEETLKRIQSHKGVIGTMVVNAEGIPIRTTLDNSTTVQYAGLLHHLTMKAKSTVRDIDPQNDLTF LRIRSKKHEIMVAPDKEYLLIVIQNPCE
Q9DAJ5	MTEVEETLKRIQSHKGVIGTMVVNAEGIPIRTTLDNSTTVQYAGLLHQLTMKAKSTVRDIDPQNDLTF LRIRSKKHEIMVAPDKEYLLIVIQNPCE
Q8K3E7	MESEQMLEGQTQVAENPHSEYGLTDSVERIVENEKINAEKSSKQKVDLQSLPTRAYLDQTVVPILLQG LAVLAKERPPNPTEFLASYLLKNKAQFEDRN
P61285	MCDRKAVIKNADMSEEMQQDSVECATQALEKYNIEKDIAAHIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
Q22799	MVDRKAVIKNADMSDDMQQDAIDCATQALEKYNIEKDIAAYIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
Q39580	MASGSSKAVIKNADMSEEMQADAVDCATQALEKYNIEKDIAAYIKKEFDRKHNPTWHCIVGRNFGSY VTHETKHFIYFYLGQVAILLFKSG
Q24117	MSDRKAVIKNADMSEEMQQDAVDCATQALEKYNIEKDIAAYIKKEFDKKYNPTWHCIVGRNFGSYVT HETRHFIYFYLGQVAILLFKSG
P63167	MCDRKAVIKNADMSEEMQQDSVECATQALEKYNIEKDIAAHIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
P63168	MCDRKAVIKNADMSEEMQQDSVECATQALEKYNIEKDIAAHIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
P63169	MCDRKAVIKNADMSEEMQQDSVECATQALEKYNIEKDIAAHIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
P63170	MCDRKAVIKNADMSEEMQQDSVECATQALEKYNIEKDIAAHIKKEFDKKYNPTWHCIVGRNFGSYVT HETKHFIYFYLGQVAILLFKSG
Q94758	MSERKAVIKNADMSEEMQEDAIHIAAGAIDKHDLEKDIAANIKKDFDRKYHPTWHCIVGRHFGSYVTH ETHNFIYFYLDDRAFLLFKSG
Q02647	MSDENKSTPIVKASDITDKLKEDILTISKDALDKYQLERDIAGTVKKQLDVKYGNTWHVIVGKNFGSY VTHEKGHFVYFYIGPLAFLVFKTA
Q96FJ2	MSDRKAVIKNADMSEDMQQDAVDCATQAMEKYNIEKDIAAYIKKEFDKKYNPTWHCIVGRNFGSYV THETKHFIYFYLGQVAILLFKSG
Q9D0M5	MSDRKAVIKNADMSEDMQQDAVDCATQAMEKYNIEKDIAAYIKKEFDKKYNPTWHCIVGRNFGSYV THETKHFIYFYLGQVAILLFKSG
Q78P75	MSDRKAVIKNADMSEDMQQDAVDCATQAMEKYNIEKDIAAYIKKEFDKKYNPTWHCIVGRNFGSYV THETKHFIYFYLGQVAILLFKSG
P11919	MAGKVTVAFFMFAMIAFLANFGYVECNPAIATGYDPMEICIENCAQCKKMLGAWFEGPLCAESCIKF KGKLIPECEDFASIAPFLNKL
O27734	MGDVVATIKVMPESPDVDLEALKKEIQERIPEGTELHKIDEEPIAFGLVALNVMVVVGDAEGGTEAAE ESLSGIEGVSNIEVTDVRRLM
P82096	FVPIWM
P82099	FVHPM
P82100	FITVH
P82101	IYEPEIA
P11925	ISINQDLKAITDMLLTEQIQARRRCLAALRQRLLDLDSDVSLFNGDLLPNGRCS
P64093	MSQIMYNYPAMMAHAGDMAGYAGTLQSLGADIASEQAVLSSAWQGDTGITYQGWQTQWNQALEDL VRAYQSMSGTHESNTMAMLARDGAEAAKWGG
P86455	CPPLC
P04204	MKSILWLCVFGLLIATLFPVSWQMAIKSRLSSEDSETDQRLFESKRHSDAIFTEEYSKLLAKLALQKYLA SILGSRTSPPPPSR
P26349	MKIILWLCVFGLFLATLFPISWQMPVESGLSSEDSASSESFASKIKRHGEGTFTSDLSKQMEEEAVRLFI EWLKNGGPSSGAPPPSG
Q3UGS4	MTSSPVSRVVYNGKRNSSPRSPTNSSEIFTPAHEENVRFIYEAWQGVERDLRSQLSSGERCLVEEYVEK VPNPSLKTFKPIDLSDLKRRNTQDAKKS
P83275	ADKNFLRF
P85451	GGNDFMRF
P85477	ASNQDFMRF
P82661	AMRNALVRF

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P83279	DGGRNFLRF
P85454	QANQDFMRF
P41863	GANDFMRF
P83308	LPLRF
P18523	QDVVHSFLRF
P20491	MISAVILFLLLLVEQAAALGEPQLCYILDAVLFLYGIVLTLLYCRLKIQVRKAAIASREKADAVYTGLNT RSQETYETLKHEKPPQ
P20411	MIPAVILFLLLLVEEAAALGEPQLCYILDAILFLYGIVLTLLYCRLKIQVRKADIASREKSDAVYTGLNTR NQETYETLKHEKPPQ
P64638	MASLIQVRDLLALRGRMEAAQISQTLNTPQPMINAMLQQLESMGKAVRIQEEPDGCLSGSCKSCPEGK ACLREWWALR
P0A3C8	MATFKVTLINEAEGTKHEIEVPDDEYILDAAEEQGYDLPFSCRAGACSTCAGKLVSGTVDQSDQSFLD DDQIEAGYVLTCVAYPTSDVVIQTHKEEDLY
P00254	MATFKVTLINEAEGTSNTIDVPDDEYILDAAEEQGYDLPFSCRAGACSTCAGKLVSGTVDQSDQSFLD DDQIEAGYVLTCVAYPTSDVTIQTHKEEDLY
P17007	MAVYKVRLICEEQGLDTTIECPDDEYILDAAEEQGIDLPYSCRAGACSTCAGKVVEGTVDQSDQSFLD DAQLAAGYVLTCVAYPSSDCTVKTHQEESLY
P07485	TIVIDHEECIGCESCVELCPEVFAMIDGEEKAMVTAPDSTAECAQDAIDACPVEAISKE
P00235	AYKTVLKTPSGEFTLDVPEGTTILDAAEEAGYDLPFSCRAGACSSCLGKVVSGSVDESEGSFLDDGQM EEGFVLTCIAIPESDLVIETHKEEELF
P00234	AYKTVLKTPSGEFTLDVPEGTTILDAAEEAGYDLPFSCRAGACSSCLGKVVSGSVDQSEGSFLDDGQM EEGFVLTCIAIPESDLVIETHKEEELF
P84873	ATYKVKLVTPDGPVEFNCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVSAGTVDQSDGNFLDDD QIADGFVLTCVAYPQSDVTIETHKEEDLTG
P00252	MATVYKVTLVDQEGTETTIDVPDDEYILDIAEDQGLDLPYSCRAGACSTCAGKIVSGTVDQSDQSFLD DDQIEKGYVLTCVAYPTSDLKIETHKEEDLY
P0A3C7	MATFKVTLINEAEGTKHEIEVPDDEYILDAAEEQGYDLPFSCRAGACSTCAGKLVSGTVDQSDQSFLD DDQIEAGYVLTCVAYPTSDVVIQTHKEEDLY
P14936	SAVYKVKLIGPDGQENEFDVPDDQYILDAAEEAGVDLPYSCRAGACSTCAGKIEKGQVDQSDGSFLED HHFEKGYVLTCVAYPQSDLVIHTHKEEELF
P0A3D3	MATYKVTLVNAAEGLNTTIDVADDTYILDAAEEQGIDLPYSCRAGACSTCAGKVVSGTVDQSDQSFL DDDQIAAGFVLTCVAYPTSDVTIETHKEEDLY
P10624	MAKYLYLDQDECMACESCVELCPEAFRMSSAGEYAEVIDPNTTAECVEDAISTCPVECIEWREE
P84874	ATYKVKLVTPDGPVEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVKAGTVDQSDGNFLDDD QMADGFVLTCVAYPQSDVTIETHKEEDLTG
P00249	MATYKVRLFNAAEGLDETIEVPDDEYILDAAEEAGLDLPFSCRSGSCSSCNGILKKGTVDQSDQNFLDD DQIAAGNVLTCVAYPTSNCEIETHREDAIA
P00231	AASYKVTFVTPSGTNTITCPADTYVLDAAEESGLDLPYSCRAGACSSCAGKVTAGAVNQEDGSFLEEE QMEAGWVLTCVAYPTSDVTIETHKEEDLTA
P00232	AASYKVTFVTPSGTKTITCPADTYVLDAAEDTGLDLPYSCRAGACSSCAGKVTAGSVNQEDGSFLDE EQMEAGWVLTCVAYPTSDVTIETHKEEDLSA
P00224	ATYKVTLVTPSGSQVIECGDDEYILDAAEEKGMDLPYSCRAGACSSCAGKVTSGSVDQSDQSFLEDGQ MEEGWVLTCIAYPTGDVTIETHKEEELTA
P00241	ASYKIHLVNKDQGIDETIECPDDQYILDAAEEQGLDLPYSCRAGACSTCAGKLLEGEVDQSDQSFLDDD QVKAGFVLTCVAYPTSNATILTHQEESLY
P81372	ATYKVKLVTPQGQQEFDCPDDVYILDQAEEEGIDLPYSCRAGSCSSCAGKVKQGEVDQSDGSFLDDE QMEQGWVLTCVAFPTSDVVIETHKEEELTA
P81373	ATYKVKLVTPSGQPLEFECPDDVYILDQAEEEGIDLPYSCRAGSCSSCAGKVKNGNVDQSDGSFLDDD QIGEGWVLTCVAYPTSDVVIETHKEEELTA
P11053	MASYQVRLINKKQDIDTTIEIDEETTILDGAEENGIELPFSCHSGSCSSCVGKVVEGEVDQSDQIFLDDE QMGKGFALLCVTYPRSNCTIKTHQEPYLA
P00250	MASYKVTLKTPDGDNVITVPDDEYILDVAEEEGLDLPYSCRAGACSTCAGKLVSGPAPDEDQSFLDDD QIQAGYILTCVAYPTGDCVIETHKEEALY
P00223	ATYKVTLITPEGKQEFEVPDDVYILDHAAEEVGDLPYSCRAGSCSSCAGKVTAGSVDQSDGSYLDDDQ MEAGWVLTCVAYPTSDVTIETHKEEELTA
P84872	ATYKVKLVTPDGPVEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVSAGTVDQSDGNFLDDD QMADGFVLTCVAYPQSDVTIETHKEEELTG
P10245	PKYTIVDKETCIACGACGAAAPDIYDYDEDGIAYVTLDDNQGIVEVPDILIDDMMDAFEGCPTDSIKVA DEPFDGDPNKFE

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P13106	ETYSVTLVNEEKNINAVIKCPDDQFILDAAEEQGIELPYSCRAGACSTCAGKVLSGTIDQSEQSFLDDD QMGAGFLLTCVAYPTSDCKVQTHAEDDLY
P83527	ASYKVKLITPDGPIEFDCPDDVYILDQAEEAGHDLPYSCRAGSCSSCAGKIAGGAVDQTDGNFLDDDQ LEEGWVLTCVAYPQSDVTIETHKEAELVG
P56408	YKVTLKTPSGEETIECPEDTYILDAAEEAGLDLPYSCRAGACSSCAGKVESGEVDQSDQSFLDDAQMG KGFVLTCVAYPTSDVTILTHQEAALY
P00222	ATYKVKLVTPSGQQEFQCPDDVYILDQAEEVGIDLPYSCRAGSCSSCAGKVKVGDVDQSDGSFLDDEQ IGEGWVLTCVAYPVSDGTIETHKEEELTA
P83520	ATYKVKLVTPDGPVEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVTAGTVDQSDGNFLDDD QMADGFVLTCVAYPQSDVTIETHKEEELTG
P68163	ATYKVKLVTPDGPVEFDCPDDVYILDRAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGNYLDDD QMAEGFVLTCVAYPQSDVTIETHKEEELTG
P68164	ATYKVKLVTPDGPVEFDCPDDVYILDRAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGNYLDDD QMAEGFVLTCVAYPQSDVTIETHKEEELTG
P68166	ATYKVKLVTPDGPVEFNCPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGNYLDDD QMADGFVLTCVAYPQSDVTIETHKEEELTG
P68165	ATYKVKLVTPDGPVEFNCPDDVYILDQAEEEGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGNYLDDD QMADGFVLTCVAYPQSDVTIETHKEEELTG
P00212	PKYTIVDKETCIACGACGAAAPDIYDYDEDGIAYVTLDDNQGIVEVPDILIDDMMDAFEGCPTESIKVA DEPFDGDPNKFD
P00233	AIFKVKFLTPDGERTIEVPDDKFILDAGEEAGLDLPYSCRAGACSSCTGKLLDGRVDQSEQSFLDDDQ MAEGFVLTCVAYPAGDITIETHAEEKL
P83522	ATYKVKLVTPEGEVELEVPDDVYILDQAEEEGIDLPYSCRAGSCSSCAGKLVSGEIDQSDQSFLDDDQ MEEGWVLTCAAYPKSDVVIETHKEEELTA
P00225	AFKVKLLTPDGPKEFECPDDVYILDQAEELGIELPYSCRAGSCSSCAGKLVEGDLDQSDQSFLDDEQIE EGWVLTCAAYPRSDVVIETHKEEELTG
P83523	ATYKVKLVTPDGPVEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVSAGTVDQSDGNFLDDD QIADGFVLTCVAYPQSDVTIETHKEEALTG
P00248	MATYKVTLINEAEGLNKTIEVPDDQYILDAAEEAGIDLPYSCRAGACSTCAGKLISGTVDQSDQSFLDD DQIEAGYVLTCVAYPTSDCVIETHKEEELY
P00253	MATFKVTLINEAEGTKHEIEVPDDEYILDAAEEEGYDLPFSCRAGACSTCAGKLVSGTVDQSDQSFLD DDQIEAGYVLTCVAYPTSDVVIQTHKEEDLY
P85121	ATYYKVKLLTPEGEKEFECPDDVYILDNAEEIGIDLPYSCRAGSCSSCAGKVVSGKVDNSDNSFLNDDN MDAGYVLTCHAYANSDVVIETHKEEEV
P83524	ATYKVKLITPDGPVVFDCPDNEYILDAAEEQGHDLPYSCRAGSCSSCAGKVTAGTVDQSDGNFLDDD QVADGFVLTCVAYPQSDVTIETHKEEELTA
P00226	ASYKVKLITPDGPQEFECPDDVYILEHAEELGIDIPYSCRAGSCSSCAGKLVAGSVDQSDQSFLDDEQIE EGWVLTCVAYPKSDVTIETHKEEELTA
P83525	ATYKVKLVTPDGPVEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVSAGTVDQSDGNFLDDD QMADGFVLTCVAYPQSDVIIETHKEEELTG
P83585	ATYKVKLVTPDGPVEFECPDDEYILDRAEEEGHDLPYSCRAGSCSSCAGKIAAGSVDQSDGNFLDDDQ IADGFVLTCVAYPQSDVTIETHKEEELTA
P83584	ASYKVKLITPDGPIEFNCPDDVYILDRAEEEGHDLPYSCRAGACSSCAGKIVDGSVDQSDNSFLDDDQI GGGFVLTCVAYPKSNVTIETHKEEALVG
P83583	ATYKVKLITPEGPVEFNCPDDVYILDSAEENGHDLPYSCRAGACSSCAGKITAGNVDQSDNSFLDDDQ VAEGFVLTCVAYPKSNVTIETHKEDDLVG
P83582	ATYKVKLVTPDGPIEFDCPDDVYILDQAEEEGHELPYSCRAGSCSSCAGKVTAGTVDQSDGNFLDDD QMADGFVLTCVAYPKSDVTIETHKEEDLTG
P00245	MATYKVTLISEAEGINETIDCDDDTYILDAAEEAGLDLPYSCRAGACSTCAGKITSGSIDQSDQSFLDDD QIEAGYVLTCVAYPTSDCTIQTHQEEGLY
P00246	MATYKVTLINEAEGINETIDCDDDTYILDAAEEAGLDLPYSCRAGACSTCAGTITSGTIDQSDQSFLDD DQIEAGYVLTCVAYPTSDCTIKTHQEEGLY
P00255	MATYKVTLVRPDGETTIDVPEDEYILDVAEEQGLDLPFSCRAGACSTCAGKLLEGEVDQSDQSFLDDD QIEKGFVLTCVAYPRSDCKILTHQEEELY
P15788	MASYKVTLINEEMGLNETIEVPDDEYILDVAEEEGIDLPYSCRAGACSTCAGKIKEGEIDQSDQSFLDD DQIEAGYVLTCVAYPASDCTIITHQEEELY
P0A3C9	MATYKVTLVRPDGSETTIDVPEDEYILDVAEEQGLDLPFSCRAGACSTCAGKLLEGEVDQSDQSFLDD DQIEKGFVLTCVAYPRSDCKILTNQEEELY
P83526	ASYKVKLITPEGAVEFDCPDDVYILDQAEEMGHDLPYSCRAGSCSSCAGKVTAGNVDQSDGNFLDDD QMADGFVLTCVAYPQSDVTIETHKEEELTA

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P44048	MARTVFCEYLKKEAEGLDFQLYPGELGKRIFDSVSKQAWGEWIKKQTMLVNEKKLNMMNAEHRKLL EQEMVNFLFEGKDVHIEGYVPPSN
Q9HU36	MSRTVMCRKYHEELPGLDRPPYPGAKGEDIYNNVSRKAWDEWQKHQTMLINERRLNMMNAEDRKF LQQEMDKFLSGEDYAKADGYVPPSA
P67617	MSRTIFCTYLQRDAEGQDFQLYPGELGKRIYNEISKDAWAQWQHKQTMLINEKKLNMMNAEHRKLL EQEMVSFLFEGKDVHIEGYTPEDKK
P24018	MNVGDRVRVKESVVVYHHPDHRNQAFDLKDAEGEIAAILTEWNGKPISANFPYLVSFSNKFRAHLRD FELEVI
Q91XV6	METVLILCSLLAPVVLASAAEKEKEKDPFYYDYQTLRIGGLVFAVVLFSVGILLILSRRCKCSFNQKPRA PGDEEAQVENLITTNAAEPQKAEN
P59649	MATPTQSPTNVPEETDPFFYDYATVQTVGMTLATIMFVLGIIIIISKKVKCRKADSRSESPTCKSCKSEL PSSAPGGGGV
P69543	MIKVEIKPSQAQFTTRSGVSRQGKPYSLNEQLCYVDLGNEYPVLVKITLDEGQPAYAPGLYTVHLSSF KVGQFGSLMIDRLRLVPAK
P69542	MIKVEIKPSQAQFTTRSGVSRQGKPYSLNEQLCYVDLGNEYPVLVKITLDEGQPAYAPGLYTVHLSSF KVGQFGSLMIDRLRLVPAK
P69544	MIKVEIKPSQAQFTTRSGVSRQGKPYSLNEQLCYVDLGNEYPVLVKITLDEGQPAYAPGLYTVHLSSF KVGQFGSLMIDRLRLVPAK
O67904	MVDREWVLKIAKLARLELKEEEIEVFQKQLSDILDFIDQLKELDTENVEPYIQEFEETPMREDEPHPSL DREKALMNAPERKDGFFVVPRVVEV
Q97SE5	MKITQEEVTHVANLSKLRFSEEETAAFATTLSKIVDMVELLGEVDTTGVAPTTTMADRKTVLRPDVA EEGIDRDRLFKNVPEKDNYYIKVPAILDNGGDA
P50151	MSSGASASALQRLVEQLKLEAGVERIKVSQAAAELQQYCMQNACKDALLVGVPAGSNPFREPRSCALL
Q28024	MSSKTASTNNIAQARRTVQQLRMEASIERIKVSKASADLMSYCEEHARNDPLLMGIPTSENPFKDKKT CTIL
Q9UBI6	MSSKTASTNNIAQARRTVQQLRLEASIERIKVSKASADLMSYCEEHARSDPLLIGIPTSENPFKDKKTCII L
Q9DAS9	MSSKTASTNSIAQARRTVQQLRLEASIERIKVSKASADLMSYCEEHARSDPLLMGIPTSENPFKDKKTCI IL
P50150	MKEGMSNNSTTSISQARKAVEQLKMEACMDRVKVSQAAADLLAYCEAHVREDPLIIPVPASENPFREK KFFCTIL
P30671	MSATNNIAQARKLVEQLRIEAGIERIKVSKASSELMSYCEQHARNDPLLVGVPASENPFKDKKPCIIL
O06721	MPAIVGAFKINAIGTSGVVHIGDCITISPQAQVRTFAGAGSFNTGDSLKVMNYQNATNVYDNDAVDQP IVANA
P0A3T7	MNFYVNQSIIINSIKIDSITTSSVFQIGTAGSIKALSKFSNTGGFTEPLRPLQAKGQIISIKPSTSSS
P99025	MPYLLISTQIRMEVGPTMVGDEHSDPELMQHLGASKRSVLGNNFYEYYVNDPPRIVLDKLECKGFRVL SMTGVGQTLVWCLHKE
P70552	MPYLLISTQIRMEVGPTMVGDEHSDPELMQQLGASKRRVLGNNFYEYYVNDPPRIVLDKLECRGFRV LSMTGVGQTLVWCLHKE
Q76CA0	MAYLKIVLVALMLVVAVSAMRLSDQEDQDISVAKRAACKCDDDGPDIRSATLTGTVDLGSCNEGWEK CASFYTILADCCRRPRG
P80665	FPDGEFLMQGCPECKLGENRFFSKPGAPVYQCTGCCFSRAYPTPLRSKKTMLVPKNITSEATCCVAK AFTKITLKDNVKIENHTECHCSTCYYHKS
P37204	YPNVDLSNMGCEECTLKKNNVFSRDRPIYQCMGCCFSRAFPTPLKAMKTMTIPKNITSEATCCVAKH SYETEVAGIRVRNHTDCHCSTCYFHKS
P73492	MAVSAKIEIYTWSTCPFCMRALALLKRKGVEFQEYCIDGDNEAREAMAARANGKRSLPQIFIDDQHIG GCDDIYALDGAGKLDPLLHS
P0AC62	MANVEIYTKETCPYCHRAKALLSSKGVSFQELPIDGNAAKREEMIKRSGRTTVPQIFIDAQHIGGCDDL YALDARGGLDPLLK
P51918	MAAKILALWLLLAGTVFPQGCCQHWSYGLSPGGKRDLDNFSDTLGNMVEEFPRVEAPCSVFGCAEES PFAKMYRVKGLLASVAERENGHRTFKK
P01148	MKPIQKLLAGLILLTWCVEGCSSQHWSYGLRPGGKRDAENLIDSFQEIVKEVGQLAETQRFECTTHQP RSPLRDLKGALESLIEEETGQKKI
P13562	MILKLMAGILLLTVCLEGCSSQHWSYGLRPGGKRNTEHLVESFQEMGKEVDQMAEPQHFECTVHWP RSPLRDLRGALESLIEEEARQKKM
P49921	MEPIPKLLAGLLLLTLCVVGCSSQHWSYGLRPGGKRNAENVIDSFQEMAKEVARLAEPQRFECTAHQ PRSPLRDLKGALESLIEEETGQKT
P07490	METIPKLMAAVVLLTVCLEGCSSQHWSYGLRPGGKRNTEHLVDSFQEMGKEEDQMAEPQNFECTVH WPRSPLRDLRGALERLIEEEAGQKKM

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P43306	MVSVCRLLLVAALLLCLQAQLSFSQHWSHGWYPGGKREIDSYSSPEISGEIKLCEAGECSYLRPLRTNIL KSILIDTLARKFQKRK
P37044	MCVSRLALLLGLLLCVGAQLSFAQHWSHGWYPGGKRELDSFGTSEISEEIKLCEAGECSYLRPQRRSIL RNILLDALARELQKRK
P51925	MCVSRLVLLLGLLLCVGAQLSNGQHWSHGWYPGGKRELDSFGTSEISEEIKLCEAGECSYLTPQRRSV LRNILLDALARELQKRK
Q9IA09	MEANSRVMVRVLLLALVVQVTLSQHWSYGWLPGGKRSVGELEATIRMMGTGEVVSLPEEASAQTQE RLRPYNVINDDSSHFDRKKRSPNK
P45652	MEAGSRVIMQVLLLALVVQVTLSQHWSYGWLPGGKRSVGELEATIRMMGTGGVVSLPDEANAQIQER LRPYNIINDDSSHFDRKKRFPNN
P30973	MDLSSKTVVQVVMLALIAQVTFSQHWSYGWLPGGKRSVGELEATIRMMDTGGVMALPEETGAHIPE RLRPYDVMSKKRMPHK
P69109	MDLSNRTVVQVVVLALVAQVTLSQHWSYGWLPGGKRSVGELEATIKMMDTGGVVALPEETSAHVSE RLRPYDVILKKWMPHK
P51921	MEASSRVTVQVLLLALVVQVTLSQHWSYGWLPGGKRSVGELEATIRMMGTGGVVSLPEEASAQTQE RLRPYNVIKDDSSHFDRKKRFPNK
P69108	MDLSNRTVVQVVVLALVAQVTLSQHWSYGWLPGGKRSVGELEATIKMMDTGGVVALPEETSAHVSE RLRPYDVILKKWMPHK
P69107	MDLSNRTVVQVVVLALVAQVTLSQHWSYGWLPGGKRSVGELEATIKMMDTGGVVALPEETSAHVSE RLRPYDVILKKWMPHK
P45653	MDLSNRTVVQVVVLALVAQVTLSQHWSYGWLPGGKRSVGELEATIKMMDTGGVVALPEETSAHFSE RLRPYDVILKKWMPHK
P51923	MEASSRVTVQVLLLALVVQVTLSQHWSYGWLPGGKRSVGELEATIRMMGTGGVVSLPEEASAQTQE RLRPYNVIKDDSSPFDRKKRFPNK
D2Y2C7	MTTVGVSLFRRSPEKITMKIAAFLGLSFLLIASYVLICEAQHPGFQELLILEENMRDPENSKERSCAKPR ENCNRMNILCCRGECVCPTFGDCFCYGD
D2Y2C8	MTTVGVSLFRRSPEKITMKIATFLGLSFLLIASYFLICEAQHPGFQELLILEENMRDPENSKERSCAKPR ENCNRMNILCCRGECVCPTFGDCFCYGD
D2Y240	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y241	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDDDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y242	MVNMKASMFLTFAGLVLLFVVCYAPESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y243	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVGNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y248	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRC
D2Y249	MVNMKASMFLTFAGLVLLFVVCYASESEKKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y250	MVNMKASMFLTFAGLVLLFVVSYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2G3	MVNMKASMFLTFAGLVLLFVVCYASGSEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2G4	MVDMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2G5	MVNMEASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2G6	MVNMKASMFLTFAGLVLLFVACYASESEKKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2M1	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDKDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2M4	MVNMKASMFLTSAGLVPLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2M5	MVNMKASMSLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWCR
D2Y2M6	MVNMKASMFLTFAGLVLLLVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2M8	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQGERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
D2Y2N0	MVNVKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2N1	MVNTKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
D2Y2N4	MVNMKALMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
Q1XF11	MVPPGNILFLLLLPVATAQMTPGSCSGCGPLSLPLLAGLVAADAVVSLLIVVVVFVCARLRSRPTQEDD KIYINMPGRG
O31796	MKPINIQDQFLNQIRKENTYVTVFLLNGFQLRGQVKGFDNFTVLLESEGKQQLIYKHAISTFAPQKNV QLELE
Q2YPW9	MAERSQNLQDLFLNSVRKQKISLTIFLINGVKLTGIVTSFDNFCVLLRRDGHSQLVYKHAISTIMPSQPV QMFEGEEA
Q9Y241	MSTDTGVSLPSYEEDQGSKLIRKAKEAPFVPVGIAGFAAIVAYGLYKLKSRGNTKMSIHLIHMRVAAQ GFVVGAMTVGMGYSMYREFWAKPKP
P83341	AEKLEESSAEAKALSYVHDATTSGHDSYQEGQKCINCLLYTDPSQEEWGGCAVFPGKLVNANGWCTA YVARG
P04168	EPRAEDGHAHDYVNEAADASGHPRYQEGQLCENCAFWGEAVQDGWGRCTHPDFDEVLVKAEGWC SVYAPAS
P33678	GTNAAMRKAFNYQDTAKNGKKCSGCAQFVPGASPTAAGGCKVIPGDNQIAPGGYCDAFIVKK
B3EBZ6	QDLPHVDPATDPTAQALKYSEDAANADRAAAARPGKPPEEQFCHNCQFVLADSGEWRPCSLFPGKA VHETGWCASWTLKAG
B3EBZ5	EVPADAVTESDPTAVALKYHRNAAESERVAAARPGLPPEEQHCENCQFMLPDQGADEWRGCSLFPG KLINLNGWCASWTLRAG
P00262	EVPANAVTESDPTAVALKYHRNAEASERVAAARPGLPPEEQHCENCQFMLPDQGADEWRGCSLFPG KLINLDGWCASWTLRAG
P59860	VPANAVTESDPAAVALKYHRDAASSERVAAARPGLPPEEQHCENCQFMNPDSAAADWKGCQLFPGK LINLSGWCASWTLRAG
P80176	AAPANAVTADDPTAIALKYNQDATKSERVAAARPGLPPEEQHCANCQFMQANVGEGDWKGCQLFP GKLINVNGWCASWTLKAG
P00261	EAPANAVAANDPTAVALKYNADATKSDRLAAARPGLPPAEQHCANCQFHLDDVAGATEEWHGCSLF PGKLINVDGWCASWTLKAG
B3EBZ4	EAPANAVTMDDPTAQALKYHPSAADSDRVAAARPGLPPEEQHCANCNFMQADVGEGDYKGCQLFP GKLINVNGWCASWTLKAG
P81492	MFSLKLFVVFLAVCICVSQAVSYTDCTESGQNYCLCVGSNVCGEGKNCQLSSSGNQCVHGEGTPKPKS QTEGDFEEIPDEDILN
Q9EWK0	MSKKTFEELFTELQHKAANGDPATSRTAELVDKGVHAIGKKVVEEAAEVWMAAEYEGKDAAAEEIS QLLYHVQVMMVARGISLDDVYAHL
P15249	AETRNFALRDKKGNEIGVFTGKQPRQAALKAANRGHKDIRLRERGTKKVHIFAGERVKVKKPKGAP AWMPNEIWKPKVKKIGVEKLDEI
P15250	AEMRNFALRDAQGNEIGVFTGKSPRQAALKAANRGYTEIKLRERGTKKVHIFSGERVQVDKPAGAPA WMPDKIWKPKVKKEGIEKLD
P15251	MIEKRNFALRDKEGNEIGVFSGKQPRQAALKAANRGFTDIRLRERGTKKVHIFQGERIQVPKPSNAPK WMPANIWKPNVKKLGVEKLEDI
P12770	SNTRNFVLRDEDGNEHGVFTGKQPRQAALKAANRGSGTKANPDIIRLRERGTKKVHVFKAWKEIVDA PKNRPAWMPEKISKPFVKKERIEKLE
P40625	GKVKDDKPAPPKRPLSAFFLFKQHNYDQVKKENPNAKITELTSMIAEKWKHVTEKEKKKYEGLQQE AKAKYEKDMQAYEKKYGKPEKVKKIKKSKKGSK
P0ABS8	MLKNLAKLDQTEMDKVNVDLAAAGVAFKERYNMPVIAEAVEREQPEHLRSWFRERLIAHRLASVNLS RLPYEPKLK
Q8R1H0	MSAQTASGPTEDQVEILEYNFNKVNKHPDPTTLCLIAAEAGLTEEQTQKWFKQRLAEWRRSEGLPSE CRSVTD
P74977	MKKIILALVLMLFSFCTLGQETASMHLDDTLSAPIAAEINRKACDTQTPSPSEENDDWCCEVCCNPAC AGC
P84261	QVNFSPNW
P84260	QVNFSPNW
P84259	QVNFSPNW
P85855	QVNFSPNW
P85570	QLNFSPNW
P85575	QLNFSPNW

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P85597	QVNFSPNW
P85585	QVNFSPNW
P85591	QVNFSPNW
P85612	QLNFSPNW
P85634	QVNFSPNW
P16353	QLTFSSGWGN
P81626	QVTFSRDWSP
P85682	QVNFSPNW
P85848	QVNFSPNW
P85703	QVNFSPNW
P85711	QVNFSPNW
P85715	QVNFSPNW
P85742	QLNFSPNW
P85752	QVNFSPNW
P85760	QVNFSPNW
P85756	QVNFSPNW
P67789	QLNFSPNW
P85791	QLNFSPNW
P67790	QLNFSPNW
P81712	STRKTSWPELVGVTAEEAEKIKEEMSGVEIQVVPPGSFVTADYKPQRVRLYVDESNKVTRTPGIG
P16063	MRSMEGSVPKYPEPTEGSIGASGAKRSWPEVVGMSAEKAKEIILRDKPDAQIEVIPVDAMVPLDFNPN RIFILVAVARTPTVG
P82977	MSSVVKKPLGGNTDTGDHHNQKTEWPELVGKSVEEAKKVILQDKSEAQIVVLPVGTIVTMEYRIDRV RLFVDSLDKIAQVPRVG
P65119	MAKQDVIELEGTVLDTLPNAMFKVELENGHEILAHVSGKIRMNYIRILPGDKVTVEMSPYDLTRGRIT YRYK
P06294	AGQNTES
P0A6X7	MALTKAEMSEYLFDKLGLSKRDAKELVELFFEEIRRALENGEQVKLSGFGNFDLRDKNQRPGRNPKT GEDIPITARRVVTFRPGQKLKSRVENASPKDE
P30787	MSEKTLTRMDLSEAVFREVGLSRNESAQLVETVLQHMSDALVRGETVKISSFGTFSVRDKTSRMGRNP KTGEEVPISPRRVLSFRPSHLMKDRVAERNAK
P0A6Y1	MTKSELIERLATQQSHIPAKTVEDAVKEMLEHMASTLAQGERIEIRGFGSFSLHYRAPRTGRNPKTGD KVELEGKYVPHFKPGKELRDRANIYG
P04482	MELKHSISDYTEAEFLEFVKKICRAEGATEEDDNKLVREFERLTEHPDGSDLIYYPRDDREDSPEGIVK EIKEWRAANGKSGFKQG
P09881	MELKNSISDYTETEFKKIIEDIINCEGDEKKQDDNLEHFISVTEHPSGSDLIYYPEGNNDGSPEAVIKEIK EWRAANGKSGFKQG
P13479	MELKHSISDYTEAEFLQLVTTICNADTSSEEELVKLVTHFEEMTEHPSGSDLIYYPKEGDDDSPSGIVNT VKQWRAANGKSGFKQG
Q9Y5U9	MAFTLYSLLQAALLCVNAIAVLHEERFLKNIGWGTDQGIGGFGEEPGIKSQLMNLIRSVRTVMRVPLII VNSIAIVLLLLFG
Q6UWN8	MKLSGMFLLLSLALFCFLTGVFSQGGQVDCGEFQDPKVYCTRESNPHCGSDGQTYGNKCAFCKAIVK SGGKISLKHPGKC
P58062	MKITGGLLLLCTVVYFCSSSEAASLSPKKVDCSIYKKYPVVAIPCPITYLPVCGSDYITYGNECHLCTES LKSNGRVQFLHDGSC
P42993	CYISNCPIG
P19873	SSCPGKSSWPHLVGVGGSVAKAIIERQNPNVKAVILEEGTPVTKDFRCNRVRIWVNKRGLVVSPPRIG
B1P1E0	MKVSVLITLAVLGVMFVWASAAELEERGSDQRDSPAWLKSMERIFRSEERECRKMFGGCSVDSDCCA HLGCKPTLKYCAWDGTFGK
B1P1F4	MKVSVVITLAVLGIMFVWASAAELEERGSDQRDSPAWLKSMERIFQSEERECTKFLGGCSEDSECCPH LGCKDVLYYCAWDGTFGK
B1P1F3	MKVSVVITLAVLGVMFVWASAAELEERGSDQRDSPAWIKSMERIFQSEERECTKFLGGCSEDSECCPH LGCKDVLYYCAWDGTFGK
P0CH56	MKVSVVITLAVLGVMFVWASAAELKERGSDQRDSPAWIKSMERIFQSEERECTKFLGGCSEDSECCPH LGCKDVLYYCAWDGTFGK

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
B1P1G2	MKVSVLITLAVLGVMFLLTSAEERGSDQMDSPAWLKSMEIIFQSEERECRWLFGGCEKDSDCCEHLGC RRAKPSWCGWDFTVGK
B1P1G0	MKVSVLITLAVLGVMFLFTSAEERGSDQMDSPAWLKSMEIIFQSEERECRWLFGGCEKDSDCCEHLGC RRAKPSWCGWDFTVGK
B1P1G3	MKVSVLITLAVLGVMFLLTSAEERGSDQMDSPAWLKSMERIFQSEERECRWLFGGCEKDSDCCEHLG CRRAKPSWCGWDFTVGK
B1P1B2	MNATIFAFLLLLLNLAMHNATEQSSETDMDDTLLIPEINRGRCIEEGKWCPKKAPCCGRLECKGPSPKQ KKCTRP
B1P1B1	MNATIFALLLLLNLAMHNAAEQSSETDMDDTLLIPEINRGRCIEEGKWCPKKAPCCGRLECKGPSPKQ KKCTRP
B1P1E4	MKVSVLITLAVLGVMFVWASAAELEERGSDHRDSPAWLKSMERIFQSEERECRKMFGGCSVHSDCCA HLGCKPTLKYCAWDGTFGK
B1P1F7	MKVLVLITLAVLGAMFVWTSAAELEERGSDQRDSPAWVKSMERIFQSEERACREWLGGCSKDADCC AHLECRKKWPYHCVWDWTVRK
B1P1B3	MNATIFALLLLLNLAMYNAAEQSSETDMDDTLLIPENYRKGCFKEGHSCPKTAPCCRPLVCKGPSPNT KKCTRP
P82688	GAQFSSWG
P0DJ39	MKSTLMTASLLILVLLSIIDYASVYAEFIDSEISLERQWINACFNVCMKISSDKKYCKYLCGKS
P0DJ40	MKSTLMTASLLILVVLFIIDYASVYAEFIDGEISLERERDIPCFETCMKLYHIPKLCYIKCRKH
Q25298	MATTYEEFSAKLDRLDEEFNRKMQEQNAKFFADKPDESTLSPEMKEHYEKFERMIREHTEKFNKKM HEHSEHFKQKFAELLEQQKAAQYPSK
Q36736	MATTYEEFSAKLDRLDQEFNRKMQEQNAKFFADKPDESTLSPEMREHYEKFERMIKEHTEKFNKKM HEHSEHFKQKFAELLEQQKAAQYPSK
P69301	MATTYEEFAAKLDRLDAEFAKKMEEQNKRFFADKPDEATLSPEMKEHYEKFEKMIQEHTDKFNKKM REHSEHFKAKFAELLEQQKNAQFPGK
P07521	SCYDLCQPCGPTPLANSCNEPCVRQCQDSRVVIQPSPVVVTLPGPILSSFPQNTAVGSTSAAVGSILSEQ GVPISSGGFSLSGLGGRSYSRYLPC
P19987	GSGFSSWG
P19988	QSSFHSWG
P19990	GADFYSWG
P21144	QDVDHVFLRF
P02869	VTSYTLNEVVPLKDVVPEWVRIGFSATTGAEFAAHEVLSWSFHSELGGTSGSQK
P07444	VTSYTLNEIVPLKDVVPEWVRIGFSATTGAEFAAHEVLSWSFHSELEETSASKQ
P02868	VTSYTLSDVVPLKDVVPEWVRIGFSATPGAEYAAHEVLSWSFHSELSGTSSKQ
P16350	SVTSYGLSAVVPLKDVVPEWVRIGFSATTGAEYAAHEVLSWSFHSELGGTSS
P35101	MNQARIWTVVKPTVGLPLLLGSVTVIAILVHFAVLSHTTWFSKYWNGKAAAIESSVNVG
P0C0X9	MSKFYKIWMIFDPRRVFVAQGVFLFLLAVMIHLILLSTPSYNWLEISAAKYNRVAVAE
P80589	SAPAQWKLWLVMDPRTVMIGTAAWLGVLALLIHFLLLGTERFNWIDTGLKEQKATAAAQA
P80259	MWKVWLLFDPRRTLVALFTFLFVLALLIHFILLSTDRFNWMQGAPTAPAQTS
P02947	MWRIWQLFDPRQALVGLATFLFVLALLIHFILLSTERFNWLEGASTKPVQTSMVMPSSDLAV
P35106	MADKTLTGLTVEESEELHKHVIDGTRIFGAIAIVAHFLAYVYSPWLH
P0C0Y1	MADKSDLGYTGLTDEQAQELHSVYMSGLWPFSAVAIVAHLAVYIWRPWF
P07368	MTDDKAGPSGLSLKEAEEIHSYLIDGTRVFGAMALVAHILSAIATPWLG
P95654	MTDDMDKVWPTGLTLAEAEEVHKQLIDGTRVFGAIALFAHFLAAIATPWLG
P35109	MVDDPNKVWPTGLTIAESEELHKHVIDGSRIFVAIAIVAHFLAYVYSPWLH
P35099	AEDRSSLSGVSDAEAKEFHALFVSSFTAFIVIAVLAHVLAWAWRPWIPGPKGWA
P0C190	EVKQESLSGITEGEAKEFHKIFTSSILVFFGVAAFAHLLVWIWRPWVPGPNGYS
P08947	MSAVPFTRVLLISGFLAHLLLSTFVTLTVCKEVTEESDDLSKRNVLQRQLWAVGSFMGKKSLENTNRR SDEDMEISALFRGSPLKVKRSD
P67803	QSDDYGHMRF
P67802	QSDDYGHMRF
Q9Y4Y9	MAANATTNPSQLLPLELVDKCIGSRIHIVMKSDKEIVGTLLGFDDFVNMVLEDVTEFEITPEGRRITKL DQILLNGNNITMLVPGGEGPEV
O42978	MSMTILPLELIDKCIGSNLWVIMKSEREFAGTLVGFDDYVNIVLKDVTEYDTVTGVTEKHSEMLLNGN GMCMLIPGGKPE

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P62312	MSLRKQTPSDFLKQIIGRPVVVKLNSGVDYRGVLACLDGYMNIALEQTEEYVNGQLKNKYGDAFIRG NNVLYISTQKRRM
P64461	MHVTLVEINVHEDKVDEFIEVFRQNHLGSVQEEGNLRFDVLQDPEVNSRFYIYEAYKDEDAVAFHKTT PHYKTCVAKLESLMTGPRKKRLFNGLMP
Q7CG46	MHVTLVEINVKEDKVDQFIEVFRANHLGSIREAGNLRFDVLRDEHIPTRFYIYEAYTDEAAVAIHKTTP HYLQCVEQLAPLMTGPRKKTVFIGLMP
Q5SH22	MVGTCPECGAELRLENPELGELVVCEDCGAELEVVGLDPLRLEPAPEEAEDWGE
P03492	MSLLTEVETPTRNGWECRCNDSSDPLIIAASIIGILHLILWILNRLFFKCIYRRLKYGLKRGPSTEGVPES MREEYRQEQQSAVDVDDGHFVNIELE
O05821	MSTNPFDDDNGAFFVLVNDEDQHSLWPVFADIPAGWRVVHGEASRAACLDYVEKNWTDLRPKSLRD AMVED
P26886	MNKLAILAIIAMVLFSANAFRLQSRLRSNMEASARDPMTCEQAMASCEHTMCGYCQGPLYMTCIGITT DPECGLP
P58152	MSLIELLFGRKQKTATVARDRLQIIIAQERAQEGQTPDYLPTLRKELMEVLSKYVNVSLDNIRISQEKQ DGMDVLELNITLPEQKKV
P42984	AYNGPLA
P81034	RRINNDCQNFIGNRAMYEKVDWICKDCANIFRKDGLLNNCRSNCFYNTEFLWCIDATENTRNKEQLE QWAAILGAGWN
P81035	RRINNDCQNFIGNRAMYEKVDWICKDCANIFRQDGLLNNCRSNCFYNTEFLWCIDATENTRNKEQLE QWAAILGAGWN
P19852	DSDSAHLIG
P19962	DSDSAQNLIG
P14948	TKVDLTVEKGSDAKTLVLNIKYTRPGDTLAEVELRQHGSEEWEPMTKKGNLWEVKSAKPLTGPMNF RFLSKGGMKNVFDEVIPTAFTVGKTYTPEYN
O05228	MFTLILQIALGIMAVSTFLYVIRVIKGPTVPDRVVALDAIGINLIAITALVSILLKTSAFLDIILLLGILSFIG TIAFSKFLEKGEIIENDRNR
P19466	MNQKHSSDFVVIKAVEDGVNVIGLTRGTDTKFHHSEKLDKGEVIIAQFTEHTSAIKVRGEALIQTAYGE MKSEKK
Q9X6J6	MYTNSDFVVIKALEDGVNVIGLTRGADTRFHHSEKLDKGEVLIAQFTEHTSAIKVRGKAYIQTRHGVIE SEGKK
P80656	MDGKAPAAFVEPGEFNEVMKRLDQIDEKVEFVNSEVAQRIGKKVGRDIGILYGGVIGLLLFLIYVQISS MFM
P0A223	MSVTVPNDDWTLSSLSETFDDGTQTLQGELTLALDKLAKNPSNPQLLAEYQSKLSEYTLYRNAQSNTV KVIKDVDAAIIQNFR
Q8DKZ3	MAVSTELLVLGVYGALAGLYLLVVPAIVYAYLNARWYVASSFERAFMYFLVTFFFPGLLLLAPFINFR PQPRSLNS
O15239	MWFEILPGLSVMGVCLLIPGLATAYIHRFTNGGKEKRVAHFGYHWSLMERDRRISGVDRYYVSKGLE NID
O43678	MAAAAASRGVGAKLGLREIRIHLCQRSPGSQGVRDFIEKRYVELKKANPDLPILIRECSDVQPKLWAR YAFGQETNVPLNNFSADQVTRALENVLSGKA
Q9CQ75	MAAAAASRAVGAKLGLREIRVHLCQRSPGSQGVRDFIVQRYVELKKAHPNLPILIRECSEVQPKLWAR YAFGQEKTVSLNNLSADEVTRAMQNVLSGKA
O95167	MAARVGAFLKNAWDKEPVLVVSFVVGGLAVILPPLSPYFKYSVMINKATPYNYPVPVRDDGNMPDV PSHPQDPQGPSLEWLKKL
Q9CQ91	MAGRISAFLKNAWAKEPVLVVSFSVWGLAIIMPMISPYTKYASMINKATPYNYPVPVRDDGNMPDVP SHPQDPLGPSLDWLKNL
O00483	MLRQIIGQAKKHPSLIPLFVFIGTGATGATLYLLRLALFNPDVCWDRNNPEPWNKLGPNDQYKFYSVN VDYSKLKKERPDF
Q62425	MLRQILGQAKKHPSLIPLFVFIGAGGTGAALYVMRLALFNPDVSWDRKNNPEPWNKLGPNEQYKFYS VNVDYSKLKKEGPDF
Q02378	MNLLQVVRDHWVHVLVPMGFVFGYYLDRKNDEKLTAFRNKSLLYKRELKPNEEVTWK
O43676	MAHEHGHEHGHHKMELPDYRQWKIEGTPLETIQKKLAAKGLRDPWGRNEAWRYMGGFAKSVSFSD VFFKGFKWGFAAFVVAVGAEYYLESLNKDKKHH
P85527	MAIFCNNVLAALPTQCNPGFLDDLPPRIRKVCVALSRIYELGSEMESYIGDKENHITGFHESIPLLDSGV KRQDVDHVFLRFGRRR
P85816	QDLDHVFMRF
P69056	CYIQNCPLG
P69044	CYIQNCPLG
P69043	CYIQNCPLG

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P69057	CYIQNCPLG
P40632	MAGASDRTGVRRPRKAKKDPNAPKRALSSYMFFAKEKRVEIIAENPEIAKDVAAIGKMIGAAWNALS DEEKKPYERMSDEDRVRYEREKAEYAQRKV
P11632	MVTPREPKKRTTRKKKDPNAPKRALSAYMFFANENRDIVRSENPDITFGQVGKKLGEKWKALTPEE KQPYEAKAQADKKRYESEKELYNATLA
Q6Q547	MHLMYTLGPDGKRIYTLKKVTESGEITKSAHPARFSPDDKYSRQRVTLKKRFGLVPGQ
Q7Q7R8	MASGTFTQRLLVALMIFALIADLSTLVAARPQDSDAASVAAAIRYLQELETKHAQHARPRFGKRGGYL NPAIFGQDEQEVDWQDSTFSR
Q03056	MNRLLQRQLFLENLLVGVNSTFHQMQKHSINTCCRSLQRILDHLILLQTIHSPVFRLDRMQLRQMQTLA CLWIHRRNHDLQVTLGAIKWISP
P03901	MPLIYMNIMLAFTISLLGMLVYRSHLMSSLLCLEGMMLSLFIMATLMTLNTHSLLANIVPIAMLVFAACE AAVGLALLVSISNTYGLDYVHNLNLLQC
P60615	MKTLLLTLVVVTIVCLDLGYTIVCHTTATSPISAVTCPPGENLCYRKMWCDAFCSSRGKVVELGCAAT CPSKKPYEEVTCCSTDKCNPHPKQRPG
P60616	MKTLLLTLVVVTIVCLDLGYTIVCHTTATSPISAVTCPPGENLCYRKMWCDVFCSSRGKVVELGCAAT CPSKKPYEEVTCCSTDKCNPHPKQRPG
Q53B59	MKTLLLTLVVMTIVCLDLGYSLICFISPHDSVTCAPGENVCFLKSWCDAWCGSRGKKLSFGCAATCPK VNPGIDIECCSTDNCNPHPKLRP
P82662	MKTLLLTLVVMTIVCLDLGYTLICFISSHDSVTCAPGENVCFLKSWCDAWCGSRGKKLSFGCAATCPK VNPGIDIECCSTDNCNPHPKLRP
Q7T3J2	MKTLLLTLVVVTIVCLDLGYTIVCHTTATSPISAVTCPPGENLCYRKMWCDAFCSSRGKVVELGCAAT CPSKKPYEEVTCCSTDKCNPHPKQRPG
P10457	MKTLLLTLVVVTVVCLDLGYTRRCYNQQSSQPKTTKSCPPGENSCYNKQWRDHRGSITERGCGCPTV KPGIKLRCCESEDCNN
P60775	MKTLLLTLVVVTIVCLDLGYTRICFNHQSSQPQTTKTCSPGESSCYNKQWSDFRGTIIERGCGCPTVKP GIKLSCCESEVCNN
P10459	MKTLLLTLVVVTIVCLDLGYTRRCFNHPSSQPQTNKSCPPGENSCYNKQWRDHRGTITERGCGCPQV KSGIKLTCCQSDDCNN
Q90VW1	MKTLLLTLVVVTIVCLDLGYTRICFNHQSSQPQTTKTCSPGESSCYHKQWSDFRGTIIERGCGCPTVKP GIKLSCCESEVCNN
Q7T2I5	MKTLLLTLVVVTIVCLDLGYTRICFNHQSSQPQTTKTCSPGESSCYHKQWSDFRGTIIERGCGCPTVKP GINLSCCESEVCNN
P10456	MKTLLLTLVVVTMVCLDLGYTRRCFNQQSSQPKTTKSCPPGENSCYNKQWRDHRGSITERGCGCPK VKPGIKLRCCESEDCNN
Q9BPB1	MQKLIILLLVAAVLMSTQALFQEKRPMKKIDFLSKGKTDAEKQQKRSCSDDWQYCESPTDCCSWDCD VVCSG
P0CJ21	MEKLIILLLVAAVLMSTQALFQEKRTMKKIDFLSKGKADAEKQRKRNCSDDWQYCESPSDCCSWDCD VVCSG
P85069	DEVKIVLD
P85159	HGGYKPTDK
P86314	YLDHGLGAPAPYVDPLEPKREVCELNPDCDELADQMGFQEAYRRFYGTT
P86313	YLDHGLGAPAPYVDPLEPKREVCELNPDCDELADQMGFQEAYRRFYGTT
P02822	MKAAALLLLAALLTFSLCRSAPDGSDARSAKAFISHRQRAEMVRRQKRHYAQDSGVAGAPPNPLEAQ REVCELSPDCDELADQIGFQEAYRRFYGPV
P15504	SFAVGSSYGAAPDPLEAQREVCELNPDCDELADHIGFQEAYRRFYGPV
P84349	MRALTLLALLALAALCIAGQAGAKPSGAESSKGAAFVSKQEGSEVVKRPRRYLYQWLGAPVPYPDPL EPRREVCELNPDCDELADHIGFQEAYRRFYGPV
P02818	MRALTLLALLALAALCIAGQAGAKPSGAESSKGAAFVSKQEGSEVVKRPRRYLYQWLGAPVPYPDPL EPRREVCELNPDCDELADHIGFQEAYRRFYGPV
P86315	YLDHGLGAPAPYVDPLEPKREVCELNPDCDELADQMGFQEAYRRFYGTT
P84348	MRALTLLALLALAALCIAGQAGAKPSGAESSKGAAFVSKQEGSEVVKRPRRYLYQWLGAPVPYPDTL EPRREVCELNPDCDELADHIGFQEAYRRFYGPV
P84350	MRALTLLALLALAALCITGQAGAKPSGADSSKGAAFVSKQEGSEVVKRPRRYLYQWLGAPVPYPDPL EPKREVCELNPDCDELADHIGFQEAYRRFYGPV
P04640	MRTLSLLTLLALTAFCLSDLAGAKPSDSESDKAFMSKQEGSKVVNRLRRYLNNGLGAPAPYPDPLEPH REVCELNPNCDELADHIGFQDAYKRIYGTTV
P40148	MKTLAFLVLCSLAAICLTSDASTGSQPASDNPADEGMFVERDQASAVVRQKRAAGQLSLTQLESLREV CELNLACEHMMDTEGIIAAYTAYYGPIPY
P42985	IAYKPE

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P42996	CYINNCPVG
P42999	CYINNCPLG
P42995	CYIQSCPIG
P42998	CFVRNCPTG
P69058	CYIQNCPLG
P80027	CYFRNCPIG
P42994	CYISNCPQG
P43000	CYIQNCPVG
P01299	MPAACRCLFLLLLSACVALLLQPPLGTRGAPLEPVYPGDDATPEQMAQYAAELRRYINMLTRPRYGK RDRGEMRDILEWGSPHAAAPRELMDE
P67298	MGKNTSFVLDEHYSAFIDGEIAAGRYRSASEVIRSALRLLEDRETQLRALREALEAGERSGSSTPFDFD GFLGRKRADASRGR
Q9A9T8	MKPYRLSRRAKADLDDIWTYSEQRWGVEQAADYARELQATIEMIAEHPGMGQPDENLRAGYRRCAS GSHVVFYRVGVRVEIIRVLHQSMNARAHLG
P0C0U1	MGQEQDTPWILSTGHISTQKRQDGQQTPKLEHRNSTRLMGHCQKTMNQVVMPKQIVYWKQWLSLR NPILVFLKTRVLKRWRLFSKHE
P0C0U0	MEQEQDTPWTQSTEHINIQKKGGGQQTQRPEHPNSTLLMDHYLKITSRAGMHKQIVYWKQWLSLKN PTQDSLKTRVLKRWKLSSKREWIS
P0A1C7	MQQEALGMVETKGLTAAIEAADAMVKSANVMLVGYEKIGSGLVTVIVRGDVGAVKAATDAGAAAAR NVGEVKAVHVIPRPHTDVEKILPKGISQ
P84863	DILRG
P61046	MDIVSLAWAALMVVFTFSLSLVVWGRSGL
P23815	KNGDLRTPVITIFDARGCKDHANKEYTGPKAGGADDEMCVKVAMQKIAVAEDAAALVLKECLSELK ARKK
Q7M4T5	SAGKFIVIFKNGVSDDKIRETKDEVIAEGGTITNEYNMPGMKGFAGELTPQSLTKFQGLQGDLIDSIEE DGIVTTQ
P35477	IEVLLGSDDGGLAFVPGNFSVSAGEKITFKNNAGFPHNVVFDEDEIPAGVDASKISMSEEDLLNGPGET YSVTLSEKGTYTFYCAPHQGAGMVGKVTVN
P00294	IEVLLGGGDGSLAFVPNDFSIAKGEKIVFKNNAGFPHNVVFDEDEIPSGVDASKISMDENDLLNAAGET YEVALTEAGTYSFYCAPHQGAGMVGKVTVN
P00300	DVTVKLGADSGALVFEPSSVTIKAGETVTWVNNAGFPHNIVFDEDEVPSGANAEALSHEDYLNAPGES YSAKFDTAGTYGYFCEPHQGAGMKGTITVQ
P00292	IEVLLGGDDGSLAFIPNDFSVAAGEKIVFKNNAGFPHNVVFDEDEIPSGVDAGKISMNEEDLLNAPGEV YKVNLTEKGSYSFYCSPHQGAGMVGKVTVN
P20422	AEVKLGADDGALVFSPSSFSVAKGEGISFKNNAGFPHNIVFDEDEVPAGVDVSKISQEDYLDGAGESFT VTLTEKGTYKFYCEPHAGAGMKGEVTVT
P00290	AEVLLGSSDGGLVFEPSTFSVASGEKIVFKNNAGFPHNVVFDEDEIPAGVDASKISMSEEDLLNAPGET YAVTLTEKGTYSFYCAPHQGAGMVGKVTVN
P00295	LDVLLGSDDGELAFVPNNFSVPSGEKITFKNNAGFPHNVVFDEDEIPSGVDASKISMDEADLLNAPGET YAVTLTEKGSYSFYCSPHQGAGMVGKVTVN
P17341	AEVKLGSDDGGLVFSPSSFTVAAGEKITFKNNAGFPHNIVFDEDEVPAGVNAEKISQPEYLNGAGETY EVTLTEKGTYKFYCEPHAGAGMKGEVTVN
P00287	LEVLLGSGDGSLVFVPSEFSVPSGEKIVFKNNAGFPHNVVFDEDEIPAGVDAVKISMPEEELLNAPGET YVVTLDTKGTYSFYCSPHQGAGMVGKVTVN
P00298	IEIKLGGDDGALAFVPGSFTVAAGEKIVFKNNAGFPHNIVFDEDEVPAGVDASKISMSEEDLLNAPGET YAVTLSEKGTYSFYCSPHQGAGMVGKVTVQ
P00291	VEILLGGEDGSLAFIPSNFSVPSGEKITFKNNAGFPHNVVFDEDEVPSGVDSAKISMSEDDLLNAPGETY SVTLTESGTYKFYCSPHQGAGMVGKVTVN
P00297	IEVLLGSDDGGLAFVPGNFSISAGEKITFKNNAGFPHNVVFDEDEIPAGVDASKISMPEEDLLNAPGET YSVTLSEKGTYSFYCSPHQGAGMVGKVTVN
P00296	LDVLLGGDDGSLAFIPGNFSVSAGEKITFKNNAGFPHNVVFDEDEIPAGVDASKISMAEEDLLNAAGET YSVTLSEKGTYTFYCAPHQGAGMVGKVTVN
P13133	AQIVKLGGDDGALAFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDAVPAGVDADAISYDDYLNSKGET VVRKLSTPGVYGVYCEPHAGAGMKMTITVQ
P56274	AQIVKLGGDDGSLAFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDAVPAGVDADAISYDDYLNSKGETV VRKLSTPGVYGVYCEPHAGAGMKMTITVQ
P00288	VEVLLGASDGGLAFVPNSFEVSAGDTIVFKNNAGFPHNVVFDEDEIPSGVDAAKISMPEEDLLNAPGE TYSVKLDAKGTYKFYCSPHQGAGMVGQVTVN
O00168	MASLGHILVFCVGLLTMAKAESPKEHDPFTYDYQSLQIGGLVIAGILFILGILIVLSRRCRCKFNQQQRT GEPDEEEGTFRSSIRRLSTRRR

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
Q9Z239	MASPGHILALCVCLLSMASAEAPQEPDPFTYDYHTLRIGGLTIAGILFILGILIILSKRCRCKFNQQQRTG EPDEEEGTFRSSIRRLSSRRR
O08589	MASPGHILIVCVCLLSMASAEAPQEPDPFTYDYHTLRIGGLTIAGILFILGILIILSKRCRCKFNQQQRTG EPDEEEGTFRSSIRRLSTRRR
P32903	MTLPGGVILVFILVGLACIAIIATIIYRKWQARQRGLQRF
Q8VWY6	MAEDPQDIADRERIFKRFDLNGDGKISSAELGETLKMLGSVTSEEVQHMMAELDTDGDGFISYEEFEE FARANRGLIKDVAKVF
Q84V36	MAAEDTPQDIADRERIFKRFDTNGDGKISSSELGDALKTLGSVTPDEVRRMMAEIDTDGDGFISFDEF TDFARANRGLVKDVSKIF
O81092	MADDPQEVAEHERIFKRFDANGDGKISSSELGETLKTLGSVTPEEIQRMMAEIDTDGDGFISFEEFTVF ARANRGLVKDVAKIF
O82040	MADDMERIFKRFDTNGDGKISLSELTDALRTLGSTSADEVQRMMAEIDTDGDGFIDFNEFISFCNANP GLMKDVAKVF
P61012	MDKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLILICLLLICIIVMLL
P26678	MEKVQYLTRSAIRRASTIEMPQQARQKLQNLFINFCLILICLLLICIIVMLL
P61014	MEKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLILICLLLICIIVMLL
P61013	MDKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLILICLLLICIIVMLL
P61015	MEKVQYLTRSAIRRASTIEMPQQARQNLQNLFINFCLILICLLLICIIVMLL
P41784	MATPWSGYLDDVSAKFDTGVDNLQTQVTEALDKLAAKPSDPALLAAYQSKLSEYNLYRNAQSNTVK VFKDIDAAIIQNFR
P31090	MVQRGSKVRILRPESYWFQDVGTVASVDQSGIKYPVIVRFDKVNYSGINTNNFAVDELIEVEAPKAKP AKK
Q7NFW6	MAIERGAKVRILRKESYWYREVGTVASVDKSEKTIYPVTVRFEKVNYSGINTNNFGVSELEEVEA
Q9WWP1	MVQRGSKVRILRPESYWFQDVGTVASVDQSGIKYPVIVRFEKVNYSGINTNNFAEDELVEVEAPKAKP KK
P0A424	MVQRGSKVKILRPESYWYNEVGTVASVDQTPGVKYPVIVRFDKVNYTGYSGSASGVNTNNFALHEVQ EVAPPKKGK
P23077	MAIARGDKVRILRPESYWFNEVGTVASVDQSGIKYPVVVRFEKVNYNGFSGSDGGVNTNNFAEAELQ VVAAAAKK
P0A423	MVQRGSKVKILRPESYWYNEVGTVASVDQTPGVKYPVIVRFDKVNYTGYSGSASGVNTNNFALHEVQ EVAPPKKGK
P0A427	MMGSYAASFLPWIFIPVVCWLMPTVVMGLLFLYIEGEA
P17230	MRDFKTYLSVAPVLSTLWFGSLAGLLIEINRFFPDALTFPFFSF
P0A429	MKHFLTYLSTAPVLAAIWMTITAGILIEFNRFYPDLLFHPL
P0A426	MVLATLPDTTWTPSVGLVVILCNLFAIALGRYAIQSRGKGPGLPIALPALFEGFGLPELLATTSFGHLL AAGVVSGLQYAGAL
P0A425	MVLATLPDTTWTPSVGLVVILCNLFAIALGRYAIQSRGKGPGLPIALPALFEGFGLPELLATTSFGHLL AAGVVSGLQYAGAL
P56779	MSGSTGERSFADIITSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTESRQGIPLITGRFDS LEQLDEFSRSF
P13554	MSGSTGERSFADIITSIRYWIIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTETRQGIPLITGRFDS LEQLDEFSRSF
P69383	MSGSTGERSFADIITSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTESRQGIPLITGRFDS LEQLDEFSRSF
P09190	MSGTTGERPFSDIVTSIRYWVIHSITIPMLFIAGWLFVSTGLAYDAFGTPRPDEYFTQTRQELPILQERY DINQEIQEFNQ
P69386	MSGSTGERSFADIITSIRYWVIHSITIPSLFIAGWLFVSTGLAYDVFGSPRPNEYFTESRQGIPLITDRFDS LEQLDEFSRSF
P62096	MTIDRTYPIFTVRWLAVHGLAVPTVSFLGSISAMQFIQR
P60128	MTIDRTYPIFTVRWLAIHGLAVPTVFFLGSISAMQFIQR
P60125	MTIDRTYPIFTVRWLAIHGLAVPTVFFLGSISAMQFIQR
P56780	MATQTVEDSSRSGPRSTTVGKLLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLD GISVN
P05146	MATQTVESSSRSRPKPTTVGALLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDG ISMN
Q8DJ43	MARRTWLGDILRPLNSEYGKVAPGWGTTPLMAVFMGLFLVFLLIILEIYNSTLILDGVNVSWKALG

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P69555	MATQTVEDSSKPRPKRTGAGSLLKPLNSEYGKVAPGWGTTPFMGVAMALFAIFLSIILEIYNSSVLLDG ILTN
P62100	MLTLKLFVYTVVIFFVSLFIFGFLSNDPGRNPGREE
P62103	MLTLKLFVYTVVIFFVSLFIFGFLSNDPGRNPGREE
Q8DJZ6	METLKITVYIVVTFFVLLFVFGFLSGDPARNPKRKDLE
Q8DIN8	MEPNPNRQPVELNRTSLYLGLLLILVLALLFSSYFFN
P12241	MEPNPNRQPVELNRTSLYLGLLLILVLALLFSSYFFN
P37256	MEALVYTFLLVGTLGIIFFSIFFRDPPRMIK
Q8DIQ0	METITYVFIFACIIALFFFAIFFREPPRITKK
Q8DHJ2	MTILFQLALAALVILSFVMVIGVPVAYASPQDWDRSKQLIFLGSGLWIALVLVVGVLNFFVV
P09974	MTLAFQLAVFALIATSLILLISVPVVFASPDGWSSNKNVVFSGTSLWIGLVFLVGILNSLIS
Q0DAS9	MVNPGRTARALCLLCLALLLLGQDTHSRKLLLQEKHSHGVGNGTTTTQEPSRENGGSTGSNNNGQLQ FDSAKWEEFHTDYIYTQDVKKP
P58261	YIYTQ
O69250	MAQKTFTVTADSGIHARPATTLVQAASKFDSDINLEFNGKTVNLKSIMGVMSLGIQKGATITISAEGSD EADALAALEDTMSKEGLGE
Q9KJV3	MEKREFNIIAETGIHARPATLLVQAASKFNSDINLEYKGKSVNLKSIMGVMSLGVGQGADVTISAEGAD EADAIAAITDTMKKEGLAE
Q9CJ83	MASKEFHIVVETGIHARPATLLVHTASKFTSEITLEYKGKSVNLKSIMGVMSLGVGQGADVTISAEGAD ADDAISTIAETMTKEGLAE
Q84F84	MKTQQFTVIDPLGIHARPASQLVAKATPFASAIEVRTEEKAANLKSILGVMGLALKQGSQFTLYVEGE DEDQAFEALATLLTEMGLAQ
P75061	MKKIQVVVKDPVGIHARPASIIAGEANKFKSELKLVSPSGVEGNIKSIINLMSLGIKQNDHITIKAEGTDE EEALNAIKAVLEKHQVI
Q9WXK8	MASKDFHIVAETGIHARPATLLVQTASKFASDITLDYKGKAVNLKSIMGVMSLGVGQGADVTISAEGA DADDALAAIEETMTKEGLA
P45596	MASKDFHIVAETGIHARPATLLVQTASKFASDITLDYKGKAVNLKSIMGVMSLGVGQGADVTITAEGA DADDAIAAINETMTKEGLA
P24366	MASKDFHIVAETGIHARPATLLVQTASKFASDITLDYKGKAVNLKSIMGVMSLGVGQGADVTISAEGA DADDAIVAIAETMTKEGLA
P37188	MKRKIIVACGGAVATSTMAAEEIKELCQNHNIPVELIQCRVNEIETYMDGVHLICTTAKVDRSFGDIPL VHGMPFISGIGIEALQNKILTILQG
O33246	MAQEQTKRGGGGGDDDDIAGSTAAGQERREKLTEETDDLLDEIDDVLEENAEDFVRAYVQKGGQ
P20116	GRLFKITACVPSQTRIRTQRELQNTYFTKLVPYENWFREQQRIQKMGGKIVKVELATGKQGINTGLA
P11396	MFGQTTLGIDSVSSSASRVFRFEVVGMRQNEENDKNKYNIRRSGSVYITVPYNRMSEEMQRIHRLGGK IVKIEPLTRAAG
P10082	MVFVRRPWPALTTVLLALLVCLGALVDAYPIKPEAPREDASPEELNRYYASLRHYLNLVTRQRYGKR DGPDTLLSKTFFPDGEDRPVRSRSEGPDLW
P0A182	MSAVAGCTATTDPGWEVDAFGGVSSLCQPMEADLYGCSDPCWWPAQVPDMMSTYQDWNAQASNS AEDWRNLGTVFPKDK
P07919	MGLEDEQKMLTESGDPEEEEEEEEELVDPLTTVREQCEQLEKCVKARERLELCDERVSSRSHTEEDC TEELFDFLHARDHCVAHKLFNNLK
P99028	MGLEDERKMLTGSGDPKEEEEEELVDPLTTVREHCEQLEKCVKARERLELCDNRVSSRSQTEEDCTE ELFDFLHARDHCVAHKLFKNLK
O14949	MGREFGNLTRMRHVISYSLSPFEQRAYPHVFTKGIPNVLRRIRESFFRVVPQFVVFYLIYTWGTEEFER SKRKNPAAYENDK
Q8R1I1	MSSPTIPSRLYSLLFRRTSTFALTIAVGALFFERAFDQGADAIYEHINEGKLWKHIKHKYENKE
P82070	VDFFA
P82071	IEFFA
P82072	IEFFT
P82073	VGFFT
O50462	MAVVPLGEVRNRLSEYVAEVELTHERITITRHGHPAAVLISADDLASIEETLEVLRTPGASEAIREGLAD VAAGRFVSNDEIRNRYTAR
O33348	MPYTVRFTTTARRDLHKLPPRILAAVVEFAFGDLSREPLRVGKPLRRELAGTFSARRGTYRLLYRIDD EHTTVVILRVDHRADIYRR
P11952	RPNWEERSRLCGRDLIRAFIYLCGGTRWTRLPNFGNYPIMEEKMGFAKKCCAIGCSTEDFRMVC

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
A6MWS9	MKIIVFLAVLMLVSAQVCLVSAAEMEHSSDNELSSRDLVKRFFLPPCAHKGTCNH
Q9RXK0	MSHYDILQAPVISEKAYSAMERGVYSFWVSPKATKTEIKDAIQQAFGVRVIGISTMNVPGKRKRVGRFI GQRNDRKKAIVRLAEGQSIEALAGQA
P04454	MKDPRDIIKRPIITENTMNLIGQKKYTFEVDVKANKTEVKDAVEKIFGVKVEKVNIMNYKGKFKRVG RYSGYTNRRKKAIVTLTPDSKEIELFEV
Q06842	MSSIIDYPLVTEKAMDEMDFQNKLQFIVDIDAAKPEIRDVVESEYDVTVVDVNTQITPEAEKKATVKLS AEDDAQDVASRIGVF
Q6N4T7	MKSIDPRHYDVIVAPVVTEKSTMASEHNKVVFKVQGGATKPQIKEAVEKLFDVKVKSVNTLVRKGKT KAFRGTFGTQSDVKRAVVTLEEGHRIDVTTGL
Q99S23	MEARDILKRPVITEKSSEAMAEDKYTFDVDTRVNKTQVKMAVEEIFNVKVASVNIMNYKPKKKRMG RYQGYTNKRRKAIVTLKEGSIDLFN
Q7A459	MEARDILKRPVITEKSSEAMAEDKYTFDVDTRVNKTQVKMAVEEIFNVKVASVNIMNYKPKKKRMG RYQGYTNKRRKAIVTLKEGSIDLFN
Q72I06	MKTAYDVILAPVLSEKAYAGFAEGKYTFWVHPKATKTEIKNAVETAFKVKVVKVNTLHVRGKKKRL GRYLGKRPDRKKAIVQVAPGQKIEALEGLI
Q5SHP0	MKTAYDVILAPVLSEKAYAGFAEGKYTFWVHPKATKTEIKNAVETAFKVKVVKVNTLHVRGKKKRL GRYLGKRPDRKKAIVQVAPGQKIEALEGLI
Q9RA57	MKTAYDVILAPVLSEKAYAGFAEGKYTFWVHPKATKTEIKNAVETAFKVKVVKVNTLHVRGKKKRL GRYLGKRPDRKKAIVQVAPGQKIEALEGLI
P0C2N0	MIREERLLKVLRAPHVSEKASAAMEKNNTIVLKVAKDATKAEIKAAVQKLFEVEVEDVNTLLVKGKS KRHGQRVGRRSDWKKAYVTLKEGQNLDFIGGAE
P0A7L8	MAHKKAGGSTRNGRDSEAKRLGVKRFGGESVLAGSIIVRQRGTKFHAGANVGCGRDHTLFAKADGK VKFEVKGPKNRKFISIEAE
P07844	MASKKGVGSTKDGRDSIAKRLGAKRADGQFVTGGSILYRQRGTKVHPGLNVGRGGDDTLYAKIDGIV RFERLGRDRKRVSVYPVSQEA
Q9HVL7	MAHKKAGGSTRNGRDSESKRLGVKLFGGQAVKAGNILVRQRGTKFHAGYGVGLGKDHTLFAKVDG VVKFETKGAFGRKYVSIVAA
Q72HR3	MAHKKGLGSTKNGRDSQAKRLGVKRYEGQVVRAGNILVRQRGTRFKPGKNVGMGRDFTLFALVDG VVEFQDRGRLGRYVHVRPLA
P60493	MAHKKGLGSTRNGRDSQAKRLGVKRYEGQVVRAGNILVRQRGTRFKPGKNVGMGRDFTLFALVDG VVEFQDRGRLGRYVHVRPLA
Q9RRG8	MSRECYLTGKKNLVVNSVIRRGKARADGGVGRKTTGITKRVQRANLHKKAIRENGQVKTVWLSANA LRTLSKGPYKGIELI
P23374	MAKCFITGKKKSFGNTRSHAMNASRRDWKANLQKVRILVDGKPKRVWVSARALKSGKVKRV
P60494	MSKVCEISGKRPIVANSIQRRGKAKREGGVGKKTTGISKRRQYPNLQKVRVRVAGQEITFRVAASHIP KVYELVERAKGLKLEGLSPKEIKKELLKLL
P0A7M6	MKAKELREKSVEELNTELLNLLREQFNLRMQAASGQLQQSHLLKQVRRDVARVKTLLNEKAGA
P66173	MKAKEIRDLTTSEIEEQIKSSKEELFNLRFQLATGQLEETARIRTVRKTIARLKTVAREREIEQSKANQ
P38514	MKASELRNYTDEELKNLLEEKKRQLMELRFQLAMGQLKNTSLIKLTKRDIARIKTILRERELGIRR
P02431	MAKKLAITLTRSVIGRPEDQRITVRTLGLRKMHQTVVHNDNPAIRGMINKVAHLVKVKEIEEE
P0A0G0	MAKLQITLTRSVIGRPETQRKTVEALGLKKTNSSVVVEDNPAIRGQINKVKHLVTVEEK
Q72I22	MPRLKVKLVKSPIGYPKDQKAALKALGLRRLQQERVLEDTPAIRGNVEKVAHLVRVEVVE
Q5SHQ6	MPRLKVKLVKSPIGYPKDQKAALKALGLRRLQQERVLEDTPAIRGNVEKVAHLVRVEVVE
P74909	MPRLKVKLVKSPIGYPKDQKAALKALGLRRLQQERVLEDTPAIRGNVEKVAHLVRVEVVE
P66195	MKQGIHPEYHQVIFLDTTTNFKFLSGSTKTSSEMMEWEDGKEYPVIRLDISSDSHPFYTGRQKFAAAD GRVERFNKKFGLKSNN
P66196	MKQGIHPEYHQVIFLDTTTNFKFLSGSTKTSSEMMEWEDGKEYPVIRLDISSDSHPFYTGRQKFAAAD GRVERFNKKFGLKSNN
Q9HUD0	MKADIHPTYEAIEATCSCGNVIKTRSTLCKPIHLDVCSECHPFYTGKQKVLDTGGRIDRFKQRFGVFG ATK
A0R551	MARNEIRPIVKLRSTAGTGYTYVTRKNRRNDPDRIVLRKYDPVLRRHVEFREER
Q9RSS4	MAKDGPRIIVKMESSAGTGFYYTTTKNRRNTQAKLELKKYDPVAKKHVVFREKKV
Q6N554	MAKAVTIKIKLVSTADTGFYYVTKKNSRTMTDKMVKKKYDPVARKHVEFKEAKIK
P75447	MKVKSAAKKRFKLTKSGQIKRKHAYTSHLAPHKTTKQKRHLRKQGTVSASDFKRIGNLI
O14455	MAVKTGIAIGLNKGKKVTQMTPAPKISYKKGAASNRTKFVRSLVREIAGLSPYERRLIDLIRNSGEKRA RKVAKKRLGSFTRAKAKVEEMNNIIAASRRH
P61513	MAKRTKKVGIVGKYGTRYGASLRKMVKKIEISQHAKYTCSFCGKTKMKRRAVGIWHCGSCMKTVAG GAWTYNTTSAVTVKSAIRRLKELKDQ

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P59289	MTKGTQSFGMRHNKSHTICRRCGKRSFHIQKSTCACCGYPAAKTRSYNWGAKAKRRRTTGTGRMSY LKKVHRSFKNGFRAGKPTSAATA
P49166	MGKGTPSFGKRHNKSHTLCNRCGRRSFHVQKKTCSSCGYPAAKTRSYNWGAKAKRRHTTGTGRMR YLKHVSRRFKNGFQTGSASKASA
P05733	MTKGTQSFGMRHNKSHTICRRCGKRSFHIQKSTCACCGYPAAKTRSYNWGAKAKRRRTTGTGRMSY LKKVHRSFKNGFRSGKPAAAVAASA
P51402	MGKGTPSFGKRHNKSHTLCNRCGRRSFHVQKKTCSSCGYPSAKTRSHNWAAKAKRRHTTGTGRMR YLKHVSRRFKNGFQTGSAKATSA
P0DJ24	MTRGTPAFGKRHQKTHTLCRRCGKATYHKQKLRCAACGYPDAKMRRYDGWGQKVRDRKGQGTG RMRYMKTIARRAKNGFRSGTQAAPKVKAATN
O75394	MFLSAVFFAKSKSKNILVRMVSEAGTGFCFNTKRNRLREKLTLLHYDPVVKQRVLFVEKKKIRSL
Q9BQ48	MAVLAGSLLGPTSRSAALLGGRWLQPRAWLGFPDAWGLPTPQQARGKARGNEYQPSNIKRKNKHG WVRRLSTPAGVQVILRRMLKGRKSLSH
Q9VC49	MIIPIRCFTCGKVIGNKWESYLGLLQAEYTEGDALDALGLKRYCCRRMLLGHVDLIEKLLNYAPLEK
P62875	MIIPVRCFTCGKIVGNKWEAYLGLLQAEYTEGDALDALGLKRYCCRRMLLAHVDLIEKLLNYAPLEK
P22139	MIVPVRCFSCGKVVGDKWESYLNLLQEDELDEGTALSRLGLKRYCCRRMILTHVDLIEKFLRYNPLEK RD
Q980Q9	MRGSSNKKIDPRIHYLVPKHEVLNIDEAYKILKELGIRPEQLPWIRASDPVARSINAKPGDIIRIIRKSQLY GEVVSYRYVISG
Q8RQE7	MAEPGIDKLFGMVDSKYRLTVVVAKRAQQLLRHGFKNTVLEPEERPKMQTLEGLFDDPNAVTWAM KELLTGRLVFGENLVPEDRLQKEMERLYPVEREE
P06507	MARKSLIQREKKRRNLEQKYHLIRRSSKQEIRKVTSLSDKWEIHGKLQSPPRNSAPARLHRRCFLTGRP RANIRDFGLSGHILREMVHTCLLPGATRSSW
P59776	MSRSLKKGPFVADHLLKKIEKLNAKGKKVVIKTWSRSSMIVPPMIGHTIGVYNGREHIPVFVSDQMVG HRLGEFSPTRTYRGHAKKDKKAKR
P06508	MTRSLKKNPFVANHLLRKIEKLNKKAEKEIIVTWSRASTIIPTMIGHTIAIHNGREHLPIYITDRMVGHK LGEFAPTLNFRGHAKNDNKSRR
P21473	MAITQERKNQLINEFKTHESDTGSPEVQIAILTDSINNLNEHLRTHKKDHHSRRGLLKMVGKRRNLLTY LRNKDVTRYRELINKLGLRR
Q8X9M2	MSLSTEATAKIVSEFGRDANDTGSTEVQVALLTAQINHLQGHFAEHKKDHHSRRGLLRMVSQRRKLLD YLKRKDVARYTRLIERLGLRR
B5YS55	MSLSTEATAKIVSEFGRDANDTGSTEVQVALLTAQINHLQGHFAEHKKDHHSRRGLLRMVSQRRKLLD YLKRKDVARYTRLIERLGLRR
P0ADZ4	MSLSTEATAKIVSEFGRDANDTGSTEVQVALLTAQINHLQGHFAEHKKDHHSRRGLLRMVSQRRKLLD YLKRKDVARYTQLIERLGLRR
Q6NCN7	MSITAERKAEVIKTSATKAGDTGSPEVQVAILSERITNLTAHFKTHTKDNHSRRGLLKLVSTRRSLLDYI KKKDEARYKALLEKHNIRR
Q7A5X8	MAISQERKNEIIKEYRVHETDTGSPEVQIAVLTAEINAVNEHLRTHKKDHHSRRGLLKMVGRRRHLLN YLRSKDIQRYRELIKSLGIRR
P62657	MPITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLL RYLQREDPERYRALIEKLGIRG
Q5SJ76	MPITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLL RYLQREDPERYRALIEKLGIRG
P80378	MPITKEEKQKVIQEFARFPGDTGSTEVQVALLTLRINRLSEHLKVHKKDHHSHRGLLMMVGQRRRLL RYLQREDPERYRALIEKLGIRG
P21474	MAVKIRLKRMGAKKSPFYRIVVADSRSPRDGRFIETVGTYNPVAKPAEVKIDEELALKWLQTGAKPS DTVRNLFSSQGIMEKFHNAKQGK
P62238	MVKIRLARFGSKHNPHYRIVVTDARRKRDGKYIEKIGYYDPRKTTPDWLKVDVERARYWLSVGAQP TDTARRLLRQAGVFRQEAREGA
Q5SJH3	MVKIRLARFGSKHNPHYRIVVTDARRKRDGKYIEKIGYYDPRKTTPDWLKVDVERARYWLSVGAQP TDTARRLLRQAGVFRQEAREGA
P80379	MVKIRLARFGSKHNPHYPHYRIVVTDARRKRDGKYIEKIGYYDPRKTTPDWLKVDVERARYWLSVG AQPTDTARRLLRQAGVFRQEAREGA
P12874	MSERNQRKVYQGRVVSDKMDKTITVVVETYKKHTLYGKRVKYSKKFKAHDENNQAKIGDIVKIMET RPLSATKRFRLVEVVEEAVII
P0AG65	MTDKIRTLQGRVVSDKMEKSIVVAIERFVKHPIYGKFIKRTTKLHVHDENNECGIGDVVEIRECRPLSK TKSWTLVRVVEKAVL
Q1R616	MTDKIRTLQGRVVSDKMEKSIVVAIERFVKHPIYGKFIKRTTKLHVHDENNECGIGDVVEIRECRPLSK TKSWTLVRVVEKAVL

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P23828	MSERNQRKVYVGRVVSDKMDKTITVLVETYKKHPLYGKRVKYSKKYKAHDEHNEAKVGDIVKIMET RPLSATKRFRLVEIVEKAVVL
Q7A462	MSERNDRKVYVGKVVSDKMDKTITVLVETYKTHKLYGKRVKYSKKYKTHDENNSAKLGDIVKIQET RPLSATKRFRIVEIVEESVII
P21475	MAGGRRGGRAKRRKVCYFTSNGITHIDYKDVDLLKKFVSERGKILPRRVTGTNAKYQRKLTAAIKRA RQMALLPYVSGE
P0A7T7	MARYFRRRKFCRFTAEGVQEIDYKDIATLKNYITESGKIVPSRITGTRAKYQRQLARAIKRARYLSLLP YTDRHQ
P10806	MAGRKGGRGKRRKVCYFTANNITHIDYKDVDLLKKFISERGKILPRRVTGTSAKYQRKLTVAIKRAR QMALLPYVADE
P66468	MAGGPRRGGRRRKKVCYFTANGITHIDYKDTELLKRFISERGKILPRRVTGTSAKYQRMLTTAIKRSR HMALLPYVKEEQ
P80382	MSTKNAKPKKEAQRRPSRKAKVKATLGEFDLRDYRNVEVLKRFLSETGKILPRRRTGLSGKEQRILA KTIKRARILGLLPFTEKLVRK
P21476	MARSLKKGPFVDGHLMTKIEKLNETDKKQVVKTWSRRSTIFPQFIGHTIAVYDGRKHVPVFISEDMV GHKLGEFAPTRTYKGHASDDKKTRR
Q0TCE5	MPRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVFVTDEMVG HKLGEFAPTRTYRGHAADKKAKKK
P0A7U3	MPRSLKKGPFIDLHLLKKVEKAVESGDKKPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVFVTDEMVG HKLGEFAPTRTYRGHAADKKAKKK
P0A5X5	MPRSLKKGPFVDEHLLKKVDVQNEKNTKQVIKTWSRRSTIIPDFIGHTFAVHDGRKHVPVFVTESMV GHKLGEFAPTRTFKGHIKDDRKSKRR
Q6N4T8	MVRSVWKGPFVEASLLKKADAARASGRHDVIKIWSRRSTILPQFVGLTFGVYNGQKHVPVSVNEEMV GHKFGEFSPTRTFHGHAGDKKSKKG
P66494	MARSIKKGPFVDEHLMKKVEAQEGSEKKQVIKTWSRRSTIFPNFIGHTFAVYDGRKHVPVYVTEDMV GHKLGEFAPTRTFKGHVADDKKTRR
P62660	MPRSLKKGVFVDDHLLEKVLELNAKGEKRLIKTWSRRSTIVPEMVGHTIAVYNGKQHVPVYITENMV GHKLGEFAPTRTYRGHGKEAKATKKK
Q5SHP2	MPRSLKKGVFVDDHLLEKVLELNAKGEKRLIKTWSRRSTIVPEMVGHTIAVYNGKQHVPVYITENMV GHKLGEFAPTRTYRGHGKEAKATKKK
P80381	MPRSLKKGVFVDDHLLEKVLELNAKGEKRLIKTWSRRSTIVPEMVGHTIAVYNGKQHVPVYITENMV GHKLGEFAPTRTYRGHGKEAKATKKK
P21477	MPNIKSAIKRTKTNNERRVHNATIKSAMRTAIKQVEASVANNEADKAKTALTEAAKRIDKAVKTGLV HKNTAARYKSRLAKKVNGLSA
Q7A5C0	MANIKSAIKRVKTTEKAEARNISQKSAMRTAVKNAKTAVSNNADNKNELVSLAVKLVDKAAQSNLIHS NKADRIKSQLMTANK
P0C0V8	MENDKGQLVELYVPRKCSATNRIIKADDHASVQINVAKVDEEGRAIPGEYVTYALSGYVRSRGESDDS LNRLAQNDGLLKNVWSYSR
Q3E754	MENDKGQLVELYVPRKCSATNRIIKADDHASVQINVAKVDEEGRAIPGEYITYALSGYVRSRGESDDSL NRLAQNDGLLKNVWSYSR
P63220	MQNDAGEFVDLYVPRKCSASNRIIGAKDHASIQMNVAEVDKVTGRFNGQFKTYAICGAIRRMGESDD SILRLAKADGIVSKNF
P05765	MQNDAGEFVDLYVPRKCSASNRIIAAKDHASIQMNVAEVDRSTGRFNGQFKTYGICGAIRRMGESDDS ILRLAKADGIVSKNF
P66521	MSKTVVRKNESLEDALRRFKRSVSKSGTIQEVRKREFYEKPSVKRKKKSEAARKRKFK
Q71UM5	MPLARDLLHPSLEEEKKKHKKKRLVQSPNSYFMDVKCPGCYKITTVFSHAQTVVLCVGCSTVLCQPT GGKARLTEGCSFRRKQH
P24051	MPLARDLLHPSLEEEKKKHKKKRLVQSPNSYFMDVKCPGCYKITTVFSHAQTVVLCVGCSTVLCQPT GGKARLTEGCSFRRKQH
Q9TXP0	MPLAVDLLHPEPQREIRCHKLKRLVQHPNSYFMDVKCSGCFKISTVFSHATTVVVCVGCNTVLCQPT RGKAKLTEGCSFRKKQ
P42677	MPLAKDLLHPSPEEEKRKHKKKRLVQSPNSYFMDVKCPGCYKITTVFSHAQTVVLCVGCSTVLCQPT GGKARLTEGCSFRRKQH
Q6ZWU9	MPLAKDLLHPSPEEEKRKHKKKRLVQSPNSYFMDVKCPGCYKITTVFSHAQTVVLCVGCSTVLCQPT GGKARLTEGCSFRRKQH
Q81JI2	MRKYEIMYIIRPGVEEEAQKALVERFAGVLTNNGAEIINTKEWGKRRLAYEINDLREGFYMILNVNAN AEAINEFDRLAKINEDILRHIVVKEEEK
P21468	MRKYEVMYIIRPNIDEESKKAVIERFNNVLTSNGAEITGTKDWGKRRLAYEINDFRDGFYQIVNVQSD AAAVQEFDRLAKISDDIIRHIVVKEEE

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P82921	MAKHLKFIARTVMVQEGNVESAYRTLNRILTMDGLIEDIKHRRYYEKPCRRRQRESYERCRRIYNME MARKINFLMRKNRADPWQGC
Q9HTK7	MKKWQCVVCGLIYDEAKGWPEEGIEAGTRWEDVPEDWLCPDCGVGKLDFEMIEIG
Q9HTK8	MRKWQCVVCGFIYDEALGLPEEGIPAGTRWEDIPADWVCPDCGVGKIDFEMIEIA
P42453	MKKYQCIVCGWIYDEAEGWPQDGIAPGTKWEDIPDDWTCPDCGVSKVDFEMIEV
P09947	MQKYVCSVCGYVYDPADGEPDDPIDPGTGFEDLPEDWVCPVCGVDKDLFEPES
P00268	MKKYTCTVCGYIYNPEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDQFEEVEE
P00269	MKKYVCTVCGYEYDPAEGDPDNGVKPGTSFDDLPADWVCPVCGAPKSEFEAA
Q9V099	MAKWRCKICGYIYDEDEGDPDNGISPGTKFEDLPDDWVCPLCGAPKSEFERIE
P24297	MAKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKLED
P19500	MEKWQCTVCGYIYDPEVGDPTQNIPPGTKFEDLPDDWVCPDCGVGKDQFEKI
P62304	MAYRGQGQKVQKVMVQPINLIFRYLQNRSRIQVWLYEQVNMRIEGCIIGFDEYMNLVLDDAEEIHSKT KSRKQLGRIMLKGDNITLLQSVSN
Q9USZ3	MSGRVQKVMIPPINFIFKLLQQHTPVSIWLFEQTDIRLQGQIRGFDEFMNIVLDDAVQVDAKNNKREL GRILLKGDNITLIQAI
P34659	MSAVQPVNPKPFLNSLTGKFVVCKLKWGMEYKGVLVAVDSYMNLQLAHAEEYIDGNSQGNLGEILIR CNNVLYVGGVDGENETSA
O59734	MSFVPVNPKPFLQGLIGKPVLVRLKWGQEYKGTLQSVDSYMNLQLLNAEELVDGVKTGDLGEILIRC NNVLWVGESTV
P62308	MSKAHPPELKKFMDKKLSLKLNGGRHVQGILRGFDPFMNLVIDECVEMATSGQQNNIGMVVIRGNSII MLEALERV
P62309	MSKAHPPELKKFMDKKLSLKLNGGRHVQGILRGFDPFMNLVIDECVEMATSGQQNNIGMVVIRGNSII MLEALERV
O29386	MPPRPLDVLNRSLKSPVIVRLKGGREFRGTLDGYDIHMNLVLLDAEEIQNGEVVRKVGSVVIRGDTVV FVSPAPGGE
O26745	MIDVSSQRVNVQRPLDALGNSLNSPVIIKLKGDREFRGVLKSFDLHMNLVLNDAEELEDGEVTRRLGT VLIRGDNIVYISP
Q9V0Y8	MAERPLDVIHRSLDKDVLVILKKGFEFRGRLIGYDIHLNVVLADAEMIQDGEVVKRYGKIVIRGDNVLA ISPTEE
P02638	MSELEKAVVALIDVFHQYSGREGDKHKLKKSELKELINNELSHFLEEIKEQEVVDKVMETLDSDGDGE CDFQEFMAFVAMITTACHEFFEHE
P04271	MSELEKAMVALIDVFHQYSGREGDKHKLKKSELKELINNELSHFLEEIKEQEVVDKVMETLDNDGDGE CDFQEFMAFVAMVTTACHEFFEHE
P04631	MSELEKAMVALIDVFHQYSGREGDKHKLKKSELKELINNELSHFLEEIKEQEVVDKVMETLDEDGDGE CDFQEFMAFVSMVTTACHEFFEHE
P02633	MSAKKSPEELKGIFEKYAAKEGDPNQLSKEELKLLLQTEFPSLLKGPSTLDELFEELDKNGDGEVSFEE FQVLVKKISQ
P25815	MTELETAMGMIIDVFSRYSGSEGSTQTLTKGELKVLMEKELPGFLQSGKDKDAVDKLLKDLDANGDA QVDFSEFIVFVAAITSACHKYFEKAGLK
P02639	MGSELETAMETLINVFHAHSGKEGDKYKLSKKELKELLQTELSGFLDAQKDADAVDKVMKELDENGD GEVDFQEYVVLVAALTVACNNFFWENS
P23297	MGSELETAMETLINVFHAHSGKEGDKYKLSKKELKELLQTELSGFLDAQKDVDAVDKVMKELDENGD GEVDFQEYVVLVAALTVACNNFFWENS
Q7LZT1	VSQLESAMESLIKVFHTYSSKEGDKYKLSKAELKSLLQGELNDFLSASKDPMVVEKIMSDLDENQDGEV DFQEFVVLVAALTVACNEFFIESMKN
P56565	MGSELESAMETLINVFHAHSGQEGDKYKLSKKELKDLLQTELSGFLDVQKDADAVDKVMKELDENGD GEVDFKEYVVLVAALTVACNNFFWETS
P35467	MGSELETAMETLINVFHAHSGKEGDKYKLSKKELKDLLQTELSSFLDVQKDADAVDKIMKELDENGD GEVDFQEFVVLVAALTVACNNFFWENS
P10462	MSSPLEQALAVMVATFHKYSGQEGDKFKLSKGEMKELLHKELPSFVGEKVDEEGLKKLMGDLDENSD QQVDFQEYAVFLALITIMCNDFFQGSPARS
P33763	METPLEKALTTMVTTFHKYSGREGSKLTLSRKELKELIKKELCLGEMKESSIDDLMKSLDKNSDQEIDF KEYSVFLTMLCMAYNDFFLEDNK
P63083	METPLEKALTTMVTTFHKYSGREGSKLTLSRKELKELIKTELSLAEKMKESSIDNLMKSLDKNSDQEID FKEYSVFLTTLCMAYNDFFLEDNK
P06703	MACPLDQAIGLLVAIFHKYSGREGDKHTLSKKELKELIQKELTIGSKLQDAEIARLMEDLDRNKDQEVN FQEYVTFLGALALIYNEALKG

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P14069	MACPLDQAIGLLVAIFHKYSGKEGDKHTLSKKELKELIQKELTIGSKLQDAEIARLMDDLDRNKDQEV NFQEYVAFLGALALIYNEALK
P30801	MASPLDQAIGLLIGIFHKYSGKEGDKHTLSKKELKELIQKELTIGSKLQDAEIVKLMDDLDRNKDQEVN FQEYITFLGALAMIYNEALKG
P05964	MACPLDQAIGLLVAIFHKYSGKEGDKHTLSKKELKELIQKELTIGAKLQDAEIARLMDDLDRNKDQEV NFQEYVAFLGALALIYNEALK
P60902	MPSQMEHAMETMMFTFHKFAGDKGYLTKEDLRVLMEKEFPGFLENQKDPLAVDKIMKDLDQCRDG KVGFQSFFSLIAGLTIACNDYFVVHMKQKGKK
P60903	MPSQMEHAMETMMFTFHKFAGDKGYLTKEDLRVLMEKEFPGFLENQKDPLAVDKIMKDLDQCRDG KVGFQSFFSLIAGLTIACNDYFVVHMKQKGKK
P05943	MPSQMEHAMETMMLTFHRFAGEKNYLTKEDLRVLMEREFPGFLENQKDPLAVDKIMKDLDQCRDG KVGFQSFLSLVAGLIIACNDYFVVHMKQKK
P80310	MTKLEDHLEGIINIFHQYSVRLGHYDTLIKRELKQLITKELPNTLKNTKDQGTIDKIFQNLDANQDEQV SFKEFVVLVTDVLITAHDNIHKE
P97352	MAAETLTELEAAIETVVSTFFTFAGREGRKGSLNINEFKELATQQLPHLLKDVGSLDEKMKTLDVNQD SELRFSEYWRLIGELAKEVRKEKALGIRKK
P02738	RSWFSFLGEAYDGARDMWRAYSDMKEANYKNSDKYFHARGNYDAAQRGPGGVWAAEVISDARENI QKLLGHGAEDT
P83569	PIDPGV
P21886	MSQHLVPEAKNGLSKFKNEVAAEMGVPFSDYNGDLSSKQCGSVGGEMVKRMVEQYEKGI
P84584	MAQNSQNGNSSNQLLVPGAAQAIDQMKFEIASEFGVNLGAETTSRANGSVGGEITKRLVSFAQQNMSG QQF
P01144	QGPPISIDLSLELLRKMIEIEKQEKEKQQAANNRLLLDTI
Q9GQW3	MNYLVMISFAFLLMTGVESVRDAYIAQNYNCVYHCARDAYCNELCTKNGAKSGSCPYLGEHKFACYC KDLPDNVPIRVPGKCHRR
Q9CQS8	MPGPTPSGTNVGSSGRSPSKAVAARAAGSTVRQRKNASCGTRSAGRTTSAGTGGMWRFYTEDSPGL KVGPVPVLVMSLLFIAAVFMLHIWGKYTRS
P60058	MDQVMQFVEPSRQFVKDSIRLVKRCTKPDRKEFQKIAMATAIGFAIMGFIGFFVKLIHIPINNIIVGG
P52870	MSSPTPPGGQRTLQKRKQGSSQKVAASAPKKNTNSNNSILKIYSDEATGLRVDPLVVLFLAVGFIFSVV ALHVISKVAGKLF
Q86SD9	MNYLVMISLALLFMIGVESARDGYIAQPNNCVYHCIPLSPGCDKLCRENGATSGKCSFLAGSGLACWC VALPDNVPIKIIGQKCTR
P45658	MNYLIMFSLALLLVIGVESGRDGYIVDSKNCVYHCYPPCDGLCKKNGAKSGSCGFLVPSGLACWCNDL PENVPIKDPSDDCHKR
Q9VM46	MSAPDKEKEKEKEETNNKSEDLGLLEEDDEFEEFPAEDFRVGDDEEELNVWEDNWDDDNVEDDFSQ QLKAHLESKKMET
Q9VHI4	MGERYNIHSQLEHLQSKYIGTGHADTTKFEWLTNQHRDSLASYMGHYDILNYFAIAENESKARVRFNL MERMLQPCGPPPEKLED
Q13296	MKLLMVLMLAALSQHCYAGSGCPLLENVISKTINPQVSKTEYKELLQEFIDDNATTNAIDELKECFLNQ TDETLSNVEVFMQLIYDSSLCDLF
Q91VW3	MSGLRVYSTSVTGSREIKSQQSEVTRILDGKRIQYQLVDISQDNALRDEMRTLAGNPKATPPQIVNGNH YCGDYELFVEAVEQDTLQEFLKLA
B1B5I9	MAYLKIVLVALMLVLGVSAMRLSDQEDQDVSVVKRAACKCDDDGPDIRSATLTGTVDFWNCNEGWE KCTAVYTAVASCCRKKKG
P68722	MKFFLLFLVVLPIMGVLGKKNGYAVDSKGKAPECFLSNYCNNECTKVHYADKGYCCLLSCYCFGLND DKKVLEISDTTKKYCDFTIIN
P68723	MKFFLLFLVVLPIMGVLGKKNGYAVDSKGKAPECFFSNYCNNECTKVHYAEKGYCCLLSCYCVGLND DKKVMEISDTRKKICDTTIIN
P68724	MKFFLLFLVVLPIMGVLGKKNGFAVDSNGKAPECFFDHYCNSECTKVYYAEKGYCCTLSCYCVGLDD DKKVLDISDTRKKLCDFTLFN
P01497	MKFLLLFLVVLPIMGVFGKKNGYAVDSSGKAPECLLSNYCNNECTKVHYADKGYCCLLSCYCFGLND DKKVLEISDTRKSYCDTTIIN
O61668	MKFFLIFLVIFPIMGVLGKKNGYAVDSSGKVSECLLNNYCNNICTKVYYATSGYCCLLSCYCFGLDDDK AVLKIKDATKSYCDVQIIG
P15147	MKFLLLFLVVLPIMGVLGKKNGYAVDSSGKAPECLLSNYCYNECTKVHYADKGYCCLLSCYCFGLND DKKVLEISDTRKSYCDTPIIN
P24336	MKLLLLLVISASMLLECLVNADGYIRKKDGCKVSCIIGNEGCRKECVAHGGSFGYCWTWGLACWCEN LPDAVTWKSSTNTCGRKK
Q26292	MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLIGNEGCDKECKAYGGSYGYCWTWGLACWCEG LPDDKTWKSETNTCGGKK

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
O77091	MKFFLIFLVIFPIMGVLGKKNGYAVDSSGKVAECLFNNYCNNECTKVYYADKGYCCLLKCYCFGLAD DKPVLDIWDSTKNYCDVQIIDLS
P85847	QSDDYGHMRF
P85811	NSDEQFDDYGYMRF
Q7TPF1	MNNLPATPSPEELMTTPVFQAPETMSPQAEEASTALIAVVITVVFLTLLSVVTLIFFYLYKNKGSYVTY EPAEGEPSAILQMETDSAKGKEKEEYFI
P61807	MSIMDHSPTTGVVTVIVILIAIAALGALILGCWCYLRLQRISQSEDEESIVGDGETKEPFLLVQYSAKGP CVERKAKLMTANSPEVHG
P28015	MEVTDVRLRRVNTDGRMRAIASITLDHEFVVHDIRVIDGNNGLFVAMPSKRTPDGEFRDITHPINSSTR GKIQDAVLNEYHRLGDTEALEFEEAGAS
Q969W0	MAGMALARAWKQMSWFYYQYLLVTALYMLEPWERTVFNSMLVSIVGMALYTGYVFMPQHIMAILH YFEIVQ
P02958	MAQQSRSRSNNNNDLLIPQAASAIEQMKLEIASEFGVQLGAETTSRANGSVGGEITKRLVRLAQQNMG GQFH
P16965	MTRTNKWTEREGKADPKYFSHTGNYGESPNHIKKQGSGKGNWGKPGDEIDDLIDNGEIPPVFKKDR RGSNLQSHEQKFENVQKE
P69179	MKKTLLIAASLSFFSASALATPDCVTGKVEYTKYNDDDTFTVKVGDKELFTNRWNLQSLLLSAQITGM TVTIKTNACHNGGGFSEVIFR
P69178	MKKTLLIAASLSFFSASALATPDCVTGKVEYTKYNDDDTFTVKVGDKELFTNRWNLQSLLLSAQITGM TVTIKTNACHNGGGFSEVIFR
Q7BQ98	MKKTLLIAASLSFFSASALATPDCVTGKVEYTKYNDDDTFTVKVGDKELFTNRWNLQSLLLSAQITGM TVTIKTNACHNGGGFSEVIFR
P19991	AAAPF
P55857	MSAAGEEDKKPAGGEGGGAHINLKVKGQDGNEVFFRIKRSTQLKKLMNAYCDRQSVDMNAIAFLFD GRRLRGEQTPDELEMEDGDEIDAMLHQTGGCLPA
Q3E8A8	MSAADKKPLIPPSHITIKIKSQDDICVYFRIKRDVELRTMMQAYSDKVGQQMSAFRFHCDGIRIKPNQT PNELDLEDGDEIDAFVDQIAGFSHRH
B3H5R8	MSSSDKKPLIPSSHITVKVKNQDDICVYFRIKRDVELRKMMHAYSDKVGVEMSTLRFLFDGNRIKLNQ TPNELGLEDEDEIEAFGEQLGGFSFFHRH
P55853	MADDAAQAGDNAEYIKIKVVGQDSNEVHFRVKYGTSMAKLKKSYADRTGVAVNSLRFLFDGRRIND DDTPKTLEMEDDDVIEVYQEQLGGF
E7CLP2	MKILIFIIASFMLIGVECKEGYPMGRDGCKISCVINNNFCKVECQAKWRQSDGYCYFWGLSCYCTNLP EDAQVWDSSTNKCGG
P68726	MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLVGNEGCDKECKAYGGSYGYCWTWGLACWCE GLPDDKTWKSETNTCGGKK
D4GVK4	MLDSIPLFPGLPGGPELLIVLLIVVLLFGANKLPQLARSSGQAMGEFRRGREEIEEELKKGAEGGDDEG ENGDEAEADDADATETEAESR
P0A843	MGEISITKLLVVAALVVLLFGTKKLRTLGGDLGAAIKGFKKAMNDDDAAAKKGADVDLQAEKLSHKE
Q8K2X8	MVNVLKGVLIECDPAMKQFLLYLDEANALGKKFIIQDIDDTHVFVIAELVNVLQERVGELMDQNAFSL TQK
Q68980	MGLSFSGARPCCCRNNVLITDDGEVVSLTAHDFDVVDIESEEEGNFYVPPDMRVVTRAPGRQRLRSSD PPSRHTHRRTPGGACPATQFPPPMSDSE
P13294	MGLAFSGARPCCCRHNVITTDGGEVVSLTAHEFDVVDIESEEEGNFYVPPDVRVVTRAPGPQYRRPS DPPSRHTRRRDPDVARPPATLTPPLSDSE
P01250	PEFLEDPSVLTKEKLKSELVANNVTLPAGEQRKDVYVELYLQSLTALKR
P01251	PEFLEDPSVLTKEKLKSELVANNVTLPAGEQRKEVYVELYLQHLTALKR
Q9C0N3	MFGLGRPQPTSAEKIAAVENELKVVAEMHSRMVKICTLKCIDKSYREGDLSKGESVCLDRCAAKFFET HQKISDQLQKETQARGGGGFGM
P62075	MDSGFGSDFGGTGGGKLDPGAIMEQVKVQIAVANAQELLQRMTDKCFRKCIGKPGGSLDNSEQKCIA MCMDRYMDAWNTVSRAYNSRLQRERANM
Q9WVA2	MESSTSSSGSALGAVDPQLQHFIEVETQKQRFQQLVHQMTELCWEKCMDKPGPKLDSRAEACFVNCV ERFIDTSQFILNRLEQTQKSKPVFSESLSD
Q9WVA1	MESSSSSSGSALAAVDPQLQHFIEVETQKQRFQQLVHQMTELCWEKCMDKPGPKLDSRAEACFVNCV ERFIDTSQFILNRLEQTQKSKPVFSESLSD
Q9Y5J9	MAELGEADEAELQRLVAAEQQKAQFTAQVHHFMELCWDKCVEKPGNRLDSRTENCLSSCVDRFIDT TLAITSRFAQIVQKGGQ
Q9WV98	MAAQIPESDQIKQFKEFLGTYNKLTETCFLDCVKDFTTREVKPEEVTCSEHCLQKYLKMTQRISVRFQ EYHIQQNEALAAKAGLLGQPR

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
Q9WV97	MAAQIPESDQIKQFKEFLGTYNKLTETCFLDCVKDFTTREVKPEEVTCSEHCLQKYLKMTQRISMRFQ EYHIQQNEALAAKAGLLGQPR
O74700	MDALNSKEQQEFQKVVEQKQMKDFMRLYSNLVERCFTDCVNDFTTSKLTNKEQTCIMKCSEKFLKH SERVGQRFQEQNAALGQGLGR
Q91W27	METCQMSRSPRERLLLLLLLLLLVPWGTGPASGVALPLAGVFSLRAPGRAWAGLGSPLSRRSLALADD AAFRERARLLAALERRRWLDSYMQKLLLLDAP
Q8I6S3	MIRVGLILCCIFIAGVFEASSADDMLTAHNLIKRSEVKPPSSSEFIGLMGRSEELTRRLIQHPGSMSETSK RGPPKKVSRRPYILKK
O76201	MWLKIQVFLLAITLITLGIQAEPNSSPNNPLIEEEARACAGLYKKCGKGASPCCEDRPCKCDLAMGNCI CKKKFIEFFGGGK
P59368	MKVAIVFLSLLVLAFASESIEENREEFPVEESARCGDINAACKEDCDCCGYTTACDCYWSKSCKCREA AIVIYTAPKKKLTC
P61103	MVNMKASMFLTFAGLVLLFVVCYASESEEKEFPKEMLSSIFAVDNDFKQEERDCAGYMRECKEKLCC SGYVCSSRWKWCVLPAPWRR
Q86C51	MKTSMFLTLTGLGLLFVVCYASESEEKEFPKELLSSIFAADSDFKVEERGCLGDKCDYNNGCCSGYVC SRTWKWCVLAGPWRR
Q9PSN1	LTCVTSKSIFGITTENCPDGQNLCFKKWYYIVPRYSDITWGCAATCPKPTNVRETIHCCETDKCNE
P80494	LTCVTSKSIFGITTENCPDGQNLCFKKWYYLNHRYSDITWGCAATCPKPTNVRETIHCCETDKCNE
P82462	LICVKEKFLFSETTETCPDGQNVCFNQAHLIYPGKYKRTRGCAATCPKLQNRDVIFCCSTDKCNL
P86419	LTCVTKDTIFGITTQNCPAGQNLCFIRRHYINHRYTEITRGCTATCPKPTNVRETIHCCNTDKCNE
P84944	VPPIGWF
Q9VVA8	MDVMQRYVSPVNPAVFPHLATVLLVIGTFFTAWFFIFVVSRKSSKESTLIKELLISLCASIFLGFGIVFLL LTVGIYV
P61960	MSKVSFKITLTSDPRLPYKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNDGIGINPAQTAGNVFLKH GSELRIIPRDRVGSC
P61961	MSKVSFKITLTSDPRLPYKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNDGIGINPAQTAGNVFLKH GSELRIIPRDRVGSC
Q8G2Q0	MNLTPREKDKLLIAMAAMVARRRLERGVKLNHPEAIALVSDFVVEGARDGRTVAELMEAGAYVITRE QVMDGVAEMIRDIQVEATFPDGTKLVTVHEPIR
P41022	MHLNPAEKEKLQIFLASELLLRRKARGLKLNYPEAVAIITSFIMEGARDGKTVAMLMEEGKHVLTRDD VMEGVPEMIDDIQAEATFPDGTKLVTVHNPIS
Q07399	MKLTSREMEKLMIVVAADLARRRKERGLKLNYPEAVAMITYEVLEGARDGKTVAQLMQYGATILTK EDVMEGVAEMIPDIQIEATFPDGTKLVTVHDPIR
Q9RHM6	MHITPREQEKLMIVVAADLARRRKDRGLKLNHPEAVALITYELIEGARDGRTVADLMSWGSTILTRDD VLEGIPEMIPDIQVEATFDDGTKLVTVHNPIR
P18316	MELTPREKDKLLLFTAALVAERRLARGLKLNYPESVALISAFIMEGARDGKSVASLMEEGRHVLTREQ VMEGVPEMIPDIQVEATFPDGSKLVTVHNPII
P0A676	MRLTPHEQERLLLSYAAELARRRARGLRLNHPEAIAVIADHILEGARDGRTVAELMASGREVLGRDD VMEGVPEMLAEVQVEATFPDGTKLVTVHQPIA
P17088	MELTPREKDKLLLFTAGLVAERRLAKGLKLNYPEAVALISCAIMEGAREGKTVAQLMSEGRTVLTAE QVMEGVPEMIKDVQVECTFPDGTKLVSIHSPIV
Q9L642	MHLSPQEKDKLLIFSAAQLAERRLNRGLKLNYPETVAFLSFQVLEGARDGKSVSQLMSEGTTWLSKKQ VMDGISEMVDEVQVEAVFPDGTKLVTIHNPIN
Q4A0J3	MHFTQREQDKLMLVIAADLARRRQQRGLKLNYPEAVAIISFELLEGARDGKTVAELMSYGKQILNEDD VMEGVADMLTEMEIEATFPDGTKLITVHHPIV
P42875	MHFTQREQDKLMLVIAADLARRRQQRGLKLNYPEAVAIISFELLEGARDGKTVAELMSYGKQILGEDD VMEGVADMLTEMEIEATFPDGTKLITVHHPIV
Q55053	MQLTMREQEKMMISLAAMIAQRRKDKGIKLNHPEAVALITDYVLEGAREGKTVAQLMDEARNLLTRE DVMEGIAEMIPMIQVEATFTDSTKLVTVHDPIQ
O87400	MHLSPQEKDKLLIVTAALLAERRLNRGLKLNHPEAVAWLSFLVLEGARDGKSVAELMQEGTTWLSRN QVMDGIPELVQEVQIEAVFPDGTKLVTLHDPIR
Q96IX5	MAGPESDAQYQFTGIKKYFNSYTLTGRMNCVLATYGSIALIVLYFKLRSKKTPAVKAT
Q78IK2	MAGAESDGQFQFTGIKKYFNSYTLTGRMNCVLATYGGIALLVLYFKLRPKKTPAVKAT
Q9JJW3	MAGPESDGQFQFTGIKKYFNSYTLTGRMNCVLATYGGIALLVLYFKLRPKKTPAVKAT
P86965	MKYVALAFVLSLVILQISAQVGAAYIPGMGLGSVGRTGAVAGASAGVGNQGRGAGILRLLSIIMELVKN NQQAQPKQDTFGAQLQSLLKKKMILEMIN
Q9TS45	EICPTFLRVIESLFLDTPSSFEAAMGFFSPDQDMSEAGAQLKKVLDTLPAKARDSIIKLMEKIDKSLLCN
Q9WUF4	MEASGSAGNDRVRNLQSEVEGVKNIMTQNVERILARGENLDHLRNKTEDLEATSEHFKTTSQKVARK FWWKNVKMIVIICVIVLIILILIILFATGTIPT

Table A.8: Xiao et al. (2013) Data Set Testing Non-AMP Sequences Continued...

Definition	Sequence
P03129	MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQAEPDRAHYNIVTFCCKCDSTLRL CVQSTHVDIRTLEDLLMGTLGIVCPICSQKP
P17387	MRGETPTLQDYVLDLQPEATDLHCYEQLPDSSDEEDVIDSPAGQAEPDTSNYNIVTFCCQCKSTLRLC VQSTQVDIRILQELLMGSFGIVCPNCSTRL
P26557	MRGNNPTLREYILDLHPEPTDLFCYEQLCDSSDEDEIGLDGPDGQAQPATANYYIVTCCYTCGTTVRL CINSTTTDVRTLQQLLMGTCTIVCPSCAQQ
P06464	MHGRHVTLKDIVLDLQPPDPVGLHCYEQLVDSSEDEVDEVDGQDSQPLKQHFQIVTCCCGCDSNVRL VVQCTETDIREVQQLLLGTLNIVCPICAPKT
P86826	YDRYEVVYR
P64877	MYRWCMSRTNIDIDDELAAEVMRRFGLTTKRAAVDLALRRLVGSPLSREFLLGLEGVGWEGDLDDLR SDRPD

Table A.9: APD Gram-Positive AMPs

Definition	Sequence
AP00004	NLCERASLTWTGNCGNTGHCDTQCRNWESAKHGACHKRGNWKCFCYFDC
AP00005	VFIDILDKVENAIHNAAQVGIGFAKPFEKLINPK
AP00012	GLFDIIKKIAESI
AP00091	GLLNTFKDWAISIAKGAGKGVLTTLSCKLDKSC
AP00093	GILDSFKQFAKGVGKDLIKGAAQGVLSTMSCKLAKTC
AP00097	VLPIIGNLLNSLL
AP00104	FLPFLAKILTGVL
AP00105	FLPLFASLIGKLL
AP00107	FLPFLASLLSKVL
AP00109	VLPLISMALGKLL
AP00110	NFLGTLINLAKKIM
AP00111	FLPILINLIHKGLL
AP00112	FLPIVGKLLSGLL
AP00140	SQLGDLGSGAGQGGGGGGSIRAAGGAFGKLEAAREEEFFYKKQKEQLERLKNDQIHQAEFHHQQIKE HEEAIQRHKDFLNNLHK
AP00171	HRHQGPIFDTRPSPFNPNQPRPGPIY
AP00182	GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCYR
AP00194	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
AP00195	RGGRLCYCRRRFCVCVGR
AP00203	QCIGNGGRCNENVGPPYCCSGFCLRQPGQGYGYCKNR
AP00204	ITSISLCTPGCKTGALMGCNMKTATCNCSIHVSK
AP00206	WKSESLCTPGCVTGALQTCFLQTLTCNCKISK
AP00207	TAGPAIRASVKQCQKTLKATRLFTVSCKGKNGCK
AP00208	RADTQTYQPYNKDWIKEKIYVLLRRQAQQAGK
AP00216	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
AP00226	VTCDLLSFKGQVNDSACAANCLSLGKAGGHCEKVGCICRKTSFKDLWDKRF
AP00267	LFCKRGTCHFGRCPSHLIKVGSCFGFRSCCKWPWDA
AP00326	GVGSFIHKVVSAIKNVA
AP00327	GWFDVVKHIASAV
AP00336	AERVGAGAPVYL
AP00354	VTCDILSVEAKGVKLNDAACAAHCLFRGRSGGYCNGKRVCVCR
AP00391	FIGTALGIASAIPAIVKLFK
AP00392	YRGGYTGPIPRPPPIGRPPLRLVVCACYRLSVSDARNCCIKFGSCCHLVK
AP00394	QVYKGGYTRPIPRPPPFVRPLPGGPIGPYNGCPVSCRGISFSQARSCCSRLGRCCHVGKGYS
AP00395	HSSGYTRPLPKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHR

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP00403	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
AP00409	ATTGCSCPQCIIFDPICASSYKNGRRGFSSGCHMRCYNRCHGTDYFQISKGSKCI
AP00410	PKRKSATKGDEPARRSARLSARPVPKPAAKPKKAAAPKKAVKGKKAAENGDAKAEAKVQAAGDGA GNAK
AP00422	QGCKGPYTRPILRPYVRPVVSYNACTLSCRGITTTQARSCCTRLGRCCHVAKGYS
AP00428	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPF GWKSIFIQC
AP00438	GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLPCTCYRCG
AP00440	VTCFCRRRGCASRERHIGYCRFGNTIYRLCCRR
AP00449	SYSMEHFRWGKPV
AP00463	FLPFIAGMAAKFLPKIFCAISKKC
AP00490	AALKGCWTKSIPPKPCSGKR
AP00497	ILGPVLGLVSDTLDDVLGIL
AP00518	QYRHRCCAWGPGRKYCKRWC
AP00519	QWGRRCCGWGPGRRYCRRWC
AP00530	VLSKSLCTPGCITGPLQTCYLCFPTFAKC
AP00536	SVRTQDNAVNRQIFGSNGPYRDFQLSDCYLPLETNPYCNEWQFAYHWNNALMDCERAIYHGCNRTR NNFITLTACKNQAGPICNRRRH
AP00542	ILGTILGLLKSL
AP00546	FLSLIPHAINAVSAIAKHN
AP00549	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
AP00558	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
AP00564	FLIGMTHGLICLISRKC
AP00569	FLPLLLAGLPLKLCFLFKKC
AP00582	GFSSLFKAGAKYLLKSVGKAGAQQLACKAANNCA
AP00584	VIDDLKKVAKKVRRELLCKKHHKKLN
AP00589	FLGALAKIISGIF
AP00594	EGTWQHGYGVSSAYSNYHHGSKTHSATVVNNNTGRQGKDTQRAGVWAKATVGRNLTEKASFYYNF W
AP00598	FLSAITSLLGKLL
AP00605	ILPILSLIGGLLGK
AP00632	KYYGNGVSCNKNGCTVDWSKAIGIIGNNAAANLTTGGAAGWNKG
AP00634	KYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGNHKC
AP00635	KYYGNGVHCTKSGCSVNWGEAASAGIHRLANGGNGFW
AP00658	FLPLVGKILSGLI
AP00659	FLPIASLLGKYL
AP00673	VGSRYLCTPGSCWKLVCFTTTVK
AP00712	GFGCPLNQGACHRHCRSIRRRGGYCAGFFKQTCCYRN
AP00729	GLPVCGETCVGGTCNTPGCTCSWPVCTRN
AP00748	DIQIPGIKKPTHRDIIIPNWNPNVRTQPWQRFGGNKS
AP00749	EADEPLWLYKGDNIERAPTTADHPILPSIIDDVKLDPNRRYA
AP00750	EIRLPEPFRFPSPTVPKPIDIDPILPHPWSPRQTYPIIARRS
AP00752	DKLIGSCVWGATNYTSDCNAECKRRGYKGGHCGSFWNVNCWCEE
AP00753	VQETQKLAKTVGANLEETNKKLAPQIKSAYDDFVKQAQEVQKKLHEAASKQ
AP00754	ETESTPDYLKNIQQQLEEYTKNFNTQVQNAFDSDKIKSEVNNFIESLGKILNTEKKEAPK
AP00812	FAEPLPSEEEGESYSKEPPEMEKRYGGFM
AP00819	FLPLLASLFSRLF
AP00824	SILPTIVSFLSKFL
AP00831	GLLSGILGAGKHIVCGLTGCAKA
AP00841	TTHSGKYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAIPGKC
AP00842	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLGGAIPGKC

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP00845	KSYGNGVHCNKKKCWVDWGSAISTIGNNSAANWATGGAAGWKS
AP00848	KNYGNGVHCTKKGCSVDWGYAWANIANNSVMNGLTGGNAGWHN
AP00849	TSYGNGVHCNKSKCWIDVSELETYKAGTVSNPKDILW
AP00850	KYYGNGVSCNSHGCSVNWGQAWTCGVNHLANGGHGVC
AP00852	KYYGNGLSCSKKGCTVNWGQAFSCGVNRVATAGHHKC
AP00853	ATRSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH
AP00867	FLPVIAGLLSKLF
AP00876	FLSIIAKVLGSLF
AP00886	FLPILASLAATLGPKLLCLITKKC
AP00900	FLSHIAGFLSNLF
AP00911	FLSLIPHIVSGVAALAKHL
AP00915	QQCGRQAGNRRCANNLCCSQYGYCGRTNEYCCTSQGCQSQCRRCG
AP00916	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
AP00930	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGVKHSSGGGGSYHC
AP00973	LLGMIPLAISAISALSKL
AP01005	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
AP01131	MSWLNFLKYIAKYGKKAVSAAWKYKGKVLEWLNVGPTLEWVWQKLKKIAGL
AP01152	SIWGDIGQGVGKAAYWVGKAMGNMSDVNQASRINRKKKH
AP01167	LTTKLWSSWGYYLGKKARWNLKHPYVQF
AP01168	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL
AP01169	NRWGDTVLSAASGAGTGIKACKSFGPWGMAICGVGGAAIGGYFGYTHN
AP01170	YSSKDCLKDIGKGIGAGTVAGAAGGGLAAGLGAIPGAFVGAHFGVIGGSAACIGGLLGN
AP01171	YSGKDCLKDMGGYALAGAGSGALWGAPAGGVGALPGAFVGAHVGAIAGGFACMGGMIGNKFN
AP01179	NGVYCNKQKCWVDWSRARSEIIDRGVKAYVNGFTKVLGGIGGR
AP01181	AYPGNGVHCGKYSCTVDKQTAIGNIGNNAA
AP01182	FTPSVSFSQNGGVVEAAAQRGYIYKKYPKGAKVPNKVKMLVNIRGKQTMRTCYLMSWTASSRTAKY YYYI
AP01183	ATYYGNGLYCNKEKCWVDWNQAKGEIGKIIVNGWVNHGPWAPRR
AP01185	ENDHRMPNNLNRPNNLSKGGAKCGAAIAGGLFGIPKGPLAWAAGLANVYSKCN
AP01186	KTYYGTNGVHCTKKSLWGKVRLKNVIPGTLCRKQSLPIKQDLKILLGWATGAFGKTFH
AP01187	MNFLKNGIAKWMTGAELQAYKKKYGCLPWEKISC
AP01188	MLAKIKAMIKKFPNPYTLAAKLTTYEINWYKQQYGRYPWERPVA
AP01189	APAGLVAKFGRPIVKKYYKQIMQFIGEGSAINKIIPWIARMWRT
AP01194	CSTNTFSLSDYWGNNGAWCTLTHECMAWCK
AP01195	KRGSGWIATITDDCPNSVFVCC
AP01198	LSCDEGMLAVGGLGAVGGPWGAAVGVLVGAALYCF
AP01199	KYYGNGVHCGKKTCYVDWGQATASIGKIIVNGWTQHGPWAHR
AP01201	KGGSGVIHTISHECNMNSWQFVFTCCS
AP01204	GKNGVFKTISHECHLNTWAFLATCCS
AP01205	STPVLASVAVSMELLPTASVLYSDVAGCFKYSAKHHC
AP01206	CTFTLPGGGGVCTLTSECIC
AP01214	GNRPVYIPPPRPPHPRL
AP01236	GLLDFAKHVIGIASKL
AP01242	GLLSFLPKVIGVIGHLIHPPS
AP01260	IIGHLIKTALGMLGL
AP01261	IIEKLVNTALGLLSGL
AP01262	GLADFLNKAVGKVVDFVKS
AP01263	FLPLVTMLLGKLF
AP01268	FLPVILPVIGKLLSGIL
AP01307	GCSRWIIGIHGQICRD

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP01317	GAILCNLCKDTVKLVENLLTVDGAQAVRQYIDNLCGKASGFLGTLCEKILSFGVDELVKLIENHVDPV VVCEKIHAC
AP01318	IPVLCPVCTSLVGKLIDLVLGGAVDKVTDYLETLCAKADGLVETLCTKIVSYGIDKLIEKILEGGSAKLI CGLIHAC
AP01326	SKGKKANKDVELARG
AP01327	LFGLIPSLIGGLVSAFK
AP01339	FLSFPTTKTYFPHFDLSHGSAQVKGHGAK
AP01340	DAECEICKFVIQQVEAFIESNHSQAEIQKELNKLCSSVPSITQTCLSIARMVPYIIKKLEEHNSPGQVCQG LHLCKSS
AP01346	FFPLVLGALGSILPKIF
AP01350	FLSLLPSIVSGAVSLAKKL
AP01358	VTCDLLSFEAKGFAANHSLCAAHCLAIGRRGGSCERGVCICRR
AP01364	ATCDLLSAFGVGHAACAAHCIGHGYRGGYCNSKAVCTCRR
AP01449	FLGAIAAALPHVINAVTNAL
AP01461	ILGKLLSTAAGLLSNL
AP01493	ASIIKTTIKVSKAVCKTLTCICTGSCSNCK
AP01495	IASKFICTPGCAKTGSFNSYCC
AP01511	TITLSTCAILSKPLGNNGYLCTVTKECMPSSCN
AP01522	TYMPVEEGEYIVNISYADQPKKNSPFTAKKQPGPKVDLSGVKAYGPG
AP01530	GSCSCSGTISPYGLRTCRATKTKPSHPTTKETHPQTLPT
AP01555	TCRYWCKTPENQTYCCEDEREIPSKVGLKPGKCPPVRPVCPPTRGFFEPPKTCSNDGSCYGADKCCF DRCLGEHVCKPIQTRG
AP01556	GCFEDWSRCSPSTSRGTGVLWRDCDSYCKVCFKADRGECFDSPSLNCPQRLPNNKQCRCINARTAKD NRNPTCWA
AP01557	DHHHDHGHDDHEHEELTLEKIKEKIKDYADKTPVDQLTERVQAGRDYLLGKGARPSHLPARVDRHLS KLTAAEKQELADYLLTFLH
AP01564	ATCDLLSKWNWNHTACAGHCIAKGFKGGYCNDKAVCVCRN
AP01569	SNDSLWYGVGQEMGKQANCITNHPVKHMIIPGYCSKILG
AP01570	GNAACVIGCIGSCVISEGIGSLVGTAFTLG
AP01576	RVPPYLGRDCKHWCRDNNQALYCCGPPGITYPPFIRKHPGKCPSVRSTCTGVRSSRPKFCPHDDACE FRSKCCYDACVKHHVCKTVEFY
AP01577	FLLFPLMCKIQGKC
AP01579	FVLPLVMCKILRKC
AP01589	KDRPKKPGLCPPRPQKPCVKECKNDDSCPGQQKCCNYGCKDECRDPIFVG
AP01593	CKQSCSFGPFTFVCDGNTK
AP01595	CANSCSYGPLTWSCDGNTK
AP01597	SVSCLRNKGVCMPGKCAPKMKQIGTCGMPQVKCCKRK
AP01600	RARAPHKAWYNCMTDAGISGAIAGAVAGCAATIEIGCVEGAIAGIGPSGIASMIAALWTCRSKY
AP01603	SSSGWLCTLTIECGTIICACR
AP01604	DAPGHPGKHYLQVNVPSDVRTIGVAGGGVQQCFRVTPGAWNDTRALVSNGAQVEVWGYTVADCAN RTTANQKYYDKAAAPSDSSTYFWFTLKNLRV
AP01606	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR
AP01607	ADRGWIKTLTKDCPNVISSICAGTIITACKNCA
AP01609	KCKWWNISCDLGNNGHVCTLSHECQVSCN
AP01612	SASIVKTTIKASKKLCRGFTLTCGCHFTGKK
AP01615	SASVLKTSIKVSKKYCKGVTLTCGCNITGGK
AP01616	SLGPAIKATRQVCPKATRFVTVSCKKSDCQ
AP01617	VTSWSLCTPGCTSPGGGSNCSFCC
AP01618	GTTVVNSTFSIVLGNKGYICTVTVECMRNCSK
AP01621	CAWYNISCRLGNKGAYCTLTVECMPSCN
AP01632	ATPATPTVAQFVIQGSTICLVC
AP01647	RCTCTTIISSSSTF
AP01651	LVATGMAAGVAKTIVNAVSAGMDIATALSLFSGAFTAAGGIMALIKKYAQKKLWKQLIAA

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP01657	VIPFVASVAAEMMHHVYCAASKRC
AP01661	ACYCRIPACLAGERRYGTCFYLGRVWAFCC
AP01663	RRTCRCRFGRCFRRESYSGSCNINGRIFSLCCR
AP01682	DTHFPICIFCCGCCHRSKCGMCCKT
AP01715	PSCVCSGFETSGIHFC
AP01721	FLPVLTGLTPSIVPKLVCLLTKKC
AP01729	GSQLVYREWVGHSNVIKGPP
AP01737	ILGPVLGLVGNALGGLIKKL
AP01738	GIGGVLLSAGKAALKGLTKVLAEKYAN
AP01741	SIGAKILGGVKTFFKGALKELAFTYLQ
AP01743	GIGGALLSVGKLALKGLANVLADKFAN
AP01744	ILGPVIKTIGGVIGGLLKNL
AP01746	FLGGILNTITGLL
AP01754	GGYYCPFFQDKCHRHCRSFGRKAGYCGGFLKKTCICV
AP01758	IPAMEPAARVKRSPGYGGCSPRWACGGYG
AP01783	FLPGVLRLVTKVGPAVVCAITRNC
AP01789	HSHACTSYWCGKFCGTASCTHYLCRVLHPGKMCACVHCSR
AP01791	FLWGLIPGAISAVTSLIKK
AP01794	FVDLKKIANIINSIFGK
AP01892	IPWKLPATFRPVERPFSKPFCRKD
AP01904	GFGSFLGKALKAALKIGADVLGGAPQQ
AP01911	SALVGCWTKSYPPNPCFGRG
AP01914	AAFRGCWTKNYSPKPCL
AP01921	ILPIIGKILSTIF
AP01941	CVHWQTNTARTSCIGP
AP01944	SLWETIKNAGKGFILNILDKIRCKVAGGCKT
AP01946	FVGPVLKIAAGILPTAICKIYKKC
AP01955	EYHLMNGANGYLTRVNGKTVYRVTKDPVSAVFGVISNCWGSAGAGFGPQH
AP01971	VTSKSLCTPGCITGVLMCLTQNSCVSCNSCIRC
AP01972	STIVCVSLRICNWSLRFCPSFKVRCPM
AP01993	TNWKKIGKCYAGTLGSAVLGFGAMGPVGYWAGAGVGYASFC
AP01995	ECELAKVDGGYTPKNCAMAVGGGMLSGAIRGGMSGTVFGVGTGNLAGAFAGAHIGLVAGGLACIGG YLGSH
AP01997	TPGGIDFISGGPHVAQDVLNAIKNFFK
AP02005	GLMSSIGKALGGLIVDVLKPKTPAS
AP02006	GLLDALSGILGL
AP02007	GLLGTLGNLLNGLGL
AP02008	GLVSSIGKVLGGLLADVVKSKGQPA
AP02010	GLFGILGSVAKHVLPHVIPVVAEHL
AP02011	GLFDVIKKVASVIGLASP
AP02019	ILGAIIPLVSGLLSHL
AP02025	DCYEDWSRCTPGTSFLTGILWKDCHSRCKELGHRGGRCVDSPSKHCPGVLKNNKQCHCY
AP02027	GFWGSLWEGVKSVV
AP02028	KRKCPKTPFDNTPGAWFAHLILGC
AP02029	DSIRDVSPTFNKIRRWFDGLFK
AP02042	DRCSQQCQHHRDPDRKQQCMRECRRHQGRSD
AP02051	AISCGQVSSAIGPCLSYARGQGSAPSAGCC
AP02052	TPALAVVTTVLPAAAVTTAKSV
AP02059	FLSTALKVAANVVPTLFCKITKKC

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP02068	YCERMMKRRSLTSPCKDVNTFIHGNKSNIKAICGANGSPYRENLRMSKSPFQVTTCKHTGGSPRPPCQ YRASAGFRHVVIACENGLPVHFDESFFSL
AP02117	VLSIVACSSGCGSGKTAASCVETCGNRCFTNVGSLC
AP02119	GFGCPGDAYQCSEHCRALGGGRTGGYCAGPWYLGHPTCTCSF
AP02143	YSLQMGATAIKQVKKLFKKKGG
AP02162	KIKIPWGKVKDFLVGGMKAV
AP02166	LVPLFLSKLICFITKKC
AP02174	FLPFLIPALTSLISSL
AP02197	PAAAAQAVAGLAPVAAEQ
AP02203	WNDTGKDADGSEY
AP02219	WSCPTLSGVCRKVCLPTEMFFGPLGCGKEFQCCVSHFF
AP02220	NKGCAICSIGAACLVDGPIPDFEIAGATGLFGLWG
AP02226	FCHLCEDLIKDGKEAGDVALDVWLDEEIGSRCKDFGVLASECFKELKVAEHDIWEAIDQEIPEDKTCK EAKLC
AP02246	MAKEFGIPAAVAGTVLNVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVI AW
AP02247	MAGFLKVVQLLAKYGSKAVQWAWANKGKILDWLNAGQAIDWVVSKIKQILGIK
AP02248	FKKKKRNIGTFVFFAIALFCTVMFAYLLLTNQYVPIDYNVPRYA
AP02249	GLLSLLSLLGKLL
AP02250	MKTILRFVAGYDIASHKKKTGGYPWERGKA
AP02251	MWGRILAFVAKYGTKAVQWAWKNKWFLLSLGEAVFDYIRSIWGG
AP02253	MGAIAKLVAKFGWPFIKKFYKQIMQFIGQGWTIDQIEKWLKRH
AP02273	FITGLIGGLMKAL
AP02304	FLAGLIGGLAKML
AP02308	GNGVVLTLTHECNLATWTKKLKCC
AP02310	LKLSPETKDTLKKVLKGAIKGAIAIASLA
AP02311	ITIPPIVKDTLKKFFKGGIAGVMGKSQ
AP02323	KRKKHRCRVYNNGMPTGMYRWC
AP02341	LVKDNPLDISPKQVQALCTDLVIRCMCCC
AP02342	DICTCCAGTKGCNTTSANGAFICEGQSDPKKPKACPLNCDPHIAYA
AP02346	GLEESPGHPGQPGPPGPPGAPGP
AP02348	TTPLCVGVIIGLTTSIKICK
AP02351	QKIAEKFSGTRRG
AP02353	LPVNSPMNKGDTEVMKCIVEVISDTLSKPSPMPVSKECFETLRGDERILSILRHQNLLKELQDLALQGA KERTHQQ
AP02356	FLFSLIPHAIGGLISAFK
AP02371	GHLGRPYIGGGGGFNRGGGFHRGGGFHRGGGFHSGGGFHRGGGFHSGGSFGYR
AP02376	GLRRLFADQLVGRRNI
AP02394	DWTCWSCLVCAACSVELLNLVTAATGASTAS
AP02396	LASTLGISTAAAKKAIDIIDAASTIASIISLIGIVTGAGAISYAIVATAKTMIKKYGKKYAAAW
AP02397	CWSCMGHSCWSCMGHSCWSCAGHSCWSCMGHSCWSCMGHSCWSCAGHCCGSCWHGGM
AP02399	ESISVAGGTWNYGYGVGQAYSHYKHDYNNHGAKVVNSNNGVKDYKNAGPGVWAKASIGTVWDPAT FYYNPTGFYSN
AP02400	FAVWGCADYRGYCRAACFAFEYSLGPKGCTEGYVCCVPNTF
AP02402	IGGALGNALNGLGTWANMMNGGGFVNQWQVYANKGKINQYRPY
AP02403	GILGKLWEGVKSIF
AP02409	AIFIFIRWLLKLGHHGRAPP
AP02433	GDINGEFTTSPACVYSVMVVSKASSAKCAAGASAVSGAILSAIRC
AP02434	TVKCGMNGKMPCKHGAFYTDTCDKNVFYRCVWGRPVKKHCGRGLVWNPRGFCDYA
AP02439	SCTTCVCTCSCCTT
AP02440	QNCPTRRGLCVTSGLTACRNHCRSCHRGDVGCVRCSNAQCTGFLGTTCTCINPCPRC

Table A.9: APD Gram-Positive AMPs Continued...

Definition	Sequence
AP02457	KKCGFFCKLKNKLKSTGSRSNIAAGTHGGTFRV
AP02460	FLPLLFGALSTLLPKIF

Table A.10: APD Gram-Negative AMPs

Definition	Sequence
AP00006	GNNRPVYIPQPRPPHPRI
AP00009	RFRPPIRRPPIRPPFYPPFRPPIRPPIFPPIRPPFRPPLGPFP
AP00010	RRIRPRPPRLPRPRPRPLPFPRPGPRPIPRPLPFPRPGPRPIPRPLPFPRPGPRPIPRPL
AP00050	GIGASILSAGKSALKGLAKGLAEHFAN
AP00116	GFLDIINKLGKTFAGHMLDKIKCTIGTCPPSP
AP00117	FLPFIARLAAKVFPSIICSVTKKC
AP00142	GLKKLLGKLLKKLGKLLLK
AP00168	GRPNPVNNKPTPHPRL
AP00196	WYVKKCLNDVGICKKKCKPEEMHVKNGWAMCGKGRDCCVPAD
AP00197	QLKKCWNNYVQGHCRKICRVNEVPEALCENGRYCCLNIKELEAC
AP00302	GCRFCCNCCPNMSGCGVCCRF
AP00356	QRFIHPTYRPPPQPRRPVIMRA
AP00357	FFPIGVFCKIFKTC
AP00364	VDKPDYRPRPWPRNMI
AP00382	GLVDVLGKVGGLIKKLLPG
AP00397	SGFVLKGYTKTSQ
AP00405	FISAIASMLGKFL
AP00408	FLFPLITSFLSKVL
AP00412	SLQPGAPNVNNKDQPWQVSPHISRDDSGNTRTDINVQRHGENNDFEAGWSKVVRGPNKAKPTWHIG GTHRW
AP00413	SLQGGAPNFPQPSQQNGGWQVSPDLGRDDKGNTRGQIEIQNKGKDHDFNAGWGKVIRGPNKAKPT WHVGGTYRR
AP00425	GCWSTVLGGLKKFAKGGLEAIVNPK
AP00426	GVFLDALKKFAKGGMNAVLNPK
AP00431	TWLKKRRWKKAKPP
AP00453	FLPAIVGAAGQFLPKIFCAISKKC
AP00455	FFPIVAGVAGQVLKKIYCTISKKC
AP00480	VGIGTPIFSYGGGAGHVPEYF
AP00501	GIGKHVGKALKGLKGLLKGLGES
AP00509	VAIALKAAHYHTHKE
AP00512	GYHGGHGGHGGGYNGGGGHGGHGGGYNGGGHHGGGGHG
AP00533	GVVDILKGAAKDIAGHLASKVMNKL
AP00543	GVVDILKGAGKDLLAHLVGKISEKV
AP00544	GVLDIFKDAAKQILAHAAEKQI
AP00579	GLLSGILGAGKHIVCGLSGLC
AP00608	KRIVQRIKDFLR
AP00615	ALFSILRGLKKLGNMGQAFVNCKIYKKC
AP00616	ALSILRGLEKLAKMGIALTNCKATKKC
AP00619	GFFSTVKNLATNVAGTVIDTLKCKVTGGCRS
AP00621	GIFPKIIGKGIKTGIVNGIKSLVKGVGMKVFKAGLNNIGNTGCNEDEC
AP00652	GIMDTVKNVAKNLAGQLLDKLKCKITAC
AP00654	GLLDTIKNTAKNLAVGLLDKIKCKMTGC

Table A.10: APD Gram-Negative AMPs Continued...

Definition	Sequence
AP00663	GFSSIFRGVAKFASKGLGKDLARLGVNLVACKISKQC
AP00707	RPDKPRPYLPRPRPPRPVR
AP00743	RYHMQCGYRGTFCTPGKCPYGNAYLGLCRPKYSCCRWL
AP00744	GLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCCRSRWYS
AP00745	MTPFWRGVSLRPVGASCRDNSECITMLCRKNRCFLRTASE
AP00805	RSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECDHYVNSKVDPIIHELEGGT APKDVCTKLNECP
AP00810	QSHLSLCRWCCNCCRSNKGC
AP00814	GLGSILGKILNVAGKVGKTIGKVADAVGNKE
AP00816	GLGSFFKNAIKIAGKVGSTIGKVADAIGNKE
AP00893	DVKGMKKAIKGILDCVIEKGYDKLAAKLKKVIQQLWE
AP00964	GLWSKIKEAAKAAGKAALNAVTGLVNQGDQPS
AP01036	GIPCGESCVWIPCISSAIGCSCKSKVCYRN
AP01158	ALYKKFKKKLLKSLKRL
AP01196	GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGGALQTVGQGLIDHGPVNVFIPVLIGPSW NGSGSGYNSATSSSGSGS
AP01227	VGIGGGGGGGGGGSCGGQGGGCGGCSNGCSGGNGGSGGSGSHI
AP01228	ASGRDIAMAIGTLSGQFVAGGIGAAAGGVAGGAIYDYASTHKPNPAMSPSGLGGTIKQKPEGIPSEAW NYAAGRLCNWSPNNLSDVCL
AP01229	GDVNWVDVGKTVATNGAGVIGGAFGAGLCGPVCAGAFAVGSSAAVAALYDAAGNSNSAKQKPEGL PPEAWNYAEGRMCNWSPNNLSDVCL
AP01230	DGNDGQAELIAIGSLAGTFISPGFGSIAGAYIGDKVHSWATTATVSPSMSPSGIGLSSQFGSGRGTSSASS SAGSGS
AP01231	GGAPATSANAAGAAAIVGALAGIPGGPLGVVVGAVSAGLTTGIGSTVGSGSASSSAGGGS
AP01232	MNLNGLPASTNVIDLRGKDMGTYIDANGACWAPDTPSIIMYPGGSGPSYSMSSSTSSANSGS
AP01258	GLMDVFKGAAKNLLASALDKIRCKVTKC
AP01266	AVDLAKIANKVLSSLF
AP01285	AGDPLADPNSQIVRQIMSNAAWGPPLVPERFRGMAVGAAGGVTQTVLQGAAAHMPVNVPIPKVPMG PSWNGSKG
AP01316	NPANPLNLKKHHGVFCDVCKALVEGGEKVGDDDLDAWLDVNIGTLCWTMLLPLHHECEEELKKVKK ELKKDIENKDSPDKACKDVDLC
AP01347	FIITGLVRGLTKLF
AP01362	ATCDLLSGFGVGDSACAAHCIARGNRGGYCNSKKVCVCPI
AP01365	AAKPMGITCDLLSLWKVGHAACAAHCLVLGDVGGYCTKEGLCVCKE
AP01366	ATCDLLSMWNVNHSACAAHCLLLGKSGGRCNDDAVCVCRK
AP01381	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
AP01382	QKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
AP01398	YQEPVLGPVRGPFPIIV
AP01400	RPKHPIKHQGLPQEVLNENLLRF
AP01405	GLLGGLLGPLLGGGGGGGGLL
AP01429	GLLDSFKNAMIGIAKSAGKTALNKIACKIDKTC
AP01455	FFPLALLCKVFKKC
AP01457	GLKDIFKAGLGSLVKGIAAHVAN
AP01476	ACDTATCVTHRLAGLLSRSGGVVKNNFVPTNVGSKAF
AP01513	GKNPTLQCMGNRGFCRPSCKKGEQAYFYCRTYQICCLQSHVRISLTGVEDNTNWSYEKHWPRIP
AP01514	GVNMYIRQIYDTCWKLKGHCRNVCGKKEIFHIFCGTQFLCCIERKEMPVLFVK
AP01520	SSFSPPRGPPGWGPPCVQQPCPKCPYDDYKCPTCDKFPECEECPHISIGCECGYFSCECPKPVCEPCE SPIAELIKKGGYKG
AP01546	GMWSKIKNAGKAAKAAAKAAGKAALGAVSEAM
AP01565	DDMTMKPTPPPQYPLNLQGGGGGGSGDGFGFAVQGHQKVWTSDNGRHEIGLNGGYGQHLGGPYGN SEPSWKVGSTYTYRFPNF
AP01566	QRPYTQPLIYYPPPPTPPRIYRA
AP01592	GIRNTVCFMQRGHCRLFMCRSGERKGDICSDPWNRCCVSSSIKNR

Table A.10: APD Gram-Negative AMPs Continued...

Definition	Sequence
AP01602	LVLKYCPKIGYCSNTCSKTQIWATSHGCKMYCCLPASWKWK
AP01619	AANFGPSVFTPEVHETWQKFLNVVVAALGKQYH
AP01620	VDKPPYLPRPPPPRRIYNNR
AP01623	GFKLKGMARISCLPNGQWSNFPPKCIRECAMVSS
AP01624	HAEHKVKIGVEQKYGQFPQGTEVTYTCSGNYFLM
AP01724	GTPGFQTPDARVISRFGFN
AP01745	ERILDLRKTKKSCKNGEVLGCVSGHGPPGCSENECGMGPRPKACFFDCHYGCWCTGKLYRRKRDRK CVPKHECLL
AP01747	SFPFFPPGICKRLKRC
AP01748	SFHVFPPWMCKSLKKC
AP01753	GIWSSIKNLASKAWNSDIGQSLRNKAAGAINKFVADKIGVTPSQAAS
AP01849	TSRCIFYRRKKCS
AP01925	GLLDAIKDTAQNLFANVLDKIKCKFTKC
AP01927	GLFNVFKKVGKNVLKNVAGSLMDNLKCKVSGEC
AP02001	GMATKAGTALGKVAKAVIGAAL
AP02003	GFWTTAAEGLKKFAKAGLASILNPK
AP02033	VSCGDVTSSIAPCLSYVMGRESSPSSSCCSGVRTLNGKASSSADRRTACSCLKNMASSFRNLNMGNAASI PSKCGVSVAFPISTSVDCSKIN
AP02053	GLSQGVEPDIGQTYFEESRINQD
AP02086	PDSVSIPITCCFNVINRKIPIQRLESYTRITNIQCPKEAVIFKTKRGKEVCADPKERWVRDSMKHLDQIF QNLKP
AP02088	QPDALNVPSTCCFTFSSKKISLQRLKSYVITTSRCPQKAVIFRTKLGKEICADPKEKWVQNYMKHLGR KAHTLKT
AP02091	GTNDAEDCCLSVTQKPIPGYIVRNFHYLLIKDGCRVPAVVFTTLRGRQLCAPPDQPWVERIIQRLQRT SAKMKRRSS
AP02104	MQFITDLIKKAVDFFKGLFGNK
AP02105	MAADIISTIGDLVKLIINTVKKFQK
AP02125	FWGKLWEGVKNAI
AP02128	SLQPGAPNFPIPGQEKQEGWKFDPSLTRGEDGNTLGSINIHHTGPNHEVGANWDKVIRGPGKAKPTY SIHGSWRW
AP02137	FLHHIVGLIHHGLSLFGDRAD
AP02150	YEALVTSILGKLTGLWHNDSVDFMGHICYFRRRPKIRRFKLYHEGKFWCPGWAPFEGRCKYCVVF
AP02153	GGWLDIVKAIVVPAARETIKTQEITLLDHYCTLSRSPYIKSLELHYRAEVTCPGWTIIRGRGSNHRNPT NSGKDALKDFMTQAVAAGLVTKEEAAPWLN
AP02154	YVDREINLFDHYCIISRSPHISRWELKWQATVTCPGWTPVKGKVRGYSNPLSAEREATRDFVQRIVQR GLVTRDEASEWL
AP02170	NPVSCVRNKGICVPIRCPGNMKQIGTCVGRAVKCCRKK
AP02173	QLPICGETCVLGGCYTPNCRCQYPICVR
AP02184	APAHRSSTFPKWVTKTERGRQPLRS
AP02224	GLLLDTVKGAAKNVAGILLNKLKCKMTGDC
AP02230	PKRKAEGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPKPKKAPAKKGEKVPKGKKGKADAGKEG NNPAENGDAKTDQAQKAEGAGDAK
AP02231	RAIGGGLSSVGGGSSTIKY
AP02237	FFRLLFHGVHHGGGYLNAA
AP02238	GWKKWFTKGERLSQRHFA
AP02239	GFLGILFHGVHHGRKKALHMNSERRS
AP02286	NLLGSLLKTGLKVGSNLL
AP02320	KTYYGTNGVHCTKNSLWGKVRLKNMKYDQNTTYMGRLQDILLGWATGAFGKTFH
AP02365	VIVKAIATLSKKLL
AP02417	MKFFTLLAALMALFAICNNFSMVSASRDSRPVQPRVQPPPPPPKQKPSIYDTPIRRPGGQKTMYA
AP02443	NIFDDIFGKVTETLVDFGTTDIAGNPCNYRLSPRLIKFELYFVGLVWCPGWTTIQGESLTRSRTRVVN KAVEDFAKKAVAAGIMTQEDADPLLNA
AP02449	GVSKILHSAGKFGKAFLGEIMKS

Table A.10: APD Gram-Negative AMPs Continued...

Definition	Sequence
AP02480	MTPLWRIMNSKPFGAYCQNNYECSTGLCRAGHCSTSHRATSETVNY

Table A.11: APD Gram-Both AMPs

Definition	Sequence
AP00001	GLWSKIKEVGKEAAKAAAKAAGKAALGAVSEAV
AP00002	YVPLPNVPQPGRRPFPTFPGQGPFNPKIKWPQGY
AP00004	NLCERASLTWTGNCGNTGHCDTQCRNWESAKHGACHKRGNWKCFCYFDC
AP00005	VFIDILDKVENAIHNAAQVGIGFAKPFEKLINPK
AP00006	GNNRPVYIPQPRPPHPRI
AP00008	RLCRIVVIRVCR
AP00009	RFRPPIRRPPIRPPFYPPFRPPIRPPIFPPIRPPFRPPLGPFP
AP00011	WNPFKELERAGQRVRDAVISAAPAVATVGQAAAIARG
AP00013	GLFDIIKKIAESF
AP00014	GLLDIVKKVVGAFGSL
AP00017	GLFDIVKKVVGTLAGL
AP00019	GLFDIAKKVIGVIGSL
AP00020	GLFDIVKKIAGHIAGSI
AP00026	FKCRRWQWRMKKLGAPSITCVRRAF
AP00027	ITPATPFTPAIITEITAAVIA
AP00030	QRFSQPTFKLPQGRLTLSRKF
AP00035	KSSAYSLQMGATAIKQVKKLFKKWGW
AP00036	DFASCHTNGGICLPNRCPGHMIQIGICFRPRVKCCRSW
AP00038	QGVRNHVTCRINRGFCVPIRCPGRTRQIGTCFGPRIKCCRSW
AP00040	QVVRNPQSCRWNMGVCIPISCPGNMRQIGTCFGPRVPCCRRW
AP00045	QGVRSYLSCWGNRGICLLNRCPGRMRQIGTCLAPRVKCCR
AP00048	SGISGPLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
AP00049	GIGALSAKGALKGLAKGLAEHFAN
AP00050	GIGASILSAGKSALKGLAKGLAEHFAN
AP00055	IIGPVLGMVGSALGGLLKKI
AP00058	GIGTKILGGVKTALKGALKELASTYAN
AP00066	IKITTMLAKLGKVLAHV
AP00069	INIKDILAKLVKVLGHV
AP00070	INVLGILGLLGKALSHL
AP00071	FLPAIFRMAAKVVPTIICSITKKC
AP00073	FLPLLAGLAANFLPKIFCKITRKC
AP00074	FLPVLAGIAAKVVPALFCKITKKC
AP00076	GILDTLKNLAISAAKGAAQGLVNKASCKLSGQC
AP00078	GILLDKLKNFAKTAGKGVLQSLLNTASCKLSGQC
AP00080	GIFSKLGRKKIKNLLISGLKNVGKEVGMDVVRTGIDIAGCKIKGEC
AP00083	GILSLVKGVAKLAGKGLAKEGGKFGLELIACKIAKQC
AP00085	SLFSLIKAGAKFLGKNLLKQGACYAACKASKQC
AP00086	GIMSIVKDVAKNAAKEAAKGALSTLSCKLAKTC
AP00088	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
AP00089	FLGALFKVASKVLPSVFCAITKKC
AP00091	GLLNTFKDWAISIAKGAGKGVLTTLSCKLDKSC
AP00094	FLPLIGRVLSGIL

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00095	LLPIVGNLLKSLL
AP00097	VLPIIGNLLNSLL
AP00099	FFPVIGRILNGIL
AP00101	FVQWFSKFLGRIL
AP00102	GSKKPVPIIYCNRRTGKCQRM
AP00104	FLPFLAKILTGVL
AP00105	FLPLFASLIGKLL
AP00107	FLPFLASLLSKVL
AP00109	VLPLISMALGKLL
AP00110	NFLGTLINLAKKIM
AP00111	FLPILINLIHKGLL
AP00113	GLLSGLKKVGKHVAKNVAVSLMDSLKCKISGDC
AP00114	SMLSVLKNLGKVGLGFVACKINKQC
AP00115	GLFLDTLKGAAKDVAGKLEGLKCKITGCKLP
AP00116	GFLDIINKLGKTFAGHMLDKIKCTIGTCPPSP
AP00117	FLPFIARLAAKVFPSIICSVTKKC
AP00118	GILDSFKGVAKGVAKDLAGKLLDKLKCKITGC
AP00121	GLMDTVKNVAKNLAGHMLDKLKCKITGC
AP00123	GLFLDTLKGLAGKLLQGLKCIKAGCKP
AP00125	KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAILS
AP00126	GGLKKLGKKLEGVGKRVFKASEKALPVAVGIKALGK
AP00129	GWLKKIGKKIERVGQNTRDATVKGLEVAQQAANVAATVR
AP00134	SWLSKTAKKLENSAKKRISEGIAIAIQGGPR
AP00136	FLPLILRKIVTAL
AP00137	LRDLVCYCRTRGCKRRERMNGTCRKGHLMYTLCCR
AP00140	SQLGDLGSGAGQGGGGGGSIRAAGGAFGKLEAAREEEFFYKKQKEQLERLKNDQIHQAEFHHQQIKE HEEAIQRHKDFLNNLHK
AP00142	GLKKLLGKLLKKLGKLLLK
AP00144	GIGKFLHSAKKFGKAFVGEIMNS
AP00145	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGAGSIGRRP
AP00146	GIGAVLKVLTTGLPALISWIKRKRQQ
AP00147	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKA
AP00149	MPCSCKKYCDPWEVIDGSCGLFNSKYICCREK
AP00150	ILPWKWPWWPWRR
AP00151	RCVCTRGFCRCVCRRGVC
AP00152	VRRFPWWWPFLRR
AP00153	RSVCRQIKICRRRGGCYYKCTNRPY
AP00154	YSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
AP00155	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG
AP00157	ALWKTMLKKLGTMALHAGKAALGAAADTISQGTQ
AP00159	ALWKNMLKGIGKLAGKAALGAVKKLVGAES
AP00160	ALWMTLLKKVLKAAAKAALNAVLVGANA
AP00163	ALWKDILKNVGKAAGKAVLNTVTDMVNQ
AP00164	ALWKTIIKGAGKMIGSLAKNLLGSQAQPES
AP00166	GWGSFFKKAAHVGKHVGKAALTHYL
AP00167	GWMSKIASGIGTFLSGMQQ
AP00168	GRPNPVNNKPTPHPRL
AP00170	VDKGSYLPRPTPPRPIYNRN
AP00171	HRHQGPIFDTRPSPFNPNQPRPGPIY
AP00172	GKPRPYSPRPTSHPRPIRV

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00174	RRCICTTRTCRFPYRRLGTCIFQNRVYTFCC
AP00175	DSHEERHHGRHGHHKYGRKFHEKHHSHRGYRSNYLYDN
AP00176	ACYCRIPACIAGERRYGTCIYQGRLWAFCC
AP00179	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
AP00180	ATCYCRTGRCATRESLSGVCEISGRLYRLCCR
AP00182	GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCYR
AP00184	RSGRGECRRQCLRRHEGQPWETQECMRRCRRRG
AP00186	GRCVCRKQLLCSYRERRIGDCKIRGVRFPFCCPR
AP00187	VVCACRRALCLPRERRAGFCRIRGRIHPLCCRR
AP00189	VSCTCRRFSCGFGERASGSCTVNGVRHTLCCRR
AP00191	QCRRLCYKQRCVTYCRGR
AP00193	DTHFPICIFCCGCCHRSKCGMCCKT
AP00194	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
AP00195	RGGRLCYCRRRFCVCVGR
AP00196	WYVKKCLNDVGICKKKCKPEEMHVKNGWAMCGKGRDCCVPAD
AP00197	QLKKCWNNYVQGHCRKICRVNEVPEALCENGRYCCLNIKELEAC
AP00198	MRILYLLFSVLFLVLQVSPGLSLPQRDMFLCRIGSCHFGRCPIHLVRVGSCFGFRSCCKSPWDV
AP00199	KYYGNGVHCTKSGCSVNWGEAFSAGVHRLANGGNGFW
AP00200	LKLKSIVSWAKKVL
AP00201	INLKALAALAKKIL
AP00203	QCIGNGGRCNENVGPPYCCSGFCLRQPGQGYGYCKNR
AP00204	ITSISLCTPGCKTGALMGCNMKTATCNCSIHVSK
AP00206	WKSESLCTPGCVTGALQTCFLQTLTCNCKISK
AP00207	TAGPAIRASVKQCQKTLKATRLFTVSCKGKNGCK
AP00208	RADTQTYQPYNKDWIKEKIYVLLRRQAQQAGK
AP00209	GVLSNVIGYLKKLGTGALNAVLKQ
AP00211	RRWCFRVCYRGFCYRKCR
AP00216	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
AP00217	GICACRRRFCPNSERFSGYCRVNGARYVRCCSRR
AP00222	VTCYCRRTRCGFRERLSGACGYRGRIYRLCCR
AP00225	ACYCRIGACVSGERLTGACGLNGRIYRLCCR
AP00228	LTCEIDRSLCLLHCRLKGYLRAYCSQQKVCRCVQ
AP00233	GWIRDFGKRIERVGQHTRDATIQTIAVAQQAANVAATLKG
AP00234	SDEKASPDKHHRFSLSRYAKLANRLANPKLLETFLSKWIGDRGNRSV
AP00236	KSCCRNTWARNCYNVCRLPGTISREICAKKCDCKIISGTTCPSDYPK
AP00239	GWASKIGQTLGKIAKVGLKELIQPK
AP00240	GLLSVLGSVAKHVLPHVVPVIAEHL
AP00249	GLVSSIGRALGGLLADVVKSKGQPA
AP00254	GLWEKIKEKASELVSGIVEGVK
AP00257	GLWQKIKSAAGDLASGIVEGIKS
AP00260	GLFVGVLAKVAAHVVPAIAEHF
AP00261	GLFVGLAKVAAHNNPAIAEHFQA
AP00262	GFVDFLKKVAGTIANVVT
AP00263	GLLQTIKEKLESLESLAKGIVSGIQA
AP00264	GRKSDCFRKSGFCAFLKCPSLTLISGKCSRFYLCCKRIR
AP00266	GKREKCLRRNGFCAFLKCPTLSVISGTCSRFQVCCKTLLG
AP00269	LSCKRGTCHFGRCPSHLIKGSCSGG
AP00272	DQYKCLQHGGFCLRSSCPSNTKLQGTCKPDKPNCCKS
AP00273	SIVPIRCRSNRDCRRFCGFRGGRCTYARQCLCGY
AP00275	GVIPCGESCVFIPCISTLLGCSCKNKVCYRN

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00276	VFQFLGKIIHHVGNFVHGFSHVF
AP00281	GLLRKGGEKIGEKLKKIGQKIKNFFQKLVPQPEQ
AP00283	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
AP00284	SHQDCYEALHKCMASHSKPFSCSMKFHMCLQQQ
AP00285	GLLCYCRKGHCKRGERVRGTCGIRFLYCCPRR
AP00288	TCESPSHKFKGPCATNRNCES
AP00291	MFFSSKKCKTVSKTFRGPCVRNAN
AP00294	DVQCGEGHFCHDQTCCRASQGGACCPYSQGVCCADQRHCCPVGF
AP00295	EVERKHPLGGSRPGRCPTVPPGTFGHCACLCTGDASEPKGQKCCSN
AP00296	LLGRCKVKSNRFHGPCLTDTHCSTVCRGEGYKGGDCHGLRRRCMCLC
AP00298	LFCKGGSCHFGGCPSHLIKVGSCFGFRSCCKWPWNA
AP00301	GILDTIKSIASKVWNSKTVQDLKRKGINWVANKLGVSPQAA
AP00303	FCTMIPIPRCY
AP00304	RVCFAIPLPICH
AP00311	CYSAAKYPGFQEFINRKYKSSRF
AP00315	SLGSFLKGVGTTLASVGKVVSDQFGKLLQAGQG
AP00316	GIVDFAKKVVGGIRNALGI
AP00322	GILDVAKTLVGKLRNVLGI
AP00323	GVLDAFRKIATVVKNVV
AP00324	GVGDLIRKAVSVIKNIV
AP00325	GVIDAAKKVVNVLKNLF
AP00326	GVGSFIHKVVSAIKNVA
AP00327	GWFDVVKHIASAV
AP00330	GWLRKAAKSVGKFYYKHKYYIKAAWQIGKHAL
AP00332	GCASRCKAKCAGRRCKGWASASFRGRCYCKCFRC
AP00333	SCASRCKGHCRARRCGYYVSVLYRGRCYCKCLRC
AP00338	PDPAKTAPKKGSKKAVTKA
AP00339	FFGWLIKGAIHAGKAIHGLIHRRRH
AP00342	AKCIKNGKGCREDQGPPFCCSGFCYRQVGWARGYCKNR
AP00346	RWKIFKKIERVGQNVRDGIIKAGPAIQVLGTAKAL
AP00350	PWNIFKEIERAVARTRDAVISAGPAVRTVAAATSVAS
AP00354	VTCDILSVEAKGVKLNDAACAAHCLFRGRSGGYCNGKRVCVCR
AP00355	ANTAFVSSAHNTQKIPAGAPFNRNLRAMLADLRQNAAFAG
AP00356	QRFIHPTYRPPPQPRRPVIMRA
AP00357	FFPIGVFCKIFKTC
AP00358	FGLPMLSILPKALCILLKRKC
AP00359	DLRFLYPRGKLPVPTPPPFNPKPIYIDMGNRY
AP00364	VDKPDYRPRPWPRNMI
AP00366	GRFKRFRKKFKKLFKKLSPVIPLLHLG
AP00367	GGLRSLGRKILRAWKKYGPIIVPIIRIG
AP00368	GLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
AP00369	RIIDLLWRVRRPQKPKFVTVWVR
AP00370	VGRFRRLRKKTRKRLKKIGKVLKWIPPIVGSIPLGCG
AP00371	GLLSRLRDFLSDRGRRLGEKIERIGQKIKDLSEFFQS
AP00374	GKVWDWIKSTAKKLWNSEPVKELKNTALNAAKNLVAEKIGATPS
AP00376	GWKDWAKKAGGWLKKKGPGMAKAALKAAMQ
AP00377	GWKDWLKKGKEWLKAKGPGIVKAALQAATQ
AP00379	DFKDWMKTAGEWLKKKGPGILKAAMAAAT
AP00382	GLVDVLGKVGGLIKKLLPG
AP00383	LLKELWTKMKGAGKAVLGKIKGLL

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00385	FKLGSFLKKAWKSKLAKKLRAKGKEMLKDYAKGLLEGGSEEVPGQ
AP00386	WLGSALKIGAKLLPSVVGLFKKKKQ
AP00388	GIWGTLAKIGIKAVPRVISMLKKKKQ
AP00389	GIWGTALKWGVKLLPKLVGMAQTKKQ
AP00390	FWGALIKGAAKLIPSVVGLFKKKQ
AP00391	FIGTALGIASAIPAIVKLFK
AP00392	YRGGYTGPIPRPPPIGRPPLRLVVCACYRLSVSDARNCCIKFGSCCHLVK
AP00394	QVYKGGYTRPIPRPPPFVRPLPGGPIGPYNGCPVSCRGISFSQARSCCSRLGRCCHVGKGYS
AP00395	HSSGYTRPLPKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHR
AP00396	RRRPRPPYLPRPRPPPFFPPRLPPRIPPGFPPRFPPRFP
AP00397	SGFVLKGYTKTSQ
AP00399	HVDKKVADKVLLLKQLRIMRLLTRL
AP00400	YPPKPESPGEDASPEEMNKYLTALRHYINLVTRQRY
AP00401	GFTQGVRNSQSCRRNKGICVPIRCPGSMRQIGTCLGAQVKCCRRK
AP00402	KTCENLANTYRGPCFTTGSCDDHCKNKEHLRSGRCRDDFRCWCTRNC
AP00403	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
AP00405	FISAIASMLGKFL
AP00408	FLFPLITSFLSKVL
AP00409	ATTGCSCPQCIIFDPICASSYKNGRRGFSSGCHMRCYNRCHGTDYFQISKGSKCI
AP00410	PKRKSATKGDEPARRSARLSARPVPKPAAKPKKAAAPKKAVKGKKAAENGDAKAEAKVQAAGDGAGNAK
AP00411	KAVAAKKSPKKAKKPATPKKAAKSPKKVKKPAAAAKKAAKSPKKATKAAKPKAAKPKAAKAKKAAPKKK
AP00412	SLQPGAPNVNNKDQPWQVSPHISRDDSGNTRTDINVQRHGENNDFEAGWSKVVRGPNKAKPTWHIGGTHRW
AP00413	SLQGGAPNFPQPSQQNGGWQVSPDLGRDDKGNTRGQIEIQNKGKDHDFNAGWGKVIRGPNKAKPTWHVGGTYRR
AP00414	SIGSALKKALPVAKKIGKIALPIAKAALP
AP00416	SLGGVISGAKKVAKVAIPIGKAVLPVVAKLVG
AP00417	SIGTAVKKAVPIAKKVGKVAIPIAKAVLSVVGQLVG
AP00418	GLRKRLRKFRNKIKEKLKKIGQKIQGFVPKLAPRTDY
AP00422	QGCKGPYTRPILRPYVRPVVSYNACTLSCRGITTTQARSCCTRLGRCCHVAKGYS
AP00424	GFLGPLLKLAAKGVAKVIPHLIPSRQQ
AP00425	GCWSTVLGGLKKFAKGGLEAIVNPK
AP00426	GVFLDALKKFAKGGMNAVLNPK
AP00427	GLLGPLLKIAAKVGSNLL
AP00428	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPFGWKSIFIQC
AP00429	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKE
AP00430	ILGKIWEGIKSLF
AP00431	TWLKKRRWKKAKPP
AP00432	KKKKPLFGLFFGLF
AP00433	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSV
AP00434	GLMSVLGHAVGNVLGGLFKS
AP00435	GWFGKAFRSVSNFYKKHKTYIHAGLSAATLL
AP00437	EFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
AP00438	GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLPCTCYRCG
AP00439	VTCFCKRPVCDSGETQIGYCRLGNTFYRLCCRQ
AP00440	VTCFCRRRGCASRERHIGYCRFGNTIYRLCCRR
AP00445	GFCRCLCRRGVCRCICTR
AP00449	SYSMEHFRWGKPV

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00450	RICRIIFLRVCR
AP00451	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK
AP00453	FLPAIVGAAGQFLPKIFCAISKKC
AP00455	FFPIVAGVAGQVLKKIYCTISKKC
AP00456	VNPIILGVLPKFVCLITKKC
AP00457	GLWETIKNFGKKFTLNILHKLKCKIGGGC
AP00459	FITLLLRKFICSITKKC
AP00461	FLPMLAGLAASMVPKLVCLITKKC
AP00470	FLPIIASVAAKVFSKIFCAISKKC
AP00474	FIHHIFRGIVHAGRSIGRFLTG
AP00475	GLNTLKKVFQGLHEAIKLINNHVQ
AP00480	VGIGTPIFSYGGGAGHVPEYF
AP00481	FFSASCVPGADKGQFPNLCRLCAGTGENKCA
AP00482	FSFKRLKGFAKKLWNSKLARKIRTKGLKYVKNFAKDMLSEGEEAPPAAEPPVEAPQ
AP00484	RGFRKHFNKLVKKVKHTISETAHVAKDTAVIAGSGAAVVAAT
AP00485	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
AP00489	SGRGKTGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAHRVGAGAPVYL
AP00490	AALKGCWTKSIPPKPCSGKR
AP00492	RQRVEELSKFSKKGAAARRRK
AP00493	NLVSGLIEARKYLEQLHRKLKNCKV
AP00494	KWKLFKKIPKFLHLAKKF
AP00496	AKKVFKRLEKLFSKIQNDK
AP00497	ILGPVLGLVSDTLDDVLGIL
AP00499	VGALAVVVWLWLWLW
AP00500	GLGSVLGKALKIGANLL
AP00501	GIGKHVGKALKGLKGLLKGLGES
AP00502	FLRFIGSVIHGIGHLVHHIGVAL
AP00503	FLGVVFKLASKVFPAVFGKV
AP00504	LAHQKPFIRKSYKCLHKRCR
AP00507	GLLSKVLGVGKKVLCGVSGLC
AP00508	GLLDSIKGMAISAGKGALQNLLKVASCKLDKTC
AP00509	VAIALKAAHYHTHKE
AP00510	ILQKAVLDCLKAAGSSLSKAAITAIYNKIT
AP00512	GYHGGHGGHGGGYNGGGGHGGHGGGYNGGGHHGGGGHG
AP00513	FLGGLIKIVPAMICAVTKKC
AP00516	IWLTALKFLGKHAAKHLAKQQLSKL
AP00517	KIKWFKTMKSIAKFIAKEQMKKHLGGE
AP00518	QYRHRCCAWGPGRKYCKRWC
AP00519	QWGRRCCGWGPGRRYCRRWC
AP00521	ILGTILGLLKGL
AP00522	INWLKLGKAIIDAL
AP00524	GIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
AP00527	ILGPVISKIGGVLGGLLKNL
AP00530	VLSKSLCTPGCITGPLQTCYLCFPTFAKC
AP00531	GKQYFPKVGGRLSGKAPLAAKTHRRLKP
AP00533	GVVDILKGAAKDIAGHLASKVMNKL
AP00535	GLGSVFGRLARILGRVIPKVAKKLGPKVAKVLPKVMKEAIPMAVEMAKSQEEQQPQ
AP00536	SVRTQDNAVNRQIFGSNGPYRDFQLSDCYLPLETNPYCNEWQFAYHWNNALMDCERAIYHGCNRTR NNFITLTACKNQAGPICNRRRH
AP00537	AEVAPAPAAAAPAKAPKKKAAAKPKKAGPS

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00538	WLNALLHHGLNCAKGVLA
AP00539	GFGCPWNRYQCHSHCRSIGRLGGYCAGSLRLTCTCYRS
AP00540	GLLDTLKGAAKNVVGSLASKVMEKL
AP00541	IDWKKLLDAAKQIL
AP00543	GVVDILKGAGKDLLAHLVGKISEKV
AP00544	GVLDIFKDAAKQILAHAAEKQI
AP00546	FLSLIPHAINAVSAIAKHN
AP00547	RSNKGFNFMVDMIQALSK
AP00548	RFGRFLRKIRRFRPKVTITIQGSARFG
AP00549	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
AP00552	GIGRKFLGGVKTTFRCGVKDFASKHLY
AP00556	GFMKYIGPLIPHAVKAISDLI
AP00557	RVKRVWPLVIRTVIAGYNLYRAIKKK
AP00558	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
AP00559	ATRVVYCNRRSGSVVGGDDTVYYEG
AP00560	TTLTLHNLCPYPVWWLVTPNNGGFPIIDNTPVVLG
AP00561	GWKIGKKLEHHGQNIRDGLISAGPAVFAVGQAATIYAAAK
AP00564	FLIGMTHGLICLISRKC
AP00567	VWPLGLVICKALKIC
AP00568	GLFSVVTGVLKAVGKNVAKNVGGSLLEQLKCKKISGGC
AP00569	FLPLLLAGLPLKLCFLFKKC
AP00570	SIITMTKEAKLPQLWKQIACRLYNTC
AP00572	GIMDTIKDTAKTVAVGLLNKLKCKITGC
AP00573	GLFSKFNKKKIKSGLIKIIKTAGKEAGLEALRTGIDVIGCKIKGEC
AP00575	GLLDTFKNLALNAAKSAGVSVLNSLSCKLSKTC
AP00576	GVLGTVKNLLIGAGKSAAQSVLKTLSCKLSNDC
AP00577	GLFTLIKGAAKLIGKTVAKEAGKTGLELMACKITNQC
AP00582	GFSSLFKAGAKYLLKSVGKAGAQQLACKAANNCA
AP00583	GVITDALKGAAKTVAAELLRKAHCKLTNSC
AP00584	VIDDLKKVAKKVRRELLCKKHHKKLN
AP00585	SIWEGIKNAGKGFLVSILDKVRCKVAGGCNP
AP00586	FLPLLFGAISHLL
AP00588	FLGSIVGALASALPSLISKIRN
AP00589	FLGALAKIISGIF
AP00594	EGTWQHGYGVSSAYSNYHHGSKTHSATVVNNNTGRQGKDTQRAGVWAKATVGRNLTEKASFYYNF W
AP00596	FLPLVTGLLSGLL
AP00598	FLSAITSLLGKLL
AP00599	GIWDTIKSMGKVFAGKILQNL
AP00600	GLLRASSVWGRKYYVDLAGCAKA
AP00601	FLSLALAALPKFLCLVFKKC
AP00605	ILPILSLIGGLLGK
AP00611	FIGPIISALASLFG
AP00612	AAEFPDFYDSEEQMGPHQEAEDEKDRADQRVLTEEEKKELENLAAMDLELQKIAEKFSQR
AP00613	RVKRFWPLVPVAINTVAAGINLYKAIRRK
AP00615	ALFSILRGLKKLGNMGQAFVNCKIYKKC
AP00616	ALSILRGLEKLAKMGIALTNCKATKKC
AP00619	GFFSTVKNLATNVAGTVIDTLKCKVTGGCRS
AP00621	GIFPKIIGKGIKTGIVNGIKSLVKGVGMKVFKAGLNNIGNTGCNEDEC
AP00624	ALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00630	GEILCNLCTGLINTLENLLTTKGADKVKDYISSLCNKASGFIATLCTKVLDFGIDKLIQLIEDKVDANAIC AKIHAC
AP00632	KYYGNGVSCNKNGCTVDWSKAIGIIGNNAAANLTTGGAAGWNKG
AP00634	KYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGNHKC
AP00639	GLIGSIGKALGGLLVDVLKPKLQAAS
AP00640	GLLGLLGSVVSHVVPAIVGHF
AP00641	GFFALIPKIISSPLFKTLLSAVGSALSSSGEQE
AP00650	GIFTKINKKKAKTGVFNIIKTIGKEAGMDVIRAGIDTISCKIKGEC
AP00651	GLFSILKGVGKIALKGLAKNMGKMGLDLVSCKISKEC
AP00654	GLLDTIKNTAKNLAVGLLDKIKCKMTGC
AP00658	FLPLVGKILSGLI
AP00659	FLPIASLLGKYL
AP00660	FWGALAKGALKLIPSLFSSFSKKD
AP00661	GILSLFTGGIKALGKTLFKMAGKAGAEHLACKATNQC
AP00662	GLFSILRGAAKFASKGLGKDLTKLGVDLVACKISKQC
AP00664	FLPAIAGILSQLF
AP00666	EGGGPQWAVGHFM
AP00667	EPHPDEFVGLM
AP00670	EPNPDEFFGLM
AP00673	VGSRYLCTPGSCWKLVCFTTTVK
AP00674	ITSVSWCTPGCTSEGGGSGCSHCC
AP00675	FELDRICGYGTARCRKKCRSQEYRIGRCPNTYACCLRKWDESLLNRTKP
AP00676	RLGNFFRKVKEKIGGGLKKVGQKIKDFLGNLVPRTAS
AP00677	GLRKKFRKTRKRIQKLGRKIGKTGRKVWKAWREYGQIPYPCRI
AP00678	RLKELITTGGQKIGEKIRRIGQRIKDFFKNLQPREEKS
AP00682	RRLRPRHQHFPSERPWPKPLPLPLPRPGPRPWPKPLPLPLPRPGLRPWPKPL
AP00684	RRLRPRRPRLPRPRPRPRPRPRSLPLPRPKPRPIPRPLPLPRPRPKPIPRPLPLPRPRPRRIPRPLPLPR PRPRPIPRPLPLPQPQPSPIPRPL
AP00686	KRFGRLAKSFLRMRILLPRRKILLAS
AP00687	KRRHWFPLSFQEFLEQLRRFRDQLPFP
AP00688	KRFHSVGSLIQRHQQMIRDKSEATRHGIRIITRPKLLLAS
AP00689	AFPPPNVPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPPFPPPIFPGPWFPPPPPFR PPPFGPPRFP
AP00691	GFFKKAWRKVKHAGRRVLDTAKGVGRHYVNNWLNRYR
AP00692	GWFKKAWRKVKNAGRRVLKGVGIHYGVGLI
AP00693	RICSRDKNCVSRPGVGSIIGRPGGGSLIGRPGGGSVIGRPGGGSPPGGGSFNDEFIRDHSDGNRFA
AP00694	AIGSILGALAKGLPTLISWIKNR
AP00696	GLFDIIKNIVSTL
AP00698	GLWQLIKDKIKDAATGFVTGIQS
AP00702	GLLGSIGNAIGAFIANKLKPK
AP00704	GLLGSIGKVLGGYLAEKLKPK
AP00707	RPDKPRPYLPRPRPPRPVR
AP00712	GFGCPLNQGACHRHCRSIRRRGGYCAGFFKQTCCYRN
AP00714	GYGCPFNQYQCHSHCSGIRGYKGGYCKGTFKQTCKCY
AP00715	RTCMKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC
AP00722	GLLNGLALRLGKRALKKIIKRLCR
AP00723	SLLSLIRKLIT
AP00727	RWCVYAYVRVRGVLVRYRRCW
AP00729	GLPVCGETCVGGTCNTPGCTCSWPVCTRN
AP00731	SFGLCRLRRGFCARGRCRFPSIPIGRCSRFVQCCRRVW
AP00736	RLGDILQKAREKIEGGLKKLVQKIKDFFGKFAPRTES

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00737	GLVTSLIKGAGKLLGGLFGSVTGGQS
AP00739	GVVTDLLKTAGKLLGNLFGSLSG
AP00741	PITYLDAILAAVRLLNQRISGPCILRLREAQPRPGWVGTLQRRREVSFLVEDGPCPPGVDCRSCEPGALQHCVGTVSIEQQPTAELRCRPLRPQ
AP00742	SPIHACRYQRGVCIPGPCRWPYYRVGSCGSGLKSCCVRNRWA
AP00743	RYHMQCGYRGTFCTPGKCPYGNAYLGLCRPKYSCCRWL
AP00744	GLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCCRSRWYS
AP00745	MTPFWRGVSLRPVGASCRDNSECITMLCRKNRCFLRTASE
AP00748	DIQIPGIKKPTHRDIIIPNWNPNVRTQPWQRFGGNKS
AP00749	EADEPLWLYKGDNIERAPTTADHPILPSIIDDVKLDPNRRYA
AP00750	EIRLPEPFRFPSPTVPKPIDIDPILPHPWSPRQTYPIIARRS
AP00752	DKLIGSCVWGATNYTSDCNAECKRRGYKGGHCGSFWNVNCWCEE
AP00753	VQETQKLAKTVGANLEETNKKLAPQIKSAYDDFVKQAQEVQKKLHEAASKQ
AP00754	ETESTPDYLKNIQQQLEEYTKNFNTQVQNAFDSDKIKSEVNNFIESLGKILNTEKKEAPK
AP00755	ENFFKEIERAGQRIRDAIISAAPAVETLAQAQKIIKGGD
AP00764	GLRSKIWLWVLLMIWQESNKFKKM
AP00765	MHDFWVLWVLLEYIYNSACSVLSATSSVSSRVLNRSLQVKVVKITN
AP00766	IYWIADQFGIHLATGTARKLLDAMASGASLGTAFAAILGVTLPAWALAAAGALGATAA
AP00767	VAGALGVQTAAATTIVNVILNAGTLVTVLGIIASIASGGAGTLMTIGWATFKATVQKLAKQSMARAIAY
AP00768	PNWTKIGKCAGSIAWAIGSGLFGGAKLIKIKKYIAELGGLQKAAKLLVGATTWEEKLHAGGYALINLAAELTGVAGIQANCF
AP00769	GLLGAMFKVASKVLPHVVPAITEHF
AP00772	FRGLAKLLKIGLKSFARVLKKVLPKAAKAGKALAKSMADENAIRQQNQ
AP00773	GKFSVFGKILRSIAKVFKGVGKVRKQFKTASDLDKNQ
AP00777	GKGRWLERIGKAGGIIIGGALDHL
AP00779	GRRKRKWLRRIGKGVKIIGGAALDHL
AP00780	GRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQA
AP00781	FLGALIKGAIHGGRFIHGMIQNHH
AP00782	GWGSIFKHGRHAAKHIGHAAVNHYL
AP00784	FFRLLFHGVHHVGKIKPRA
AP00785	GWKSVFRKAKKVGKTVGGLALDHYL
AP00787	GWRLLLKKAEVKTVGKLALKHYL
AP00788	AGWGSIFKHIFKAGKFIHGAIQAHND
AP00789	GFWGKLFKLGLHGIGLLHLHL
AP00791	GWKKWLRKGAKHLGQAAIKGLAS
AP00792	FLGLLFHGVHHVGKWIHGLIHGHH
AP00796	IIGPVLGLVGKPLESLLE
AP00805	RSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECDHYVNSKVDPIIHELEGGTAPKDVCTKLNECP
AP00806	HHQELCTKGDDALVTELECIRLRISPETNAAFDNAVQQLNCLNRACAYRKMCATNNLEQAMSVYFTNEQIKEIHDAATACDPEAHHEHDH
AP00807	NRWYCNSAAGGVGGAAGCVLAGYVGEAKENIAGEVRKGWGMAGGFTHNKACKSFPGSGWASG
AP00809	GIKCRFCCGCCTPGICGVCCRF
AP00812	FAEPLPSEEEGESYSKEPPEMEKRYGGFM
AP00814	GLGSILGKILNVAGKVGKTIGKVADAVGNKE
AP00816	GLGSFFKNAIKIAGKVGSTIGKVADAIGNKE
AP00817	FLPLLASLFSRLL
AP00818	FLPLIGKILGTIL
AP00822	GIFNVFKGALKTAGKHVAGSLLNQLKCKVSGEC
AP00824	SILPTIVSFLSKFL

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP00834	KVNANAIKKGGKAIGKGFKVISAASTAHDVYEHIKNRRH
AP00835	GKIPVKAIKKGGQIIGKALRGINIASTAHDIISQFKPKKKKNH
AP00836	KVPIGAIKKGGKIIKKGLGVIGAAGTAHEVYSHVKNRH
AP00840	KGIGSALKKGGKIIKGGLGALGAIGTGQQVYEHVQNRQ
AP00841	TTHSGKYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAIPGKC
AP00842	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLGGAIPGKC
AP00845	KSYGNGVHCNKKKCWVDWGSAISTIGNNSAANWATGGAAGWKS
AP00848	KNYGNGVHCTKKGCSVDWGYAWANIANNSVMNGLTGGNAGWHN
AP00849	TSYGNGVHCNKSKCWIDVSELETYKAGTVSNPKDILW
AP00850	KYYGNGVSCNSHGCSVNWGQAWTCGVNHLANGGHGVC
AP00852	KYYGNGLSCSKKGCTVNWGQAFSCGVNRVATAGHHKC
AP00853	ATRSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH
AP00857	SSMKLSFRARAYGFRGPGPQL
AP00863	FLPIVGKLLSGLSGLL
AP00866	FLPIIAKVLSGLL
AP00867	FLPVIAGLLSKLF
AP00869	ILPLVGNLLNDLL
AP00875	FLSSIGKILGNLL
AP00876	FLSIIAKVLGSLF
AP00877	FLGSLIGAAIPAIKQLLGLKK
AP00878	FLPILASLAAKFGPKLFCLVTKKC
AP00879	GRLRNLIEKAGQNIRGKIQGIGRRIKDILKNLQPRPQV
AP00884	QLKVDLWGTRSGIQPEQHSSGKSDVRRWRSRY
AP00891	IIGLVSKGTCVLVKTVCKKVLKQG
AP00892	PDITKLNIKKLTKATCKVISKGASMCKVLFDKKKQE
AP00893	DVKGMKKAIKGILDCVIEKGYDKLAAKLKKVIQQLWE
AP00894	GLLDFVTGVGKDIFAQLIKQI
AP00895	KRFKKFFKKLKNSVKKRAKKFFKKPRVIGVSIPF
AP00898	FLSGIVGMLGKLF
AP00900	FLSHIAGFLSNLF
AP00911	FLSLIPHIVSGVAALAKHL
AP00915	QQCGRQAGNRRCANNLCCSQYGYCGRTNEYCCTSQGCQSQCRRCG
AP00916	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
AP00927	IYFIADKMGIQLAPAWYQDIVNWVSAGGTLTTGFAIIVGVTVPAWIAEAAAAFGIASA
AP00928	NKGCATCSIGAACLVDGPIPDFEIAGATGLFGLWG
AP00929	MAKEFGIPAAVAGTVINVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVIA W
AP00930	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGVKHSSGGGGSYHC
AP00931	LAGYTGIASGTAKKVVDAIDKGAAAFVIISIISTVISAGALGAVSASADFIILTVKNYISRNLKAQAVIW
AP00964	GLWSKIKEAAKAAGKAALNAVTGLVNQGDQPS
AP00969	GLWSKIKEAAKTAGLMAMGFVNDMV
AP00973	LLGMIPLAISAISALSKL
AP00987	SRWPSPGRPRPFPGRPKPIFRPRPCNCYAPPCPCDRW
AP00990	RNCESLSHRFKGPCTRDSN
AP00993	GIFSSRKCKTPSKTFKGICTRDSNCDTSCRYEGYPAGDCKGIRRRCMCSKPC
AP00994	GIFSNMYARTPAGYFRGPAGYAAN
AP00996	ISLEICAIFHDN
AP00998	ALPKKLKYLNLFNDGFNYMGVV
AP01001	NRWWQGVVPTVSYECRMNSWQHVFTCC
AP01003	FKSWSFCTPGCAKTGSFNSYCC

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01005	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
AP01007	KWKLFKKIGAVLKVL
AP01009	DYDWSLRGPPKCATYGQKCRTWSPRNCCWNLRCKAFRCRPR
AP01010	SMWSGMWRRKLKKLRNALKKKLKGE
AP01011	GLFGKLIKKFGRKAISYAVKKARGKH
AP01012	SWKSMAKKLKEYMEKLKQRA
AP01014	GLKDKFKSMGEKLKQYIQTWKAKF
AP01016	GFFGKMKEYFKKFGASFKRRFANLKKRL
AP01018	QAFQTFKPDWNKIRYDAMKMQTSLGQMKKRFNL
AP01019	GETFDKLKEKLKTFYQKLVEKAEDLKGDLKAKLS
AP01128	GSKILHSAGKFGKAFLGEINKS
AP01129	GLGSLVGNALRIGAKLL
AP01130	GMASKAGSVLGKVAKVALKAAL
AP01131	MSWLNFLKYIAKYGKKAVSAAWKYKGKVLEWLNVGPTLEWVWQKLKKIAGL
AP01142	KPYCSCKWRCGIGEEEKGICHKFPIVTYVCCRRP
AP01148	IATQCRIRGGFCRVGSCRFPHIAIGKCATFISCCGRAY
AP01151	GTWDDIGQGIGRVAYWVGKALGNLSDVNQASRINRKKKH
AP01153	YLAFRCGRYSPCLDDGPNVNLYSCCSFYNCHKCLARLENCPKGLHYNAYLKVCDWPSKAGCT
AP01154	AIKLVQSPNGNFAASFVLDGTKWIFKSKYYDSSKGYWVGIYEVWDRK
AP01155	ESVFSKIGNAVGPAAYWILKGLGNMSDVNQADRINRKKH
AP01156	NKLAYNMGHYAGKATIFGLAAWALLA
AP01157	QRGSRGQRCGPGEVFNQCGSACPRVCGRPPAQACTLQCVSGCFCRRGYIRTQRGGCIPERQCHQR
AP01158	ALYKKFKKKLLKSLKRL
AP01160	QDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
AP01161	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRVCRTGRSRWRDVCRNFMRRYQSRVTQGLVAGET AQQICEDLR
AP01167	LTTKLWSSWGYYLGKKARWNLKHPYVQF
AP01168	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL
AP01169	NRWGDTVLSAASGAGTGIKACKSFGPWGMAICGVGGAAIGGYFGYTHN
AP01170	YSSKDCLKDIGKGIGAGTVAGAAGGGLAAGLGAIPGAFVGAHFGVIGGSAACIGGLLGN
AP01171	YSGKDCLKDMGGYALAGAGSGALWGAPAGGVGALPGAFVGAHVGAIAGGFACMGGMIGNKFN
AP01172	KRGPNCVGNFLGGLFAGAAAGVPLGPAGIVGGANLGMVGGALTCL
AP01174	KVSGGEAVAAIGICATASAAIGGLAGATLVTPYCVGTWGLIRSH
AP01175	GMSGYIQGIPDFLKGYLHGISAANKHKKGRLGY
AP01176	TTPACFTIGLGVGALFSAKFC
AP01177	FNRGGYNFGKSVRHVVDAIGSVAGILKSIR
AP01178	GAWKNFWSSLRKGFYDGEAGRAIRR
AP01179	NGVYCNKQKCWVDWSRARSEIIDRGVKAYVNGFTKVLGGIGGR
AP01180	NPKVAHCASQIGRSTAWGAVSGA
AP01181	AYPGNGVHCGKYSCTVDKQTAIGNIGNNAA
AP01182	FTPSVSFSQNGGVVEAAAQRGYIYKKYPKGAKVPNKVKMLVNIRGKQTMRTCYLMSWTASSRTAKY YYYI
AP01183	ATYYGNGLYCNKEKCWVDWNQAKGEIGKIIVNGWVNHGPWAPRR
AP01185	ENDHRMPNNLNRPNNLSKGGAKCGAAIAGGLFGIPKGPLAWAAGLANVYSKCN
AP01186	KTYYGTNGVHCTKKSLWGKVRLKNVIPGTLCRKQSLPIKQDLKILLGWATGAFGKTFH
AP01187	MNFLKNGIAKWMTGAELQAYKKKYGCLPWEKISC
AP01188	MLAKIKAMIKKFPNPYTLAAKLTTYEINWYKQQYGRYPWERPVA
AP01189	APAGLVAKFGRPIVKKYYKQIMQFIGEGSAINKIIPWIARMWRT
AP01192	SDCNINSNTAADVILCFNQVGSCALCSPTLVGGPVP
AP01193	DIDITGCSACKYAAGQVCTIGCSAAGGFICGLLGITIPVAGLSCLGFVEIVCTVADEYSGCGDAVAKEA CNRAGLC

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01194	CSTNTFSLSDYWGNNGAWCTLTHECMAWCK
AP01195	KRGSGWIATITDDCPNSVFVCC
AP01196	GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGGALQTVGQGLIDHGPVNVFIPVLIGPSW NGSGSGYNSATSSSGSGS
AP01198	LSCDEGMLAVGGLGAVGGPWGAAVGVLVGAALYCF
AP01199	KYYGNGVHCGKKTCYVDWGQATASIGKIIVNGWTQHGPWAHR
AP01201	KGGSGVIHTISHECNMNSWQFVFTCCS
AP01204	GKNGVFKTISHECHLNTWAFLATCCS
AP01205	STPVLASVAVSMELLPTASVLYSDVAGCFKYSAKHHC
AP01206	CTFTLPGGGGVCTLTSECIC
AP01213	EFRGSIVIQGTKEGKSRPSLDIDYKQRVYDKNGMTGDAYGGLNIRPGQPSRQHAGFEFGKEYKNGFIK GQSEVQRGPGGRLSPYFGINGGFRF
AP01215	FVPYNPPRPYQSKPFPSFPGHGPFNPKIQWPYPLPNPGH
AP01218	GFRDVLKGAAKAFVKTVAGHIANI
AP01220	GIKDWIKGAAKKLIKTVASNIANQ
AP01227	VGIGGGGGGGGGGSCGGQGGGCGGCSNGCSGGNGGSGGSGSHI
AP01228	ASGRDIAMAIGTLSGQFVAGGIGAAAGGVAGGAIYDYASTHKPNPAMSPSGLGGTIKQKPEGIPSEAW NYAAGRLCNWSPNNLSDVCL
AP01229	GDVNWVDVGKTVATNGAGVIGGAFGAGLCGPVCAGAFAVGSSAAVAALYDAAGNSNSAKQKPEGL PPEAWNYAEGRMCNWSPNNLSDVCL
AP01230	DGNDGQAELIAIGSLAGTFISPGFGSIAGAYIGDKVHSWATTATVSPSMSPSGIGLSSQFGSGRGTSSASS SAGSGS
AP01231	GGAPATSANAAGAAAIVGALAGIPGGPLGVVVGAVSAGLTTGIGSTVGSGSASSSAGGGS
AP01232	MNLNGLPASTNVIDLRGKDMGTYIDANGACWAPDTPSIIMYPGGSGPSYSMSSSTSSANSGS
AP01233	QKKPPRPPQWAVGHFM
AP01234	FSKYERQKDKRPYSERKNQYTGPQFLYPPERIPPQKVIKWNEEGLPIYEIPGEGGHAEPAAA
AP01235	FNKLKQGSSKRTCAKCFRKIMPSVHELDERRRGANRWAAGFRKCVSSICRY
AP01236	GLLDFAKHVIGIASKL
AP01238	NPLIPAIYIGATVGPSVWAYLVALVGAAAVTAANIRRASSDNHSCAGNRGWCRSKCFRHEYVDTYYSA VCGRYFCCRSR
AP01240	ALKAALLAILKIVRVIKK
AP01241	FASLLGKALKALAKQ
AP01242	GLLSFLPKVIGVIGHLIHPPS
AP01243	KGAPCAKKPCCGPLGHYKVDCSTIPDYPCCSKYGFCGSGPQYCG
AP01245	AVTCNTVVSSLAPCVPFFAGSAAQPTAACCNGVRSLNSAARTTPDRRTACNCIKSSASSIGLNYNKAA KLPSRCTVNVTVPISPSVNCAT
AP01246	FLPKTLRKFFCRIRGGRCAVLNCLGKEEQIGRCSNSGRKCCRKKK
AP01247	FMPIIGRLMSGSL
AP01248	INMKASAAVAKKLL
AP01249	GILDAIKAIAKAAG
AP01253	GLMSLFKGVLKTAGKHIFKNVGGSLLDQAKCKITGEC
AP01258	GLMDVFKGAAKNLLASALDKIRCKVTKC
AP01260	IIGHLIKTALGMLGL
AP01261	IIEKLVNTALGLLSGL
AP01262	GLADFLNKAVGKVVDFVKS
AP01263	FLPLVTMLLGKLF
AP01264	RIGVLLARLPKLFSLFKLMGKKV
AP01266	AVDLAKIANKVLSSLF
AP01267	RRTCHCRSRCLRRESNSGSCNINGRIFSLCCR
AP01268	FLPVILPVIGKLLSGIL
AP01269	GFLSILKKVLPKVMAHMK
AP01277	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01283	MRKEFHNVLSSGQLLADKRPARDYNRK
AP01285	AGDPLADPNSQIVRQIMSNAAWGPPLVPERFRGMAVGAAGGVTQTVLQGAAAHMPVNVPIPKVPMG PSWNGSKG
AP01294	GLGGAKKNFIIAANKTAPQSVKKTFSCKLYNG
AP01296	FMPILSCSRFKRC
AP01298	GLFTLIKCAYQLIAPTVACN
AP01299	GLFTLIKGAAKLIGKTVPKKQARLGMNLWLVKLPTNVKT
AP01300	ATAVDFGPHGLLPIRPIRIRPLCGKDKS
AP01303	VIPFVASVAAEMMQHVYCAASKKC
AP01304	GLLSGILGAGKHIVCGLSGPCQSLNRKSSDVEYHLAKC
AP01305	FLPPSPWKETFRTS
AP01306	TSRCYIGYRRKVVCS
AP01307	GCSRWIIGIHGQICRD
AP01308	GLLSGTSVRGSI
AP01315	ARLKKCFNKVTGYCRKKCKVGERYEIGCLSGKLCCAN
AP01316	NPANPLNLKKHHGVFCDVCKALVEGGEKVGDDDLDAWLDVNIGTLCWTMLLPLHHECEEELKKVKK ELKKDIENKDSPDKACKDVDLC
AP01317	GAILCNLCKDTVKLVENLLTVDGAQAVRQYIDNLCGKASGFLGTLCEKILSFGVDELVKLIENHVDPV VVCEKIHAC
AP01318	IPVLCPVCTSLVGKLIDLVLGGAVDKVTDYLETLCAKADGLVETLCTKIVSYGIDKLIEKILEGGSAKLI CGLIHAC
AP01319	DPVTYIRNGGICQYRCIGLRHKIGTCGSPFKCCK
AP01321	APGNKAECEREKGYCGFLKCSFPFVVSGKCSRFFFCCKNIW
AP01322	IPRPLDPCIAQNGRCFTGICRYPYFWIGTCRNGKSCCRRR
AP01323	LPVNEAQCRQVGGYCGLRICNFPSRFLGLCTRNHPCCSRVWV
AP01324	GPDSCNHDRGLCRVGNCNPGEYLAKYCFEPVILCCKPLSPTPTKT
AP01325	QPFIPRPIDTCRLRNGICFPGICRRPYYWIGTCNNGIGSCCARGWRS
AP01326	SKGKKANKDVELARG
AP01327	LFGLIPSLIGGLVSAFK
AP01328	GFIFHIIKGLFHAGKMIHGLV
AP01331	IFGAILPLALGALKNLIK
AP01339	FLSFPTTKTYFPHFDLSHGSAQVKGHGAK
AP01340	DAECEICKFVIQQVEAFIESNHSQAEIQKELNKLCSSVPSITQTCLSIARMVPYIIKKLEEHNSPGQVCQG LHLCKSS
AP01343	TESYFVFSVGM
AP01345	FFGTALKIAANVLPTAICKILKKC
AP01346	FFPLVLGALGSILPKIF
AP01347	FIITGLVRGLTKLF
AP01348	SLSRFLSFLKIVYPPAF
AP01350	FLSLLPSIVSGAVSLAKKL
AP01353	FWGHIWNAVKRVGANALHGAVTGALS
AP01354	GFWKKVGSAAWGGVKAAAKGAAVGGLNALAKHIQ
AP01355	RESPSSRMECYEQAERYGYGGYGGGRYGGGYGSGRGQPVGQGVERSHDDNRNQPR
AP01356	KTCMTKKEGWGRCLIDTTCAHSCRKYGYMGGKCQGITRRCYCLLNC
AP01358	VTCDLLSFEAKGFAANHSLCAAHCLAIGRRGGSCERGVCICRR
AP01364	ATCDLLSAFGVGHAACAAHCIGHGYRGGYCNSKAVCTCRR
AP01365	AAKPMGITCDLLSLWKVGHAACAAHCLVLGDVGGYCTKEGLCVCKE
AP01367	VTCNIGEWVCVAHCNSKSKKSGYCSRGVCYCTN
AP01368	ATCDLFSFRSKWVTPNHAACAAHCLLRGNRGGRCKGTICHCRK
AP01372	SKCKCSRKGPKIRYSDVKKLEMKPKYPHCEEKMVIITTKSVSRYRGQEHCLHPKLQSTKRFIKWYNA WNEKRRVYEE

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01374	NLAKGKEESLDSDLYAELRCMCIKTTSGIHPKNIQSLEVIGKGTHCNQVEVIATLKDGRKICLDPDAPRI KKIVQKKLAGDES
AP01376	YENPYGCPTDEGKCFDRCNDSEFEGGYCGGSYRATCVCYRT
AP01377	FFRHLFRGAKAIFRGARQGWRAHKVVSRYRNRDVPETDNNQEEP
AP01378	AREASKSLIGTASCTCRRAWICRWGERHSGKCIDQKGSTYRLCCRR
AP01379	ILENLLARSTNEDREGSIFDTGPIRRPKPRPRPRPEG
AP01380	YDLSKNCRLRGGICYIGKCPRRFFRSGSCSRGNVCCLRFG
AP01381	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
AP01382	QKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
AP01398	YQEPVLGPVRGPFPIIV
AP01400	RPKHPIKHQGLPQEVLNENLLRF
AP01405	GLLGGLLGPLLGGGGGGGGLL
AP01407	HNKQEGRDHDKSKGHFHRVVIHHKGGKAH
AP01423	FLPAVLRVAAKIVPTVFCAISKKC
AP01425	GLLGSLFGAGKKVACALSGLC
AP01428	GFKGAFKNVMFGIAKSAGKSALNALACKIDKSC
AP01429	GLLDSFKNAMIGIAKSAGKTALNKIACKIDKTC
AP01432	FMGGLIKAATKIVPAAYCAITKKC
AP01434	FFGSVLKLIPKIL
AP01440	FFPIIAGMAAKVICAITKKC
AP01445	FMGSALRIAAKVLPAALCQIFKKC
AP01447	FLPGLIAGIAKML
AP01448	FLPIALKALGSIFPKIL
AP01449	FLGAIAAALPHVINAVTNAL
AP01454	IIPLPLGYFAKKT
AP01455	FFPLALLCKVFKKC
AP01456	VGKTWIKVIRGIGKSKIKWQ
AP01457	GLKDIFKAGLGSLVKGIAAHVAN
AP01461	ILGKLLSTAAGLLSNL
AP01465	VNWKKVLGKIIKVAK
AP01470	AQRCGDQARGAKCPNCLCCGKYGFCGSGDAYCGAGSCQSQCRGCR
AP01471	RPKPQQFFGLM
AP01472	QLYENKPRRPYIL
AP01474	YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY
AP01475	ECWMDGHCRLLCKDGEDSIIRCRNRKRCC
AP01476	ACDTATCVTHRLAGLLSRSGGVVKNNFVPTNVGSKAF
AP01477	HSDAVFTDNYTRLRKQMAVKKYLNSILN
AP01479	YRQSMNNFQGLRSFGCRFGTCTVQKLAHQIYQFTDKDKDNVAPRSKISPQGY
AP01491	QWGYGGYGRGYGGYGGYGRGYGGYGGYGRGYGGYGRGMYGGYGRPYGGYGWGK
AP01493	ASIIKTTIKVSKAVCKTLTCICTGSCSNCK
AP01496	IPPFIKKVLTTVF
AP01509	FLPKMSTKLRVPYRRGTKDYH
AP01510	GILKKFMLHRGTKVYKMRTLSKRSH
AP01511	TITLSTCAILSKPLGNNGYLCTVTKECMPSSCN
AP01513	GKNPTLQCMGNRGFCRPSCKKGEQAYFYCRTYQICCLQSHVRISLTGVEDNTNWSYEKHWPRIP
AP01514	GVNMYIRQIYDTCWKLKGHCRNVCGKKEIFHIFCGTQFLCCIERKEMPVLFVK
AP01515	AACSDRAHGHICESFKSFCKDSGRNGVKLRANCKKTCGLC
AP01516	LNLKGIFKKVASLLT
AP01517	INLLKIAKGIIKSL

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01520	SSFSPPRGPPGWGPPCVQQPCPKCPYDDYKCPTCDKFPECEECPHISIGCECGYFSCECPKPVCEPCE SPIAELIKKGGYKG
AP01521	RFRLPFRRPPIRIHPPPFYPPFRRFL
AP01522	TYMPVEEGEYIVNISYADQPKKNSPFTAKKQPGPKVDLSGVKAYGPG
AP01523	AVDFSSCARMDVPGLSKVAQGLCISSCKFQNCGTGHCEKRGGRPTCVCDRCGRGGGEWPSVPMPKG RSSRGRRHS
AP01524	DIDFSTCARMDVPILKKAAQGLCITSCSMQNCGTGSCKKRSGRPTCVCYRCANGGGDIPLGAL
AP01529	GAARKSIRLHRLYTWKATIYTR
AP01530	GSCSCSGTISPYGLRTCRATKTKPSHPTTKETHPQTLPT
AP01531	GKWGWIYITILFADVGGFKSSRHPEERRVQERRFKRITRGPD
AP01533	KRRGSVTTRYQFLMIHLLRPKKLFA
AP01539	SILSGNFGVGKKIVCGLSGLC
AP01540	AGANDLCQECEDIVHLLTKMTKEDAFQDTIRKFLEQECDILPLKLLVPRCRQVLDVYLPLVIDYFQGQI KPKAICSHVGLC
AP01541	AVLDILKDVGKGLLSHFMEKV
AP01542	AVLDFIKAAGKGLVTNIMEKVG
AP01543	KPWRFRRAIRRVRWRKVAPYIPFVVKTVGKK
AP01544	IFGAIAGLLKNIF
AP01545	FFGHLFKLATKIIPSLFQ
AP01547	GVIKSVLKGVAKTVALGML
AP01548	ADTLACRQSHQSCSFVACRAPSVDIGTCRGGKLKCCKWAPSS
AP01549	VLLFLFQAAPGSADAPFADTAACRSQGNFCRAGACPPTFAASGSCHGGLLNCCAK
AP01550	SAVGRHGRRFGLRKHRKH
AP01552	QIVDCWETWSRCTKWSQGGTGTLWKSCNDRCKELGRKRGQCEEKPSRCPLSKKAWTCICY
AP01554	DFGCGQGMIFMCQRRCMRLYPGSTGFCRGFRCMCDTHIPLRPPFMVG
AP01555	TCRYWCKTPENQTYCCEDEREIPSKVGLKPGKCPPVRPVCPPTRGFFEPPKTCSNDGSCYGADKCCF DRCLGEHVCKPIQTRG
AP01556	GCFEDWSRCSPSTSRGTGVLWRDCDSYCKVCFKADRGECFDSPSLNCPQRLPNNKQCRCINARTAKD NRNPTCWA
AP01557	DHHHDHGHDDHEHEELTLEKIKEKIKDYADKTPVDQLTERVQAGRDYLLGKGARPSHLPARVDRHLS KLTAAEKQELADYLLTFLH
AP01558	LGAWLAGKVAGTVATYAWNRYV
AP01559	ATYDGKCYKKDNICKYKAQSGKTAICKCYVKVCPRDGAKCEFDSYKGKCYC
AP01564	ATCDLLSKWNWNHTACAGHCIAKGFKGGYCNDKAVCVCRN
AP01565	DDMTMKPTPPPQYPLNLQGGGGGGSGDGFGFAVQGHQKVWTSDNGRHEIGLNGGYGQHLGGPYGN SEPSWKVGSTYTYRFPNF
AP01566	QRPYTQPLIYYPPPPTPPRIYRA
AP01569	SNDSLWYGVGQEMGKQANCITNHPVKHMIIPGYCSKILG
AP01570	GNAACVIGCIGSCVISEGIGSLVGTAFTLG
AP01571	IFGSIYHRKCVVKNRCETVSGHKTCKDLTCCRAVIFRHERPEVCRPQT
AP01572	WNPFKKIANRNCYPKTTCETAGGKKTCKDFSCCQIVLFGKKTRAKCTVVTS
AP01573	GWFKKTFHKVSHAVKSGIHAGQRGCSALGF
AP01574	SWFSRTVHNVGNAVRKGIHAGQGVCSGLGL
AP01575	NLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGE GCDRDGKYGFYTHVFRLKKWIQKVIDQFGE
AP01576	RVPPYLGRDCKHWCRDNNQALYCCGPPGITYPPFIRKHPGKCPSVRSTCTGVRSSRPKFCPHDDACE FRSKCCYDACVKHHVCKTVEFY
AP01577	FLLFPLMCKIQGKC
AP01578	GIHDILKYGKPS
AP01579	FVLPLVMCKILRKC
AP01580	AQEPVKGPVSTKPGSCPIILIRCAMLNPPNRCLKDTDCPGIKKCCEGSCGMACFVPQ
AP01582	CIAKGNGCQPSGVQGNCCSGHCHKEPGWVAGYCK
AP01583	GWANTLKNVAGGLCKITGAA

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01585	KSCCRSTQARNIYNAPRFAGGSRPLCALGSGCKIVDDKKTPPND
AP01586	KSCCRSTTARNIYNGCRVPGTARPVCAKKSGCKIQEAKKCEPPYD
AP01587	AQCGAQGGGATCPGGLCCSQWGWCGSTPKYCGAGCQSNCK
AP01589	KDRPKKPGLCPPRPQKPCVKECKNDDSCPGQQKCCNYGCKDECRDPIFVG
AP01590	DHYICAKKGGTCNFSPCPLFNRIEGTCYSGKAKCCIR
AP01591	KCWNLRGSCREKCIKNEKLYIFCTSGKLCCLKPKFQPNMLQR
AP01592	GIRNTVCFMQRGHCRLFMCRSGERKGDICSDPWNRCCVSSSIKNR
AP01593	CKQSCSFGPFTFVCDGNTK
AP01595	CANSCSYGPLTWSCDGNTK
AP01597	SVSCLRNKGVCMPGKCAPKMKQIGTCGMPQVKCCKRK
AP01599	WNPFRKLYRKECNDVTSCDTVSGVKTCTKKNCCHRKFFGKTILKAPECTVIS
AP01600	RARAPHKAWYNCMTDAGISGAIAGAVAGCAATIEIGCVEGAIAGIGPSGIASMIAALWTCRSKY
AP01601	YVPPVQKPHPNGPKFPTFP
AP01602	LVLKYCPKIGYCSNTCSKTQIWATSHGCKMYCCLPASWKWK
AP01603	SSSGWLCTLTIECGTIICACR
AP01604	DAPGHPGKHYLQVNVPSDVRTIGVAGGGVQQCFRVTPGAWNDTRALVSNGAQVEVWGYTVADCAN RTTANQKYYDKAAAPSDSSTYFWFTLKNLRV
AP01606	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR
AP01607	ADRGWIKTLTKDCPNVISSICAGTIITACKNCA
AP01609	KCKWWNISCDLGNNGHVCTLSHECQVSCN
AP01612	SASIVKTTIKASKKLCRGFTLTCGCHFTGKK
AP01613	LPRDTSRCVGYHGYCIRSKVCPKPFAAFGTCSWRQKTCCVDTTSDFHTCQDKGGHCVSPKIRCLEEQ LGLCPLKRWTCCKEI
AP01614	WRSLGRTLLRLSHALKPLARRSGW
AP01615	SASVLKTSIKVSKKYCKGVTLTCGCNITGGK
AP01616	SLGPAIKATRQVCPKATRFVTVSCKKSDCQ
AP01618	GTTVVNSTFSIVLGNKGYICTVTVECMRNCSK
AP01619	AANFGPSVFTPEVHETWQKFLNVVVAALGKQYH
AP01620	VDKPPYLPRPPPPRRIYNNR
AP01621	CAWYNISCRLGNKGAYCTLTVECMPSCN
AP01622	GLGSLLGKAFKIGLKTVGKMMGGAPREQ
AP01623	GFKLKGMARISCLPNGQWSNFPPKCIRECAMVSS
AP01624	HAEHKVKIGVEQKYGQFPQGTEVTYTCSGNYFLM
AP01625	LQDAALGWGRRCPQCPRCPSCPSCPRCPRCPRCKCNPK
AP01632	ATPATPTVAQFVIQGSTICLVC
AP01633	RRWVRRVRRWVRRVVRVVRRWVRR
AP01634	INWKKIFEKVKNLV
AP01637	INWKKIASIGKEVLKAL
AP01638	INWKKIAEVGGKILSSL
AP01641	IDWLKLGKMVMDVL
AP01642	LCLDQKPEMEPFRKDAQQALEPSRQRRWLHRRCLSGRGFCRAICSIFEEPVRGNIDCYFGYNCCRRMF SHYRTS
AP01644	GAFGNFLKGVAKKAGLKILSIAQCKLSGTC
AP01647	RCTCTTIISSSSTF
AP01648	GKLNLFLSRLEILKLFVGAL
AP01650	YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG
AP01651	LVATGMAAGVAKTIVNAVSAGMDIATALSLFSGAFTAAGGIMALIKKYAQKKLWKQLIAA
AP01652	LIDHLGAPRWAVDTILGAIAVGNLASWVLALVPGPGWAVKAGLATAAAIVKHQGKAAAAAW
AP01658	NALSMPRNKCNRALMCFG
AP01660	WNSNRRFRVGRPPVVGRPGCVCFRAPCPCSNY
AP01663	RRTCRCRFGRCFRRESYSGSCNINGRIFSLCCR

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01667	RRICRCRIGRCLGLEVYFGVCFLHGRLARRCCR
AP01673	GICRCICTRGFCRCICVL
AP01676	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
AP01693	KFCEKPSGTWSGVCGNSGACKDQCIRLEGAKHGSCNYKPPAHRCICYYEC
AP01696	SPAGCRFCCGCCPNMRGCGVCCRF
AP01701	MKTFSVAVAVAIVLAFICTQESSALPVTGVEELVELVSSDDPVADHQELPVELGERLFNIRKKRASPKC TPYCYPTRDGVFCGVRCDF
AP01708	SFLTTFKDLAIKAAKSAGQSVLSTLSCKLSNTC
AP01710	GIFSTVFKAGKGIVCGLTGLC
AP01713	SRSGRGSGKGGRGGSRGSSGSRGSKGPSGSRGSSGSRGSKGSRGGRSGRGSTIAGNGNRNNGGTRTA
AP01715	PSCVCSGFETSGIHFC
AP01718	FKVQNQHGQVVKIFHH
AP01720	RPRPNYRPRPIYRP
AP01724	GTPGFQTPDARVISRFGFN
AP01729	GSQLVYREWVGHSNVIKGPP
AP01739	GIGGVLLGAGKATLKGLAKVLAEKYAN
AP01743	GIGGALLSVGKLALKGLANVLADKFAN
AP01745	ERILDLRKTKKSCKNGEVLGCVSGHGPPGCSENECGMGPRPKACFFDCHYGCWCTGKLYRRKRDRK CVPKHECLL
AP01746	FLGGILNTITGLL
AP01747	SFPFFPPGICKRLKRC
AP01748	SFHVFPPWMCKSLKKC
AP01749	LVQRGRFGRFLKKVRRFIPKVIIAAQIGSRFG
AP01752	VTCELLMFGGVVGDSACAANCLSMGKAGGSCNGGLCDCRKTTFKELWDKRFG
AP01753	GIWSSIKNLASKAWNSDIGQSLRNKAAGAINKFVADKIGVTPSQAAS
AP01754	GGYYCPFFQDKCHRHCRSFGRKAGYCGGFLKKTCICV
AP01756	PDPGQPWQVKAGRPPCYSIPCRKHDECRVGSCSRCNNGLWGDRTCR
AP01757	SPRVSRRYGRPFGGRPFVGGQFGGRPGCVCIRSPCPCANYG
AP01758	IPAMEPAARVKRSPGYGGCSPRWACGGYG
AP01759	RMRRSKSGKGSGGSKGSGSKGSKGSKGSGSKGSGSKGGSRPGGGSSIAGGGSKGKGGTQTA
AP01762	SPPNQPSIMTFDYAKTNK
AP01763	SPPSEQLGKSFNF
AP01765	APPPGYAMESDSFS
AP01766	FPPPGESAVDMSFFYALSNP
AP01768	QLGELIQQGGQKIVEKIQKIGQRIRDFFSNLRPRQEA
AP01769	KSLRPRCWIKIKFRCKSLKF
AP01771	ILPLLLGKVVCAITKKC
AP01778	DSMGAVKLAKLLIDKMKCEVTKAC
AP01783	FLPGVLRLVTKVGPAVVCAITRNC
AP01786	GKLQAFLAKMKEIAAQTL
AP01789	HSHACTSYWCGKFCGTASCTHYLCRVLHPGKMCACVHCSR
AP01790	HPHVCTSYYCSKFCGTAGCTRYGCRNLHRGKLCFCLHCSR
AP01791	FLWGLIPGAISAVTSLIKK
AP01793	GWINEEKIQKKIDEKIGNNILGGMAKAVVHKLAKGEFQCVANIDTMGNCETHCQKTSGEKGFCHGTK CKCGKPLSY
AP01794	FVDLKKIANIINSIFGK
AP01795	QIINNPITCMTNGAICWGPCPTAFRQIGNCGHFKVRCCKIR
AP01796	ASFPWSCPSLSGVCRKVCLPTELFFGPLGCGKGFLCGVSHFL
AP01797	GLWNSIKIAGKKLFVNVLDKIRCKVAGGCKTSPDVE
AP01798	SPRPDDKKNQGSASVDVQNERGEGTKVDARVRQELWRSDDGRTRAQAYGHWDRTYGGRNHGERSY GGGMRIEHTWGN

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01799	KRGFGKKLRKRLKKFRNSIKKRLKNFNVVIPIPLPG
AP01800	KRGLWESLKRKATKLGDDIRNTLRNFKIKFPVPRQG
AP01801	RTKRRIKLIKNGVKKVKDILKNNNIIILPGSNEK
AP01802	RPWAGNGSVHRYTVLSPRLKTQ
AP01804	GIRCPKSWKCKAFKQRVLKRLLAMLRQHAF
AP01815	DFGCARGMIFVCMRRCARMYPGSTGYCQGFRCMCDTMIPIRRPPFIMG
AP01824	FLPKLFAKITKKNMAHIR
AP01842	FIFPKKNIINSLFGR
AP01849	TSRCIFYRRKKCS
AP01850	SFLSTFKELAINAAKNAGQSILHTLSCKLDKTC
AP01855	GLFSKFVGKGIKNFLIKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01860	ATNIPFKVHFRCKAAFC
AP01876	GIFGKILGVGKKTLCELSGMC
AP01883	GILGNIVGMGKKVVCGLSGLC
AP01886	VVKCSYRLGSPDSQCN
AP01891	RFIYMKGFGKPRFGKR
AP01892	IPWKLPATFRPVERPFSKPFCRKD
AP01893	AAPRGGKGFFCKLFKDC
AP01895	FLGSLLGLVGKVVPTLFCKISKKC
AP01896	GLMSTLKDFGKTAAKEIAQSLLSTASCKLAKTC
AP01898	RRSRRGRGGGRRGGSGGRGGRGGGGRSGAGSSIAGVGSRGGGGGRHYA
AP01899	FLKPLFNAALKLLP
AP01900	FLPVLAGVLSRA
AP01901	GLASFLGKALKAGLKIGSHLLGGAPQQ
AP01906	GIGSLLAKAAKLGANLL
AP01911	SALVGCWTKSYPPNPCFGRG
AP01914	AAFRGCWTKNYSPKPCL
AP01916	GTRCGETCFVLPCWSAKFGCYCQKGFCYRN
AP01917	GLWDSIKNFGKTIALNVMDKIKCKIGGGCPP
AP01919	FTLKKSQLLLFFLGTINFSLCEEERNAEEERRDYPEEKDVEVEKR
AP01921	ILPIIGKILSTIF
AP01922	GMWSKILGHLIR
AP01923	GKWMSLLKHILK
AP01925	GLLDAIKDTAQNLFANVLDKIKCKFTKC
AP01927	GLFNVFKKVGKNVLKNVAGSLMDNLKCKVSGEC
AP01929	GIFALIKTAAKFVGKNLLKQAGKAGLEHLACKANNQC
AP01939	CVISAGWNHKIRCKLTGNC
AP01940	FKTWKRPPFQTSCWGIIKE
AP01941	CVHWQTNTARTSCIGP
AP01943	SLWETIKNAGKGFIQNLDKIR
AP01947	FLGPIIKIATGILPTAICKFLKKC
AP01948	SIRDKIKTIAIDLAKSAGTGVLKTLICKLDKSC
AP01952	FFPLLFGALSSHLPKLF
AP01953	FALGAVTKLLPSLLCMITRKC
AP01955	EYHLMNGANGYLTRVNGKTVYRVTKDPVSAVFGVISNCWGSAGAGFGPQH
AP01956	GFGMALKLLKKVL
AP01957	GTGLPMSERRKIMLMMR
AP01958	GLPRKILCAIAKKKGKCKGPLKLVCKC
AP01959	AILTTLANWARKFL
AP01963	ACQCPDAISGWTHTDYQCHGLENKMYRHVYAICMNGTQVYCRTEWGSSC

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP01964	IKLSPETKDNLKKVLKGAIKGAIAVAKMV
AP01965	LKIPGFVKDTLKKVAKGIFSAVAGAMTPS
AP01968	IKIPPIVKDTLKKVAKGVLSTIAGALST
AP01969	GPVGLLSSPGSLPPVGGAP
AP01971	VTSKSLCTPGCITGVLMCLTQNSCVSCNSCIRC
AP01972	STIVCVSLRICNWSLRFCPSFKVRCPM
AP01973	MLCKLSMFGAVLGVPACAIDCLPMGKTGGSCEGGVCGCRKLTFKILWDKKFG
AP01974	YGQSTHAVIYAQGYTYSSDWR
AP01975	KQIMTQFFNFARSPAVKD
AP01977	FLFSLIPSAISGLISAFK
AP01978	FIGAIARLLSKIF
AP01979	VAKCTEESGGKYFVFCCYKPTRICYMNEQKCESTCIGK
AP01981	GGKCTVDWGGQGGGRRLPSPLFCCYKPTRICYLNQETCETETCP
AP01982	ANKCIIDCMKVKTTCGDECKGAGFKTGGCALPPDIMKCCHNC
AP01993	TNWKKIGKCYAGTLGSAVLGFGAMGPVGYWAGAGVGYASFC
AP01995	ECELAKVDGGYTPKNCAMAVGGGMLSGAIRGGMSGTVFGVGTGNLAGAFAGAHIGLVAGGLACIGG YLGSH
AP01997	TPGGIDFISGGPHVAQDVLNAIKNFFK
AP02001	GMATKAGTALGKVAKAVIGAAL
AP02003	GFWTTAAEGLKKFAKAGLASILNPK
AP02006	GLLDALSGILGL
AP02007	GLLGTLGNLLNGLGL
AP02011	GLFDVIKKVASVIGLASP
AP02012	VKVGINGFGRIGRLVTRAAFHGKKVEVVAIND
AP02013	FIGKLISAASGLLSHL
AP02016	GKIPVKAIKQAGKVIGKGLRAINIAGTTHDVVSFFRPKKKKH
AP02019	ILGAIIPLVSGLLSHL
AP02020	FLSTLLKVAFKVVPTLFCPITKKC
AP02021	FFPIVGKRLYGLL
AP02022	FLPLFLPKIICVITKKC
AP02025	DCYEDWSRCTPGTSFLTGILWKDCHSRCKELGHRGGRCVDSPSKHCPGVLKNNKQCHCY
AP02026	GFWGKLWEGVKSAI
AP02028	KRKCPKTPFDNTPGAWFAHLILGC
AP02029	DSIRDVSPTFNKIRRWFDGLFK
AP02030	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLV LRLR
AP02032	RWKPFKKELKVGRNIRDGIIKAGPAVAVIGQATSIARPTGK
AP02033	VSCGDVTSSIAPCLSYVMGRESSPSSSCCSGVRTLNGKASSSADRRTACSCLKNMASSFRNLNMGNAASI PSKCGVSVAFPISTSVDCSKIN
AP02038	FLGLIFHGLVHAGKLIHGLIHRNRG
AP02040	KSCCRSTLGRNCYNLCRARGAQKLCAGVCRCKISSGLSCPKGFPK
AP02042	DRCSQQCQHHRDPDRKQQCMRECRRHQGRSD
AP02043	AVPLAVPLVYMGASVSPAVWNWLLVTFGAAAVAAAAVTVSDNDSHSCANNRGWCRSRCFSHEYIDS WHSDVCGSYDCCRPRY
AP02049	FFHHIFRGIVHVGKTIHKLVTGT
AP02051	AISCGQVSSAIGPCLSYARGQGSAPSAGCC
AP02052	TPALAVVTTVLPAAAVTTAKSV
AP02053	GLSQGVEPDIGQTYFEESRINQD
AP02064	GILDKLKEFGISAARGVAQSLLNTTASCKLAKTC
AP02068	YCERMMKRRSLTSPCKDVNTFIHGNKSNIKAICGANGSPYRENLRMSKSPFQVTTCKHTGGSPRPPCQ YRASAGFRHVVIACENGLPVHFDESFFSL
AP02075	SNFDCCLGYTDRILHPKFIVGFTRQLANEGCDINAIIFHTKKKLSVCANPKQTWVKYIVRLLSKKVKNM

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP02076	ASVATELRCQCLQTLQGIHPKNIQSVNVKSPGPHCAQTEVIATLKNGRKACLNPASPIVKKIIEKMLNS DKSN
AP02080	VPLSRTVRCTCISISNQPVNPRSLEKLEIIPASQFCPRVEIIATMKKKGEKRCLNPESKAIKNLLKAVSKE RSKRSP
AP02081	FPMFKRGRCLCIGPGVKAVKVADIEKASIMYPSNNCDKIEVIITLKENKGQRCLNPKSKQARLIIKKVE RKNF
AP02082	KPVSLSYRCPCRFFESHVARANVKHLKILNTPNCALQIVARLKNNNRQVCIDPKLKWIQEYLEKALNK
AP02083	VLEVYYTSLRCRCVQESSVFIPRRFIDRIQILPRGNGCPRKEIIVWKKNKSIVCVDPQAEWIQRMMEVL RKRSSSTLPVPVFKRKIP
AP02084	VGSEVSDKRTCVSLTTQRLPVSRIKTYTITEGSLRAVIFITKRGLKVCADPQATWVRDVVRSMDRKSN TRNNMIQTKPTGTQQSTNTAVTLTG
AP02085	KSMQVPFSRCCFSFAEQEIPLRAILCYRNTSSICSNEGLIFKLKRGKEACALDTVGWVQRHRKMLRHCP SKRK
AP02086	PDSVSIPITCCFNVINRKIPIQRLESYTRITNIQCPKEAVIFKTKRGKEVCADPKERWVRDSMKHLDQIF QNLKP
AP02087	GPASVPTTCCFNLANRKIPLQRLESYRRITSGKCPQKAVIFKTKLAKDICADPKKKWVQDSMKYLDQK SPTPKP
AP02088	QPDALNVPSTCCFTFSSKKISLQRLKSYVITTSRCPQKAVIFRTKLGKEICADPKEKWVQNYMKHLGR KAHTLKT
AP02089	ARGTNVGRECCLEYFKGAIPLRKLKTWYQTSEDCSRDAIVFVTVQGRAICSDPNNKRVKNAVKYLQS LERS
AP02090	AQVGTNKELCCLVYTSWQIPQKFIVDYSETSPQCPKPGVILLTKRGRQICADPNKKWVQKYISDLKLN A
AP02091	GTNDAEDCCLSVTQKPIPGYIVRNFHYLLIKDGCRVPAVVFTTLRGRQLCAPPDQPWVERIIQRLQRT SAKMKRRSS
AP02094	GPYGANMEDSVCCRDYVRYRLPLRVVKHFYWTSDSCPRPGVVLLTFRDKEICADPRVPWVKMILNK LSQ
AP02096	KVHGSLARAGKVRGQTPKVAKQEKKKKKTGRAKRRMQYNRRFVNVVPTFGKKKGPNANS
AP02098	SGTSEKERESGRLLGVVKRLIVCFRSPFP
AP02104	MQFITDLIKKAVDFFKGLFGNK
AP02105	MAADIISTIGDLVKLIINTVKKFQK
AP02116	FFPTIAGLTKLFCAITKKC
AP02117	VLSIVACSSGCGSGKTAASCVETCGNRCFTNVGSLC
AP02118	FLPAALAGIGGILGKLF
AP02119	GFGCPGDAYQCSEHCRALGGGRTGGYCAGPWYLGHPTCTCSF
AP02120	YVPKIPKPQPNKPNFPSFPGHGPFNPHASRFPRSPKDNGKIVFDAKKEGGKTQWNVETQQKVWGNK HGSIHVSAGAGKQPGGKPQGQVGIGGSFSWGK
AP02121	GFGCPFNENECHAHCLSIGRKFGFCAGPLRATCTCGKQ
AP02122	CLRIGMRGRELMGGIGKTM
AP02123	NVTPATKPTPSKPGYCRVMDELILCPDPPLSKDLCKNDSDCPGAQKCCYRTCIMQCLPPIFRE
AP02127	IWSAIWSGIKGLL
AP02128	SLQPGAPNFPIPGQEKQEGWKFDPSLTRGEDGNTLGSINIHHTGPNHEVGANWDKVIRGPGKAKPTY SIHGSWRW
AP02135	FIHHIIGGLFSAGKAIHRLIRRRRR
AP02136	FIHHIIGWISHGVRAIHRAIHG
AP02137	FLHHIVGLIHHGLSLFGDRAD
AP02140	IWDAIFHGAKHFLHRLVNPGGKDAVKDVQQKQ
AP02141	LLRHVVKILEKYL
AP02142	GFLDIIKDTGKEFAVKILNNLKCKLAGGCPP
AP02146	QGWEAVAAAVASKIVGLWRNEKTELLGHECKFTVKPYLKRFQVYYKGRMWCPGWTAIRGEASTRS QSGVAGKTAKDFVRKAFQKGLISQQEANQWLSS
AP02148	FFLLFLQGAAGNSVLCRIRGGRCHVGSCHFPERHIGRCSGFQACCIRTWG
AP02149	QYEALTAAILTKLSKMWHSDTLNFLGHTCHVSRTPTVKRFKLYWKGKFWCPGWAPFSGTSRTKSRS GSAREATKSFVDQALQRRLITQQEADLWLKG
AP02150	YEALVTSILGKLTGLWHNDSVDFMGHICYFRRRPKIRRFKLYHEGKFWCPGWAPFEGRCKYCVVF

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP02153	GGWLDIVKAIVVPAARETIKTQEITLLDHYCTLSRSPYIKSLELHYRAEVTCPGWTIIRGRGSNHRNPT NSGKDALKDFMTQAVAAGLVTKEEAAPWLN
AP02154	YVDREINLFDHYCIISRSPHISRWELKWQATVTCPGWTPVKGKVRGYSNPLSAEREATRDFVQRIVQR GLVTRDEASEWL
AP02159	FFGHLYRGITSVVKHVHGLLSG
AP02160	TDTNVIGECFDEWSRCHRQTRWWTKILFQSCENRCKCKVQLMGNCIKVPFKCFLWKQKRFMCECY GPISGTKPWYCGWEL
AP02162	KIKIPWGKVKDFLVGGMKAV
AP02163	GFFGNTWKKIKGKADKIMLKKAVKIMVKKEGISKEEAQAKVDAMSKKQIRLYLLKYYGKKALQKASE KL
AP02165	ITSFIGCTPGCGKTGSFNSFCC
AP02169	AKISGPEETSELPEVVSEERVPATATEPMADLRHGVTREPISPASKDSLRDKFKEKLDKWFHRPNLLS KRD
AP02171	PGGPGSAPPATCRYWCRTPQGQAYCCEGVDEPEGPVGVKIGSCPRVRPQCPPVRTFGPPSPCSNDFK CFGSDKCCYDICLEQHVCKPLSFFG
AP02172	FFGSLLSLGSKLLPSVFKLFQRKKE
AP02173	QLPICGETCVLGGCYTPNCRCQYPICVR
AP02174	FLPFLIPALTSLISSL
AP02175	RRSKARGGSRGSKMGRKDSKGGSRGRPGSGSRPGGGSSIAGASRGDRGGTRNA
AP02178	LRVRRTLQCSCRRVCRNTCSCIRLSRSTYAS
AP02179	LNWGAILKHIIK
AP02180	LDVKKIICVACKIKPNPACKKICPK
AP02182	DRCTKRYGRCKRDCLESEKQIDICSLPRKICCTEKLYEEDDMF
AP02184	APAHRSSTFPKWVTKTERGRQPLRS
AP02185	GPVSAVLTELRCTCLRVTLRVNPKTIGKLQVFPAGPQCSKVEVVASLKNGKQVCLDPEAPFLKKVIQK ILDSGNKKN
AP02193	YSKSLPLSVLNP
AP02196	KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY
AP02197	PAAAAQAVAGLAPVAAEQ
AP02198	KAYSMPRCKGGFRAVMCWL
AP02199	KAYSTPRCKGLFRALMCWL
AP02202	RKCNFLCKLKEKLRTVITSHIDKVLRPQG
AP02203	WNDTGKDADGSEY
AP02205	MVFAYAPTCARCKSIGARYCGYGYLNRKGVSCDGQTTINSCEDCKRKFGRCSDGFITECFL
AP02208	SPIEPKGEILHRFRRSFCDYNLCVVSCKDSGFIGGYCSELDLCSCTIGWQ
AP02211	FLNALKNFAKTAGKRLKSLLN
AP02213	GILGKLWEGFKSIV
AP02214	IFGAIWKGISSLL
AP02215	FLSTIWNGIKSLL
AP02216	FLGALWNVAKSVF
AP02217	FLSTLWNAAKSIF
AP02222	FIPLVSGLFSRLL
AP02223	VIPIVSGLLSSLL
AP02224	GLLLDTVKGAAKNVAGILLNKLKCKMTGDC
AP02225	YDTGIQGWTCGSRGLCRKHCYAQEHTVGYHGCPRRYRCCALRF
AP02226	FCHLCEDLIKDGKEAGDVALDVWLDEEIGSRCKDFGVLASECFKELKVAEHDIWEAIDQEIPEDKTCK EAKLC
AP02227	NGIECEMCKMSVKIVVPMLGEDTESIKKAVDAECKKEFHSIPFGTQECKKFIDTKLDPIIHELENGTAP SDVCTKLGMC
AP02229	GKCSVLKKVACAAAIAGAVAACGGIDLPCVLAALKAAEGCASCFCEDHCHGVCKDLHLC
AP02230	PKRKAEGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPKPKKAPAKKGEKVPKGKKGKADAGKEG NNPAENGDAKTDQAQKAEGAGDAK
AP02231	RAIGGGLSSVGGGSSTIKY

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP02232	AATAKKGAKKADAPAKPKKATKPKSPKKAAKKAGAKKGVKRAGKKGAKKTTKAKK
AP02233	GFGCPNDYSCSNHCRDSIGCRGGYCKYQLICTCYGCKKRRSIQE
AP02237	FFRLLFHGVHHGGGYLNAA
AP02238	GWKKWFTKGERLSQRHFA
AP02239	GFLGILFHGVHHGRKKALHMNSERRS
AP02241	KYALMKKIAELIPNLKSRQVK
AP02242	TWATIGKTIVQSVKKCRTFTCGCSLGSCSNCN
AP02244	FQSHSLPTPADERNLLQQIDCGTSCSARCRLSSRPRLCKRACGTCCARCNCVPSGTAGNLDECPCYAN MTTHGNKRKCP
AP02247	MAGFLKVVQLLAKYGSKAVQWAWANKGKILDWLNAGQAIDWVVSKIKQILGIK
AP02248	FKKKKRNIGTFVFFAIALFCTVMFAYLLLTNQYVPIDYNVPRYA
AP02249	GLLSLLSLLGKLL
AP02250	MKTILRFVAGYDIASHKKKTGGYPWERGKA
AP02251	MWGRILAFVAKYGTKAVQWAWKNKWFLLSLGEAVFDYIRSIWGG
AP02253	MGAIAKLVAKFGWPFIKKFYKQIMQFIGQGWTIDQIEKWLKRH
AP02258	CARLNCVPKGTSGNTETCPCYASLHSCRKYG
AP02259	QYEALVASILGKLSGLWHSDTVDFMGHTCHIRRRPKFRKFKLYHEGKFWCPGWTHLEGNSRTKSRSG SARDAIKDFVYKALQNKLITENNAAAWLKG
AP02260	RILTMTKRVKMPQLYKQIVCRLFKTC
AP02262	FLGGLLASLLGKI
AP02263	YPELQQDLIARLL
AP02264	FLSGILKLAFKIPSVLCAVLKNC
AP02265	AKAWGIPPHVIPQIVPVRIRPLCGNV
AP02266	GFWDSVKEGLKNAAVTILNKIKCKISECPPA
AP02267	FIPGLRRLFATVVPTVVCAINKLPPG
AP02268	GLLDSVKEGLKKVAGQLLDTLKCKISGCTPA
AP02270	GFFDRIKALTKNVTLELLNTITCKLPVTPP
AP02273	FITGLIGGLMKAL
AP02281	GVLDTLKNVAIGVAKGAGTGVLKALLCQLDKSC
AP02283	SLFGTFAKMALKGASKLIPHLLPSRQQ
AP02285	KLGFENFLVKALKTVMHVPTSPLL
AP02287	GLKEVAHSAKKFAKGFISGLTGS
AP02288	GWASSIGSILGKFAKGGAQAFLQPK
AP02289	GWLPTFGKILRKAMQLGPKLIQPI
AP02292	GLLSNVAGLLKQFAKGGVNAVLNPK
AP02293	GFMSKVANFAKKFAKGGVNAIMNQK
AP02294	FIGALLRPALKLLAGK
AP02296	ILPIRSLIKKLL
AP02297	FLPLKKLRFGLL
AP02301	RISKKKGKGSWIKNGLIKGIKGLGKEISLDVIRTGIDIAGCKIKGEC
AP02303	SLWENFKNAGKQFILNILDKIRCRVAGGCRT
AP02304	FLAGLIGGLAKML
AP02306	PPCRGIFCRRVGSSSAIARPGKTLSTFITV
AP02307	GKCNVLCQLKQKLRSIGSGSHIGSVVLPRG
AP02308	GNGVVLTLTHECNLATWTKKLKCC
AP02311	ITIPPIVKDTLKKFFKGGIAGVMGKSQ
AP02317	ITIPPIVKNTLKKFIKGAVSALMS
AP02318	IKIPSFFRNILKKVGKEAVSLIAGALKQS
AP02319	GIFPIFAKLLGKVIKVASSLISKGRTE
AP02321	TNYGNGVGVPDAIMAGIIKLIFIFNIRQGYNFGKKAT

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP02323	KRKKHRCRVYNNGMPTGMYRWC
AP02324	VFHAYSARGNYYGNCPANWPSCRNNYKSAGGK
AP02332	CETPSKHFNGLCIRSSNCASVCHGEHFTDGRCQGVRRRCMCLKPC
AP02341	LVKDNPLDISPKQVQALCTDLVIRCMCCC
AP02342	DICTCCAGTKGCNTTSANGAFICEGQSDPKKPKACPLNCDPHIAYA
AP02343	IQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLLKNGERIEKVEHSDLSFSKDWSFYLLYYTEF TPTEKDEYACRVNHVTLSQPKIVKWDRDM
AP02346	GLEESPGHPGQPGPPGPPGAPGP
AP02347	KVTKSVKSIPVKI
AP02348	TTPLCVGVIIGLTTSIKICK
AP02351	QKIAEKFSGTRRG
AP02352	YPGPQAKEDSEGPSQGPASREK
AP02353	LPVNSPMNKGDTEVMKCIVEVISDTLSKPSPMPVSKECFETLRGDERILSILRHQNLLKELQDLALQGA KERTHQQ
AP02354	KINNPVSCLRKGGRCWNRCIGNTRQIGSCGVPFLKCCKRK
AP02358	FRFGSFLKKVWKSKLAKKLRSKGKQLLKDYANKVLNGPEEEAAAPAE
AP02360	MVALLKSLERRRLMITISTMLQFGLFLIALIGLVIKLIELSNKK
AP02365	VIVKAIATLSKKLL
AP02367	INLKAIAALARNY
AP02370	FCKSLPLPLSVK
AP02371	GHLGRPYIGGGGGFNRGGGFHRGGGFHRGGGFHSGGGFHRGGGFHSGGSFGYR
AP02373	RRRRRRPPCEDVNGQCQPRGNPCLRLRGACPRGSRCCMPTVAAH
AP02374	GLVGTLLGHIGKAILG
AP02376	GLRRLFADQLVGRRNI
AP02384	AWLDKLKSLGKVVGKVALGVAQNYLNPQQ
AP02385	TKPTLLGLPLGAGPAAGPGKR
AP02386	KLSPSLGPVSKGKLLAGQR
AP02387	RLGTALPALLKTLLAGLNG
AP02389	GKGLEVIKWKLKHVIQL
AP02390	LFAKINGLKVGPLKIQIV
AP02391	FSLFFPYAALKWLRKLLKK
AP02392	VKLEILGSKGGAKI
AP02393	VSKIKKYLKYKDRI
AP02394	DWTCWSCLVCAACSVELLNLVTAATGASTAS
AP02395	DWTFANWSCLVCDDCSVNLTV
AP02396	LASTLGISTAAAKKAIDIIDAASTIASIISLIGIVTGAGAISYAIVATAKTMIKKYGKKYAAAW
AP02397	CWSCMGHSCWSCMGHSCWSCAGHSCWSCMGHSCWSCMGHSCWSCAGHCCGSCWHGGM
AP02399	ESISVAGGTWNYGYGVGQAYSHYKHDYNNHGAKVVNSNNGVKDYKNAGPGVWAKASIGTVWDPAT FYYNPTGFYSN
AP02400	FAVWGCADYRGYCRAACFAFEYSLGPKGCTEGYVCCVPNTF
AP02401	VAPIAKYLATALAKWALKQGFAKLKS
AP02402	IGGALGNALNGLGTWANMMNGGGFVNQWQVYANKGKINQYRPY
AP02405	GGYKNFYGSALRKGFYEGEAGRAIRR
AP02406	TVKCGMNGKMPCKHGAFYTDTCDKNVFYRCVWGRPVKKACGRGLVWNPRGFCDYA
AP02407	SDYLNNNPLFPRYDIGNVELSTAYRSFANQKAPGRLNQNWALTADYTYR
AP02409	AIFIFIRWLLKLGHHGRAPP
AP02414	FLKGVINLASKIPSMLCAVLKTC
AP02415	GLFDSITQGLKDTAVKLLDKIKCKLSACPPA
AP02416	FIVPSIFLLKKAFCIALKKNC
AP02417	MKFFTLLAALMALFAICNNFSMVSASRDSRPVQPRVQPPPPPPKQKPSIYDTPIRRPGGQKTMYA
AP02420	GVLSAFKNALPGIMKIIV

Table A.11: APD Gram-Both AMPs Continued...

Definition	Sequence
AP02429	KTKQQFLIKAQTQLFKVFGYTL
AP02430	RLFRHAFKAVLRL
AP02432	APVPFSCTRGCLTHLV
AP02433	GDINGEFTTSPACVYSVMVVSKASSAKCAAGASAVSGAILSAIRC
AP02435	FWGAVWKILSKVLPHIPGTVKWLQEKV
AP02436	GTDSGRFCSSICGQRCSKAGMKDRCMKFCGICCGKCKCVPSGTYGNKHECPCYRDMKNSKGKPKCP
AP02437	FFGRLKSVWSAVKHGWKAAKSR
AP02439	SCTTCVCTCSCCTT
AP02440	QNCPTRRGLCVTSGLTACRNHCRSCHRGDVGCVRCSNAQCTGFLGTTCTCINPCPRC
AP02442	QVLEGLAAAVTGKLAGLWRNGEVELLGHYCSYSVTPTIRRWQLYFRGRMWCPGWTSIRGEAMTRSN SGVQGDTTRDFVTKALNAGLISQQEAQAWLDG
AP02443	NIFDDIFGKVTETLVDFGTTDIAGNPCNYRLSPRLIKFELYFVGLVWCPGWTTIQGESLTRSRTRVVN KAVEDFAKKAVAAGIMTQEDADPLLNA
AP02444	GWLDRIIGTAVDSVAEFGTTNIVDQICNTRVMPTIKKFELYFRGRVWCPGWTTIQGESLTRSRTRVVN KAVEDFARKAVAAGLMTQEDANPLLNA
AP02446	DTFDYKKFGYRYDSLELEGRSISRIDELIQQRQEKDRTFAGFLLKGFGTSAS
AP02451	VVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGVIFTTKKGQQSCGDPKQEWVQRYMKNLDAKQ KKASPRARAVA
AP02452	TRGSDISKTCCFQYSHKPLPWTWVRSYEFTSNSCSQRAVIFTTKRGKKVCTHPRKKWVQKYISLLKTP KQL
AP02453	MDSFQKIEKIGEGTYGVVYKAKDKVSGRLVALKKIRLENESEGVPSTA
AP02456	KKCKFFCKVKKKIKSIGFQIPIVSIPFK
AP02457	KKCGFFCKLKNKLKSTGSRSNIAAGTHGGTFRV
AP02458	FLPVLGKVIKLVGGLL
AP02461	FLPGLIKAAVGVGSTILCKITKKC
AP02466	SILSTLKDVGISAIKNAGSGVLKTLLCKLNKNCEK
AP02470	GFMDTAKNVAKNVAKNVAVTLLDKLRCKVTGGC
AP02476	SIMSTLKQFGISAIKGAAQNVLGVLSCKIAKTC
AP02478	QSHISLCRWCCNCCKANKGCGFCCKF
AP02479	GMKCKFCCNCCNLNGCGVCCRF
AP02480	MTPLWRIMNSKPFGAYCQNNYECSTGLCRAGHCSTSHRATSETVNY

Table A.12: DBAASP Data Set AMPs and Median MIC Values

Definition	*E.coli* Median MIC	*S.aureus* Median MIC	Sequence
Seq265	3.00	72.00	KWKSFLKTFKSAKKTVLHTALKAISS
Seq317	8.00	16.00	KIKGAIKWKGAIKIKGAI
Seq321	8.00	16.00	KKLAGLAKKWAGLAKKLAGLA
Seq322	32.00	256.00	KLKAGLAKWKAGLAKLKAGLA
Seq326	8.00	16.00	KLAGLAKKWAGLAKKLAGLAK
Seq360	31.25	50.00	GIGKFLHSAKKFGKAFVGEIMNS
Seq1231	8.00	8.00	RQRVEELSKFSKKGAAARRRK
Seq1484	8.21	4.11	GVLGAVKDLLIGAGKSAAQSVLKTLSCKLSNDC
Seq1503	9.37	4.68	ATAWDFGPHGLLPIRPIRIRPLCGKDKS
Seq1519	42.50	21.25	FLPPSPWKETFRTS
Seq1520	17.50	35.00	TSRCYIGYRRKVVCS
Seq1522	36.00	18.00	GLLSGTSVRGSI
Seq1837	100.00	25.00	GLFAVIKKVASVIKKL

Table A.12: DBAASP Data Set AMPs and Median MIC Values Continued...

Definition	***E. coli* Median MIC**	***S. aureus* Median MIC**	**Sequence**
Seq1838	100.00	25.00	GLFAVIKKVAAVIKKL
Seq1841	100.00	50.00	KLFAVIKKVAAVIGGL
Seq2251	15.00	5.00	FLPKMSTKLRVPYRRGTKDYH
Seq3226	170.00	170.00	TESYFVFSVGM
Seq3949	64.00	6.00	FLGVVFKLASKVFPAVFGKV
Seq3951	16.00	32.00	FLGKVFKLASKVFPAVFGKV
Seq3953	128.00	8.00	FLGVVFKLASKVFKAVFGKV
Seq3954	64.00	4.00	FLGVVFKLASKVFPAVFKKV
Seq3959	16.00	128.00	FLGKVFKLASKVFPAVFKKV
Seq3974	256.00	32.00	FLGVVFKLASKVFPAVVGKV
Seq3982	32.00	256.00	FLGKVFKKAVKVFPAVFGKV
Seq3987	128.00	16.00	FLKVVFKLASKVFGAVFGKV
Seq4106	300.00	300.00	CLRIGMRGRELMGGIGKTM
Seq4333	64.00	128.00	VKRFKKFFRKLKKSVKKL
Seq4337	128.00	128.00	LVKRFKKFFRKLKKSVKK
Seq4380	32.00	128.00	KLVKRFKKFFRKLKKS
Seq4536	7.04	14.07	RKCNFLCKLKEKLRTVITSHIDKVLRPQG
Seq4867	3.10	3.1	FFHHIFRGIVHVGKTIHRLVTG
Seq5137	10.00	5.00	VRRFAWWWAFLRR
Seq5295	64.00	64.00	VKLKYPKVKLYP
Seq5385	16.00	32.00	KRFKKFFKKLKNSVKKRVKKFFRKPRVIGVTFPF
Seq5387	16.00	64.00	KRFKKFFMKLKKSVKKRVMKFFKKPMVIGVTFPF
Seq5670	8.00	64.00	YSKSLPLSVLNP
Seq6187	64.00	64.00	VKLKYPVKLYP
Seq6190	16.00	4.00	VKLKYPLKVKLYP
Seq6191	64.00	4.00	VKLVYPLKVKLYP
Seq6193	4.00	8.00	VKLKVKYPKLKVKLYP
Seq6194	8.00	2.00	VKLKVYPKLKVKLYP
Seq6195	32.00	64.00	VKLKKYPKLKVKLYP
Seq6196	16.00	32.00	VKLVKYPKLKVKLYP
Seq7011	32.00	16.00	KFFRKLKKSVKKRAKKFFKKPRVIGVSIPF
Seq7012	16.00	16.00	KFFRKLAKSVKKAAKEFFKKPRVIGVSIPF
Seq7702	12.50	9.375	KFFKKLKKAVKKGFKKFAKV
Seq8154	16.00	4.00	FLFKLIPKAIKGLVKAIRK
Seq4477	0.025	0.006	RVRRFWPLVPVAINTVAAGINLYKAIRRK
Seq2025	0.008	0.042	RWCVYAYVRVRGVLVRYRRCW
Seq3836	0.002	0.008	GFLSALKKYLPIVLKHV
Seq894	0.003	0.002	AKKVFKRLEKLFSKIWNWK
Seq897	0.003	0.005	AKKVFKRLEKLFSKIFNFK
Seq898	0.003	0.006	AKKVFKRLEKLFSKIYNYK
Seq4969	0.004	0.005	FKRLKKLFKKIWNWK
Seq5243	0.004	0.003	RRIRPRPPRLPRPRPRPLPFPRPGPRPIPRPLPFP
Seq7151	0.006	0.012	FLPLIGRVLSGILGWKRKRFG
Seq3301	0.007	0.005	GKWWSLLKHILK
Seq7553	0.003	0.009	AKKVFKRLEKLFSKIWNDK
Seq4965	0.041	0.009	WKKVFKRLEKLFSKIWNWK
Seq3291	0.006	0.011	GKWMKLLKHILK
Seq3262	0.003	0.012	GMWSKILGHLIR

Table A.12: DBAASP Data Set AMPs and Median MIC Values Continued...

Definition	***E. coli* Median MIC**	***S. aureus* Median MIC**	**Sequence**
Seq7551	0.036	0.013	AKKVFKRLEKLFSKIQNDK
Seq5546	10.00	6.00	IIKVPLKKFKSMREVMRDHGIKAPVVDPATKY
Seq354	2.00	1.00	RLLRRLLRRLLRRLLRRLLR
Seq1141	100.00	100.00	GLLSVLGSVAQHVLPHVVPVIAEHL
Seq3971	256.00	8.00	FLGVVFKLVSKVFPAVFGKV
Seq8153	16.00	4.00	FLFKLIPKVIKGLVKAIRK
Seq665	200.00	25.00	FLGLLFKVASK
Seq645	100.00	25.00	FWGALFKVASK
Seq2015	0.13	0.50	RRWCFRVCYRGFCYRKCR
Seq2012	0.50	4.00	KWKSFLKTFKSAVKTVLHTALKAISS
Seq643	12.50	1.60	WLGALFKVASKVL
Seq3950	32.00	4.00	FLKVVFKLASKVFPAVFGKV
Seq5143	20.00	5.00	VRRFPWWWAFLRR
Seq5140	20.00	5.00	VRRYPWWWPYLRR
Seq937	81.70	1.40	GLPALILWIKRKRQQ
Seq646	100.00	12.50	FLWALFKVASK
Seq3991	256.00	8.00	FLGVVFKLASKVFGAVFKKV
Seq663	100.00	12.50	FLGALFKVAWK
Seq3995	64.00	16.00	FLFRVASKVFPALIGKFKKK
Seq5549	50.00	100.00	AVVKVPLKKFKSIRETMKEKGLLEDF
Seq1506	3.50	3.50	VIPFVASVAAEMMQHVYCAASKKC
Seq2013	0.50	32.00	KWKSFLRTLKSPAKTVFHTALKAISS
Seq668	100.00	25.00	FLGALFKLASK
Seq3930	4.70	2.30	FALGAVTKRLPSLFCLITRKC
Seq1507	6.25	3.12	GLLSGILGAGKHIVCGLSGPCQSLNRKSSDVEYHLAKC
Seq670	100.00	50.00	FLGALFKFASK
Seq5141	6.00	6.00	VRRFPFFFPFLRR
Seq5383	16.00	32.00	KRFKKFFKKVKKSVKKRLKKIFKKPMVIGVTIPF
Seq5142	20.00	4.00	VRRFAWWWPFLRR
Seq1494	4.68	9.37	FMPILSCSRFKRC
Seq5569	3.10	3.10	FIHHIFRGIVHAGRSIGRFLTG
Seq1689	128.00	32.00	VKLKVYPLKVKLYP
Seq6835	200.00	100.00	TLKQFAKGVGKWLVK
Seq1323	100.00	25.00	GLFAVIKKVASVIGGL
Seq653	200.00	100.00	FLGALWKVASK
Seq2044	0.033	0.002	GLWSTIKQKGKEAAIAAAKAAGQAALGAL
Seq3961	32.00	256.00	FLGVVFKKASKVFKAVFGKV
Seq926	25.00	3.10	FLGWLFKVASK
Seq6188	64.00	16.00	VKLYPKVKLYP
Seq3955	4.00	64.00	FLKKVFKLASKVFPAVFGKV
Seq1499	35.60	17.80	GLFTLIKCAYQLIAPTVACN
Seq2252	5.00	1.00	GILKKFMLHRGTKVYKMRTLSKRSH
Seq5620	0.004	0.012	SPIHACRYQRGVCIPGPCRWPYYRVGSCGSGLKSCCVRNRWA
Seq626	12.50	3.10	FLGALFKVASKVLPSVKCAITKKC
Seq934	76.70	1.30	GLPLLISWIKRKRQQ
Seq667	100.00	12.50	FLGFLFKVASK
Seq3997	256.00	64.00	LGALFRVASKVFPAVISMVK
Seq933	6.25	3.10	GLSALISWIKRKRQQ

Table A.12: DBAASP Data Set AMPs and Median MIC Values Continued...

Definition	***E. coli* Median MIC**	***S. aureus* Median MIC**	**Sequence**
Seq4794	200.00	20.00	ILPWKWPWAPARR
Seq640	6.30	3.10	FLKALFKVASKVL
Seq918	12.50	3.10	FLGALFKVASKVLPSVFCAITKKC
Seq4725	0.040	0.004	RIKRFWPVVIRTVVAGYNLYRAIKKK
Seq888	16.00	16.00	RAGLQFPVGGIGKFLHSAKKFGK
Seq666	200.00	200.00	FLGKLFKVASK
Seq641	50.00	3.10	FLKALFKVALKVL
Seq4480	0.007	0.009	GIRCPKSWKCKAFKQRVLKRLLAMLRQHAF
Seq1483	3.21	5.83	AALKGCWTKSIPPKPCFGKR
Seq3957	16.00	4.00	FLGKVFKLASKVFKAVFGKV
Seq2427	11.40	2.60	GLPALISWIKRKRL
Seq5390	18.00	32.00	KRFKKFFKKLKNSVKKRAKKFFKKPRVIGVSIPF
Seq1466	18.10	16.00	FLFSLIPSAISGLISAFK
Seq671	12.50	3.20	FLGWLFKWASK
Seq5547	8.00	4.00	SGRGKQGGKVRAKAKTRSSRAGLQFPVGRVHRLLRKGNY

Table A.13: Server Annotation Data Set

Definition	**Sequence**
AP00001	GLWSKIKEVGKEAAKAAAKAAGKAALGAVSEAV
AP00002	YVPLPNVPQPGRRPFPTFPGQGPFNPKIKWPQGY
AP00003	DGVKLCDVPSGTWSGHCGSSSKCSQQCKDREHFAYGGACHYQFPSVKCFCKRQC
AP00005	VFIDILDKVENAIHNAAQVGIGFAKPFEKLINPK
AP00006	GNNRPVYIPQPRPPHPRI
AP00008	RLCRIVVIRVCR
AP00011	WNPFKELERAGQRVRDAVISAAPAVATVGQAAAIARG
AP00013	GLFDIIKKIAESF
AP00014	GLLDIVKKVVGAFGSL
AP00017	GLFDIVKKVVGTLAGL
AP00019	GLFDIAKKVIGVIGSL
AP00020	GLFDIVKKIAGHIAGSI
AP00023	AACARFIDDFCDTLTPNIYRPRDNGQRCYAVNGHRCDFTVFNTNNGGNPIRASTPNCKTVLRTAANR CPTGGRGKINPNAPFLFAIDPNDGDCSTNF
AP00025	HGVSGHGQHGVHG
AP00026	FKCRRWQWRMKKLGAPSITCVRRAF
AP00027	ITPATPFTPAIITEITAAVIA
AP00028	CLGIGSCNDFAGCGYAVVCFW
AP00030	QRFSQPTFKLPQGRLTLSRKF
AP00031	DKLIGSCVWGAVNYTSDCNGECKRRGYKGGHCGSFANVNCWCET
AP00035	KSSAYSLQMGATAIKQVKKLFKKWGW
AP00036	DFASCHTNGGICLPNRCPGHMIQIGICFRPRVKCCRSW
AP00038	QGVRNHVTCRINRGFCVPIRCPGRTRQIGTCFGPRIKCCRSW
AP00040	QVVRNPQSCRWNMGVCIPISCPGNMRQIGTCFGPRVPCCRRW
AP00045	QGVRSYLSCWGNRGICLLNRCPGRMRQIGTCLAPRVKCCR
AP00048	SGISGPLSCGRNGGVCIPIRCPVPMRQIGTCFGRPVKCCRSW
AP00050	GIGASILSAGKSALKGLAKGLAEHFAN
AP00058	GIGTKILGGVKTALKGALKELASTYAN
AP00064	ILGPVLGLVGNALGGLIKNE

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00066	IKITTMLAKLGKVLAHV
AP00069	INIKDILAKLVKVLGHV
AP00070	INVLGILGLLGKALSHL
AP00071	FLPAIFRMAAKVVPTIICSITKKC
AP00073	FLPLLAGLAANFLPKIFCKITRKC
AP00074	FLPVLAGIAAKVVPALFCKITKKC
AP00076	GILDTLKNLAISAAKGAAQGLVNKASCKLSGQC
AP00078	GILLDKLKNFAKTAGKGVLQSLLNTASCKLSGQC
AP00080	GIFSKLGRKKIKNLLISGLKNVGKEVGMDVVRTGIDIAGCKIKGEC
AP00083	GILSLVKGVAKLAGKGLAKEGGKFGLELIACKIAKQC
AP00085	SLFSLIKAGAKFLGKNLLKQGACYAACKASKQC
AP00086	GIMSIVKDVAKNAAKEAAKGALSTLSCKLAKTC
AP00088	GILDTLKQFAKGVGKDLVKGAAQGVLSTVSCKLAKTC
AP00089	FLGALFKVASKVLPSVFCAITKKC
AP00091	GLLNTFKDWAISIAKGAGKGVLTTLSCKLDKSC
AP00094	FLPLIGRVLSGIL
AP00095	LLPIVGNLLKSLL
AP00097	VLPIIGNLLNSLL
AP00099	FFPVIGRILNGIL
AP00101	FVQWFSKFLGRIL
AP00102	GSKKPVPIIYCNRRTGKCQRM
AP00103	RECKAQGRHGTCFRDANCVQVCEKQAGWSHGDCRAQFKCKCIFEC
AP00104	FLPFLAKILTGVL
AP00105	FLPLFASLIGKLL
AP00107	FLPFLASLLSKVL
AP00109	VLPLISMALGKLL
AP00110	NFLGTLINLAKKIM
AP00111	FLPILINLIHKGLL
AP00113	GLLSGLKKVGKHVAKNVAVSLMDSLKCKISGDC
AP00114	SMLSVLKNLGKVGLGFVACKINKQC
AP00116	GFLDIINKLGKTFAGHMLDKIKCTIGTCPPSP
AP00117	FLPFIARLAAKVFPSIICSVTKKC
AP00119	GILSSIKGVAKGVAKNVAAQLLDTLKCKITGC
AP00121	GLMDTVKNVAKNLAGHMLDKLKCKITGC
AP00123	GLFLDTLKGLAGKLLQGLKCIKAGCKP
AP00125	KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAILS
AP00126	GGLKKLGKKLEGVGKRVFKASEKALPVAVGIKALGK
AP00134	SWLSKTAKKLENSAKKRISEGIAIAIQGGPR
AP00137	LRDLVCYCRTRGCKRRERMNGTCRKGHLMYTLCCR
AP00140	SQLGDLGSGAGQGGGGGGSIRAAGGAFGKLEAAREEEFFYKKQKEQLERLKNDQIHQAEFHHQQIKE HEEAIQRHKDFLNNLHK
AP00142	GLKKLLGKLLKKLGKLLLK
AP00144	GIGKFLHSAKKFGKAFVGEIMNS
AP00145	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGAGSIGRRP
AP00146	GIGAVLKVLTTGLPALISWIKRKRQQ
AP00147	AKIPIKAIKTVGKAVGKGLRAINIASTANDVFNFLKPKKRKA
AP00148	ATYNGKCYKKDNICKYKAQSGKTAICKCYVKKCPRDGAKCEFDSYKGKCYC
AP00149	MPCSCKKYCDPWEVIDGSCGLFNSKYICCREK
AP00150	ILPWKWPWWPWRR
AP00151	RCVCTRGFCRCVCRRGVC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00152	VRRFPWWWPFLRR
AP00153	RSVCRQIKICRRRGGCYYKCTNRPY
AP00154	YSRCQLQGFNCVVRSYGLPTIPCCRGLTCRSYFPGSTYGRCQRY
AP00155	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG
AP00157	ALWKTMLKKLGTMALHAGKAALGAAADTISQGTQ
AP00159	ALWKNMLKGIGKLAGKAALGAVKKLVGAES
AP00160	ALWMTLLKKVLKAAAKAALNAVLVGANA
AP00164	ALWKTIIKGAGKMIGSLAKNLLGSQAQPES
AP00166	GWGSFFKKAAHVGKHVGKAALTHYL
AP00168	GRPNPVNNKPTPHPRL
AP00170	VDKGSYLPRPTPPRPIYNRN
AP00171	HRHQGPIFDTRPSPFNPNQPRPGPIY
AP00172	GKPRPYSPRPTSHPRPIRV
AP00174	RRCICTTRTCRFPYRRLGTCIFQNRVYTFCC
AP00175	DSHEERHHGRHGHHKYGRKFHEKHHSHRGYRSNYLYDN
AP00176	ACYCRIPACIAGERRYGTCIYQGRLWAFCC
AP00179	VCSCRLVFCRRTELRVGNCLIGGVSFTYCCTRV
AP00180	ATCYCRTGRCATRESLSGVCEISGRLYRLCCR
AP00181	AFTCHCRRSCYSTEYSYGTCTVMGINHRFCCL
AP00182	GFGCPLDQMQCHRHCQTITGRSGGYCSGPLKLTCTCYR
AP00184	RSGRGECRRQCLRRHEGQPWETQECMRRCRRRG
AP00186	GRCVCRKQLLCSYRERRIGDCKIRGVRFPFCCPR
AP00187	VVCACRRALCLPRERRAGFCRIRGRIHPLCCRR
AP00189	VSCTCRRFSCGFGERASGSCTVNGVRHTLCCRR
AP00190	HPLKQYWWRPSI
AP00191	QCRRLCYKQRCVTYCRGR
AP00193	DTHFPICIFCCGCCHRSKCGMCCKT
AP00194	VGECVRGRCPSGMCCSQFGYCGKGPKYCGR
AP00195	RGGRLCYCRRRFCVCVGR
AP00196	WYVKKCLNDVGICKKKCKPEEMHVKNGWAMCGKGRDCCVPAD
AP00197	QLKKCWNNYVQGHCRKICRVNEVPEALCENGRYCCLNIKELEAC
AP00198	MRILYLLFSVLFLVLQVSPGLSLPQRDMFLCRIGSCHFGRCPIHLVRVGSCFGFRSCCKSPWDV
AP00199	KYYGNGVHCTKSGCSVNWGEAFSAGVHRLANGGNGFW
AP00200	LKLKSIVSWAKKVL
AP00203	QCIGNGGRCNENVGPPYCCSGFCLRQPGQGYGYCKNR
AP00204	ITSISLCTPGCKTGALMGCNMKTATCNCSIHVSK
AP00206	WKSESLCTPGCVTGALQTCFLQTLTCNCKISK
AP00207	TAGPAIRASVKQCQKTLKATRLFTVSCKGKNGCK
AP00208	RADTQTYQPYNKDWIKEKIYVLLRRQAQQAGK
AP00209	GVLSNVIGYLKKLGTGALNAVLKQ
AP00211	RRWCFRVCYRGFCYRKCR
AP00216	ATCDLLSGTGINHSACAAHCLLRGNRGGYCNGKGVCVCRN
AP00217	GICACRRRFCPNSERFSGYCRVNGARYVRCCSRR
AP00222	VTCYCRRTRCGFRERLSGACGYRGRIYRLCCR
AP00225	ACYCRIGACVSGERLTGACGLNGRIYRLCCR
AP00228	LTCEIDRSLCLLHCRLKGYLRAYCSQQKVCRCVQ
AP00236	KSCCRNTWARNCYNVCRLPGTISREICAKKCDCKIISGTTCPSDYPK
AP00237	KSCCPTTTARNIYNTCRFGGGSRPVCAKLSGCKIISGTKCDSGWNH
AP00238	GGKPDLRPCHPPCHYIPRPKPR
AP00239	GWASKIGQTLGKIAKVGLKELIQPK

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00249	GLVSSIGRALGGLLADVVKSKGQPA
AP00254	GLWEKIKEKASELVSGIVEGVK
AP00257	GLWQKIKSAAGDLASGIVEGIKS
AP00261	GLFVGLAKVAAHNNPAIAEHFQA
AP00262	GFVDFLKKVAGTIANVVT
AP00263	GLLQTIKEKLESLESLAKGIVSGIQA
AP00264	GRKSDCFRKSGFCAFLKCPSLTLISGKCSRFYLCCKRIR
AP00266	GKREKCLRRNGFCAFLKCPTLSVISGTCSRFQVCCKTLLG
AP00272	DQYKCLQHGGFCLRSSCPSNTKLQGTCKPDKPNCCKS
AP00273	SIVPIRCRSNRDCRRFCGFRGGRCTYARQCLCGY
AP00276	VFQFLGKIIHHVGNFVHGFSHVF
AP00283	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK
AP00284	SHQDCYEALHKCMASHSKPFSCSMKFHMCLQQQ
AP00285	GLLCYCRKGHCKRGERVRGTCGIRFLYCCPRR
AP00286	QKLCERPSGTWSGVCGNNNACKNQCINLEKARHGSCNYVFPAHKCICYFPC
AP00291	MFFSSKKCKTVSKTFRGPCVRNAN
AP00295	EVERKHPLGGSRPGRCPTVPPGTFGHCACLCTGDASEPKGQKCCSN
AP00296	LLGRCKVKSNRFHGPCLTDTHCSTVCRGEGYKGGDCHGLRRRCMCLC
AP00298	LFCKGGSCHFGGCPSHLIKVGSCFGFRSCCKWPWNA
AP00301	GILDTIKSIASKVWNSKTVQDLKRKGINWVANKLGVSPQAA
AP00303	FCTMIPIPRCY
AP00304	RVCFAIPLPICH
AP00313	LSKKLICYCRIRGCKRRERVFGTCRNLFLTFVFCCS
AP00315	SLGSFLKGVGTTLASVGKVVSDQFGKLLQAGQG
AP00316	GIVDFAKKVVGGIRNALGI
AP00322	GILDVAKTLVGKLRNVLGI
AP00323	GVLDAFRKIATVVKNVV
AP00324	GVGDLIRKAVSVIKNIV
AP00325	GVIDAAKKVVNVLKNLF
AP00326	GVGSFIHKVVSAIKNVA
AP00327	GWFDVVKHIASAV
AP00330	GWLRKAAKSVGKFYYKHKYYIKAAWQIGKHAL
AP00332	GCASRCKAKCAGRRCKGWASASFRGRCYCKCFRC
AP00333	SCASRCKGHCRARRCGYYVSVLYRGRCYCKCLRC
AP00338	PDPAKTAPKKGSKKAVTKA
AP00339	FFGWLIKGAIHAGKAIHGLIHRRRH
AP00342	AKCIKNGKGCREDQGPPFCCSGFCYRQVGWARGYCKNR
AP00346	RWKIFKKIERVGQNVRDGIIKAGPAIQVLGTAKAL
AP00350	PWNIFKEIERAVARTRDAVISAGPAVRTVAAATSVAS
AP00354	VTCDILSVEAKGVKLNDAACAAHCLFRGRSGGYCNGKRVCVCR
AP00355	ANTAFVSSAHNTQKIPAGAPFNRNLRAMLADLRQNAAFAG
AP00356	QRFIHPTYRPPPQPRRPVIMRA
AP00357	FFPIGVFCKIFKTC
AP00358	FGLPMLSILPKALCILLKRKC
AP00359	DLRFLYPRGKLPVPTPPPFNPKPIYIDMGNRY
AP00364	VDKPDYRPRPWPRNMI
AP00366	GRFKRFRKKFKKLFKKLSPVIPLLHLG
AP00367	GGLRSLGRKILRAWKKYGPIIVPIIRIG
AP00368	GLFRRLRDSIRRGQQKILEKARRIGERIKDIFRG
AP00369	RIIDLLWRVRRPQKPKFVTVWVR

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00370	VGRFRRLRKKTRKRLKKIGKVLKWIPPIVGSIPLGCG
AP00371	GLLSRLRDFLSDRGRRLGEKIERIGQKIKDLSEFFQS
AP00374	GKVWDWIKSTAKKLWNSEPVKELKNTALNAAKNLVAEKIGATPS
AP00376	GWKDWAKKAGGWLKKKGPGMAKAALKAAMQ
AP00377	GWKDWLKKGKEWLKAKGPGIVKAALQAATQ
AP00379	DFKDWMKTAGEWLKKKGPGILKAAMAAAT
AP00382	GLVDVLGKVGGLIKKLLPG
AP00383	LLKELWTKMKGAGKAVLGKIKGLL
AP00385	FKLGSFLKKAWKSKLAKKLRAKGKEMLKDYAKGLLEGGSEEVPGQ
AP00386	WLGSALKIGAKLLPSVVGLFKKKKQ
AP00388	GIWGTLAKIGIKAVPRVISMLKKKKQ
AP00389	GIWGTALKWGVKLLPKLVGMAQTKKQ
AP00390	FWGALIKGAAKLIPSVVGLFKKKQ
AP00391	FIGTALGIASAIPAIVKLFK
AP00392	YRGGYTGPIPRPPPIGRPPLRLVVCACYRLSVSDARNCCIKFGSCCHLVK
AP00394	QVYKGGYTRPIPRPPPFVRPLPGGPIGPYNGCPVSCRGISFSQARSCCSRLGRCCHVGKGYS
AP00395	HSSGYTRPLPKPSRPIFIRPIGCDVCYGIPSSTARLCCFRYGDCCHR
AP00396	RRRPRPPYLPRPRPPPFFPPRLPPRIPPGFPPRFPPRFP
AP00397	SGFVLKGYTKTSQ
AP00399	HVDKKVADKVLLLKQLRIMRLLTRL
AP00400	YPPKPESPGEDASPEEMNKYLTALRHYINLVTRQRY
AP00401	GFTQGVRNSQSCRRNKGICVPIRCPGSMRQIGTCLGAQVKCCRRK
AP00402	KTCENLANTYRGPCFTTGSCDDHCKNKEHLRSGRCRDDFRCWCTRNC
AP00403	ACNFQSCWATCQAQHSIYFRRAFCDRSQCKCVFVRG
AP00404	YGPGDGHGGGHGGGHGGGHGNGQGGGHGHGPGGGFGGGHGGGHGGGGRGGGGSGGGGSPGHGA GGGYPGGHGGGHHGGYQTHGY
AP00405	FISAIASMLGKFL
AP00408	FLFPLITSFLSKVL
AP00409	ATTGCSCPQCIIFDPICASSYKNGRRGFSSGCHMRCYNRCHGTDYFQISKGSKCI
AP00410	PKRKSATKGDEPARRSARLSARPVPKPAAKPKKAAAPKKAVKGKKAAENGDAKAEAKVQAAGDGA GNAK
AP00411	KAVAAKKSPKKAKKPATPKKAAKSPKKVKKPAAAAKKAAKSPKKATKAAKPKAAKPKAAKAKKAA PKKK
AP00412	SLQPGAPNVNNKDQPWQVSPHISRDDSGNTRTDINVQRHGENNDFEAGWSKVVRGPNKAKPTWHIG GTHRW
AP00413	SLQGGAPNFPQPSQQNGGWQVSPDLGRDDKGNTRGQIEIQNKGKDHDFNAGWGKVIRGPNKAKPT WHVGGTYRR
AP00414	SIGSALKKALPVAKKIGKIALPIAKAALP
AP00416	SLGGVISGAKKVAKVAIPIGKAVLPVVAKLVG
AP00417	SIGTAVKKAVPIAKKVGKVAIPIAKAVLSVVGQLVG
AP00418	GLRKRLRKFRNKIKEKLKKIGQKIQGFVPKLAPRTDY
AP00424	GFLGPLLKLAAKGVAKVIPHLIPSRQQ
AP00425	GCWSTVLGGLKKFAKGGLEAIVNPK
AP00426	GVFLDALKKFAKGGMNAVLNPK
AP00427	GLLGPLLKIAAKVGSNLL
AP00428	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPF GWKSIFIQC
AP00429	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAI CVDIKICKE
AP00430	ILGKIWEGIKSLF
AP00431	TWLKKRRWKKAKPP
AP00432	KKKKPLFGLFFGLF

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00433	SSLLEKGLDGAKKAVGGLGKLGKDAVEDLESVGKGAVHDVKDVLDSV
AP00434	GLMSVLGHAVGNVLGGLFKS
AP00435	GWFGKAFRSVSNFYKKHKTYIHAGLSAATLL
AP00436	LCNERPSQTWSGNCGNTAHCDKQCQDWEKASHGACHKRENHWKCFCYFNC
AP00437	EFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS
AP00438	GFGCPNNYQCHRHCKSIPGRCGGYCGGWHRLPCTCYRCG
AP00439	VTCFCKRPVCDSGETQIGYCRLGNTFYRLCCRQ
AP00440	VTCFCRRRGCASRERHIGYCRFGNTIYRLCCRR
AP00445	GFCRCLCRRGVCRCICTR
AP00446	GADFQECMKEHSQKQHQHQG
AP00449	SYSMEHFRWGKPV
AP00450	RICRIIFLRVCR
AP00451	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK
AP00453	FLPAIVGAAGQFLPKIFCAISKKC
AP00455	FFPIVAGVAGQVLKKIYCTISKKC
AP00456	VNPIILGVLPKFVCLITKKC
AP00457	GLWETIKNFGKKFTLNILHKLKCKIGGGC
AP00459	FITLLLRKFICSITKKC
AP00461	FLPMLAGLAASMVPKLVCLITKKC
AP00470	FLPIIASVAAKVFSKIFCAISKKC
AP00474	FIHHIFRGIVHAGRSIGRFLTG
AP00475	GLNTLKKVFQGLHEAIKLINNHVQ
AP00479	AGCIKNGGRCNASAGPPYCCSSYCFQIAGQSYGVCKNR
AP00480	VGIGTPIFSYGGGAGHVPEYF
AP00481	FFSASCVPGADKGQFPNLCRLCAGTGENKCA
AP00482	FSFKRLKGFAKKLWNSKLARKIRTKGLKYVKNFAKDMLSEGEEAPPAAEPPVEAPQ
AP00483	KTCEHLADTYRGVCFTNASCDDHCKNKAHLISGTCHNWKCFCTQNC
AP00485	GFGALFKFLAKKVAKTVAKQAAKQGAKYVVNKQME
AP00489	SGRGKTGGKARAKAKTRSSRAGLQFPVGRVHRLLRKGNYAHRVGAGAPVYL
AP00490	AALKGCWTKSIPPKPCSGKR
AP00492	RQRVEELSKFSKKGAAARRRK
AP00493	NLVSGLIEARKYLEQLHRKLKNCKV
AP00495	EQCGRQAGGKLCPNNLCCSQYGWCGSSDDYCSPSKNCQSNCKGGG
AP00496	AKKVFKRLEKLFSKIQNDK
AP00497	ILGPVLGLVSDTLDDVLGIL
AP00499	VGALAVVVWLWLWLW
AP00501	GIGKHVGKALKGLKGLLKGLGES
AP00502	FLRFIGSVIHGIGHLVHHIGVAL
AP00503	FLGVVFKLASKVFPAVFGKV
AP00504	LAHQKPFIRKSYKCLHKRCR
AP00506	KLAKLAKKLAKLAK
AP00507	GLLSKVLGVGKKVLCGVSGLC
AP00508	GLLDSIKGMAISAGKGALQNLLKVASCKLDKTC
AP00509	VAIALKAAHYHTHKE
AP00510	ILQKAVLDCLKAAGSSLSKAAITAIYNKIT
AP00514	FLGGLMKAFPALICAVTKKC
AP00516	IWLTALKFLGKHAAKHLAKQQLSKL
AP00517	KIKWFKTMKSIAKFIAKEQMKKHLGGE
AP00518	QYRHRCCAWGPGRKYCKRWC
AP00519	QWGRRCCGWGPGRRYCRRWC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00522	INWLKLGKAIIDAL
AP00524	GIGDPVTCLKSGAICHPVFCPRRYKQIGTCGLPGTKCCKKP
AP00527	ILGPVISKIGGVLGGLLKNL
AP00530	VLSKSLCTPGCITGPLQTCYLCFPTFAKC
AP00531	GKQYFPKVGGRLSGKAPLAAKTHRRLKP
AP00533	GVVDILKGAAKDIAGHLASKVMNKL
AP00535	GLGSVFGRLARILGRVIPKVAKKLGPKVAKVLPKVMKEAIPMAVEMAKSQEEQQPQ
AP00536	SVRTQDNAVNRQIFGSNGPYRDFQLSDCYLPLETNPYCNEWQFAYHWNNALMDCERAIYHGCNRTR NNFITLTACKNQAGPICNRRRH
AP00537	AEVAPAPAAAAPAKAPKKKAAAKPKKAGPS
AP00538	WLNALLHHGLNCAKGVLA
AP00539	GFGCPWNRYQCHSHCRSIGRLGGYCAGSLRLTCTCYRS
AP00540	GLLDTLKGAAKNVVGSLASKVMEKL
AP00541	IDWKKLLDAAKQIL
AP00543	GVVDILKGAGKDLLAHLVGKISEKV
AP00544	GVLDIFKDAAKQILAHAAEKQI
AP00547	RSNKGFNFMVDMIQALSK
AP00548	RFGRFLRKIRRFRPKVTITIQGSARFG
AP00549	GFGCNGPWDEDDMQCHNHCKSIKGYKGGYCAKGGFVCKCY
AP00552	GIGRKFLGGVKTTFRCGVKDFASKHLY
AP00556	GFMKYIGPLIPHAVKAISDLI
AP00557	RVKRVWPLVIRTVIAGYNLYRAIKKK
AP00558	GFGCPGNQLKCNNHCKSISCRAGYCDAATLWLRCTCTDCNGKK
AP00559	ATRVVYCNRRSGSVVGGDDTVYYEG
AP00560	TTLTLHNLCPYPVWWLVTPNNGGFPIIDNTPVVLG
AP00561	GWKIGKKLEHHGQNIRDGLISAGPAVFAVGQAATIYAAAK
AP00564	FLIGMTHGLICLISRKC
AP00567	VWPLGLVICKALKIC
AP00568	GLFSVVTGVLKAVGKNVAKNVGGSLLEQLKCKKISGGC
AP00569	FLPLLLAGLPLKLCFLFKKC
AP00570	SIITMTKEAKLPQLWKQIACRLYNTC
AP00572	GIMDTIKDTAKTVAVGLLNKLKCKITGC
AP00573	GLFSKFNKKKIKSGLIKIIKTAGKEAGLEALRTGIDVIGCKIKGEC
AP00575	GLLDTFKNLALNAAKSAGVSVLNSLSCKLSKTC
AP00576	GVLGTVKNLLIGAGKSAAQSVLKTLSCKLSNDC
AP00577	GLFTLIKGAAKLIGKTVAKEAGKTGLELMACKITNQC
AP00582	GFSSLFKAGAKYLLKSVGKAGAQQLACKAANNCA
AP00583	GVITDALKGAAKTVAAELLRKAHCKLTNSC
AP00584	VIDDLKKVAKKVRRELLCKKHHKKLN
AP00585	SIWEGIKNAGKGFLVSILDKVRCKVAGGCNP
AP00588	FLGSIVGALASALPSLISKIRN
AP00589	FLGALAKIISGIF
AP00591	FLPLLGNLLRGLL
AP00594	EGTWQHGYGVSSAYSNYHHGSKTHSATVVNNNTGRQGKDTQRAGVWAKATVGRNLTEKASFYYNF W
AP00596	FLPLVTGLLSGLL
AP00598	FLSAITSLLGKLL
AP00599	GIWDTIKSMGKVFAGKILQNL
AP00600	GLLRASSVWGRKYYVDLAGCAKA
AP00601	FLSLALAALPKFLCLVFKKC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00605	ILPILSLIGGLLGK
AP00606	GLLSAVKGVLKGAGKNVAGSLMDKLKCKLFGGC
AP00611	FIGPIISALASLFG
AP00612	AAEFPDFYDSEEQMGPHQEAEDEKDRADQRVLTEEEKKELENLAAMDLELQKIAEKFSQR
AP00613	RVKRFWPLVPVAINTVAAGINLYKAIRRK
AP00614	ALFSILRGLKKLGKMGQAFVNCEIYKKC
AP00616	ALSILRGLEKLAKMGIALTNCKATKKC
AP00618	GFLSTVKNLATNVAGTVLDTIRCKVTGGCRP
AP00621	GIFPKIIGKGIKTGIVNGIKSLVKGVGMKVFKAGLNNIGNTGCNEDEC
AP00630	GEILCNLCTGLINTLENLLTTKGADKVKDYISSLCNKASGFIATLCTKVLDFGIDKLIQLIEDKVDANAIC AKIHAC
AP00632	KYYGNGVSCNKNGCTVDWSKAIGIIGNNAAANLTTGGAAGWNKG
AP00634	KYYGNGVTCGKHSCSVDWGKATTCIINNGAMAWATGGHQGNHKC
AP00639	GLIGSIGKALGGLLVDVLKPKLQAAS
AP00640	GLLGLLGSVVSHVVPAIVGHF
AP00641	GFFALIPKIISSPLFKTLLSAVGSALSSSGEQE
AP00646	FFPNVASVPGQVLLKKIFCAISKKC
AP00650	GIFTKINKKKAKTGVFNIIKTIGKEAGMDVIRAGIDTISCKIKGEC
AP00651	GLFSILKGVGKIALKGLAKNMGKMGLDLVSCKISKEC
AP00654	GLLDTIKNTAKNLAVGLLDKIKCKMTGC
AP00658	FLPLVGKILSGLI
AP00660	FWGALAKGALKLIPSLFSSFSKKD
AP00661	GILSLFTGGIKALGKTLFKMAGKAGAEHLACKATNQC
AP00662	GLFSILRGAAKFASKGLGKDLTKLGVDLVACKISKQC
AP00664	FLPAIAGILSQLF
AP00666	EGGGPQWAVGHFM
AP00667	EPHPDEFVGLM
AP00670	EPNPDEFFGLM
AP00672	DCLSGRYKGPCAVWDNETCRRVCKEEGRSSGHCSPSLKCWCEGC
AP00673	VGSRYLCTPGSCWKLVCFTTTVK
AP00674	ITSVSWCTPGCTSEGGGSGCSHCC
AP00675	FELDRICGYGTARCRKKCRSQEYRIGRCPNTYACCLRKWDESLLNRTKP
AP00676	RLGNFFRKVKEKIGGGLKKVGQKIKDFLGNLVPRTAS
AP00677	GLRKKFRKTRKRIQKLGRKIGKTGRKVWKAWREYGQIPYPCRI
AP00678	RLKELITTGGQKIGEKIRRIGQRIKDFFKNLQPREEKS
AP00682	RRLRPRHQHFPSERPWPKPLPLPLPRPGPRPWPKPLPLPLPRPGLRPWPKPL
AP00684	RRLRPRRPRLPRPRPRPRPRPRSLPLPRPKPRPIPRPLPLPRPRPKPIPRPLPLPRPRPRRIPRPLPLPR PRPRPIPRPLPLPQPQPSPIPRPL
AP00686	KRFGRLAKSFLRMRILLPRRKILLAS
AP00687	KRRHWFPLSFQEFLEQLRRFRDQLPFP
AP00688	KRFHSVGSLIQRHQQMIRDKSEATRHGIRIITRPKLLLAS
AP00689	AFPPPNVPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPRFPPPNFPGPPFPPPIFPGPWFPPPPPFR PPPFGPPRFP
AP00694	AIGSILGALAKGLPTLISWIKNR
AP00696	GLFDIIKNIVSTL
AP00698	GLWQLIKDKIKDAATGFVTGIQS
AP00702	GLLGSIGNAIGAFIANKLKPK
AP00704	GLLGSIGKVLGGYLAEKLKPK
AP00707	RPDKPRPYLPRPRPPRPVR
AP00714	GYGCPFNQYQCHSHCSGIRGYKGGYCKGTFKQTCKCY
AP00716	RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00717	RTCESQSHRFHGTCVRESNCASVCQTEGFIGGNCRAFRRRCFCTRNC
AP00718	KICRRRSAGFKGPCMSNKNCAQVCQQEQWQQQNCDQPFRRCKCIRQC
AP00722	GLLNGLALRLGKRALKKIIKRLCR
AP00723	SLLSLIRKLIT
AP00727	RWCVYAYVRVRGVLVRYRRCW
AP00730	GSVLNCGETCLLGTCYTTGCTCNKYRVCTKD
AP00731	SFGLCRLRRGFCARGRCRFPSIPIGRCSRFVQCCRRVW
AP00736	RLGDILQKAREKIEGGLKKLVQKIKDFFGKFAPRTES
AP00737	GLVTSLIKGAGKLLGGLFGSVTGGQS
AP00739	GVVTDLLKTAGKLLGNLFGSLSG
AP00741	PITYLDAILAAVRLLNQRISGPCILRLREAQPRPGWVGTLQRRREVSFLVEDGPCPPGVDCRSCEPGALQHCVGTVSIEQQPTAELRCRPLRPQ
AP00743	RYHMQCGYRGTFCTPGKCPYGNAYLGLCRPKYSCCRWL
AP00745	MTPFWRGVSLRPVGASCRDNSECITMLCRKNRCFLRTASE
AP00748	DIQIPGIKKPTHRDIIIPNWNPNVRTQPWQRFGGNKS
AP00749	EADEPLWLYKGDNIERAPTTADHPILPSIIDDVKLDPNRRYA
AP00750	EIRLPEPFRFPSPTVPKPIDIDPILPHPWSPRQTYPIIARRS
AP00753	VQETQKLAKTVGANLEETNKKLAPQIKSAYDDFVKQAQEVQKKLHEAASKQ
AP00754	ETESTPDYLKNIQQQLEEYTKNFNTQVQNAFDSDKIKSEVNNFIESLGKILNTEKKEAPK
AP00755	ENFFKEIERAGQRIRDAIISAAPAVETLAQAQKIIKGGD
AP00756	ALWKDILKNAGKAALNEINQLVNQ
AP00764	GLRSKIWLWVLLMIWQESNKFKKM
AP00765	MHDFWVLWVLLEYIYNSACSVLSATSSVSSRVLNRSLQVKVVKITN
AP00767	VAGALGVQTAAATTIVNVILNAGTLVTVLGIIASIASGGAGTLMTIGWATFKATVQKLAKQSMARAIAY
AP00768	PNWTKIGKCAGSIAWAIGSGLFGGAKLIKIKKYIAELGGLQKAAKLLVGATTWEEKLHAGGYALINLAAELTGVAGIQANCF
AP00769	GLLGAMFKVASKVLPHVVPAITEHF
AP00772	FRGLAKLLKIGLKSFARVLKKVLPKAAKAGKALAKSMADENAIRQQNQ
AP00773	GKFSVFGKILRSIAKVFKGVGKVRKQFKTASDLDKNQ
AP00777	GKGRWLERIGKAGGIIIGGALDHL
AP00779	GRRKRKWLRRIGKGVKIIGGAALDHL
AP00780	GRRRRSVQWCAVSQPEATKCFQWQRNMRKVRGPPVSCIKRDSPIQCIQA
AP00781	FLGALIKGAIHGGRFIHGMIQNHH
AP00782	GWGSIFKHGRHAAKHIGHAAVNHYL
AP00784	FFRLLFHGVHHVGKIKPRA
AP00787	GWRLLLKKAEVKTVGKLALKHYL
AP00788	AGWGSIFKHIFKAGKFIHGAIQAHND
AP00789	GFWGKLFKLGLHGIGLLHLHL
AP00791	GWKKWLRKGAKHLGQAAIKGLAS
AP00792	FLGLLFHGVHHVGKWIHGLIHGHH
AP00796	IIGPVLGLVGKPLESLLE
AP00805	RSALSCQMCELVVKKYEGSADKDANVIKKDFDAECKKLFHTIPFGTRECDHYVNSKVDPIIHELEGGTAPKDVCTKLNECP
AP00806	HHQELCTKGDDALVTELECIRLRISPETNAAFDNAVQQLNCLNRACAYRKMCATNNLEQAMSVYFTNEQIKEIHDAATACDPEAHHEHDH
AP00807	NRWYCNSAAGGVGGAAGCVLAGYVGEAKENIAGEVRKGWGMAGGFTHNKACKSFPGSGWASG
AP00809	GIKCRFCCGCCTPGICGVCCRF
AP00812	FAEPLPSEEEGESYSKEPPEMEKRYGGFM
AP00813	CIKNGNGCQPNGSQNGCCSGYCHKQPGWVAGYCRRK
AP00814	GLGSILGKILNVAGKVGKTIGKVADAVGNKE

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00816	GLGSFFKNAIKIAGKVGSTIGKVADAIGNKE
AP00817	FLPLLASLFSRLL
AP00822	GIFNVFKGALKTAGKHVAGSLLNQLKCKVSGEC
AP00823	SILPTIVSFLSKVF
AP00830	GLLLDTLKGAAKDIAGIALEKLKCKITGCKP
AP00833	CLAGRLDKQCTCRRSQPSRRSGHEVGRPSPHCGPSRQCGCHMD
AP00834	KVNANAIKKGGKAIGKGFKVISAASTAHDVYEHIKNRRH
AP00835	GKIPVKAIKKGGQIIGKALRGINIASTAHDIISQFKPKKKKNH
AP00836	KVPIGAIKKGGKIIKKGLGVIGAAGTAHEVYSHVKNRH
AP00840	KGIGSALKKGGKIIKGGLGALGAIGTGQQVYEHVQNRQ
AP00841	TTHSGKYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAIPGKC
AP00842	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLGGAIPGKC
AP00845	KSYGNGVHCNKKKCWVDWGSAISTIGNNSAANWATGGAAGWKS
AP00848	KNYGNGVHCTKKGCSVDWGYAWANIANNSVMNGLTGGNAGWHN
AP00850	KYYGNGVSCNSHGCSVNWGQAWTCGVNHLANGGHGVC
AP00852	KYYGNGLSCSKKGCTVNWGQAFSCGVNRVATAGHHKC
AP00853	ATRSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH
AP00857	SSMKLSFRARAYGFRGPGPQL
AP00863	FLPIVGKLLSGLSGLL
AP00866	FLPIIAKVLSGLL
AP00867	FLPVIAGLLSKLF
AP00869	ILPLVGNLLNDLL
AP00871	FLPFLKSILGKIL
AP00872	FLPFFASLLGKLL
AP00875	FLSSIGKILGNLL
AP00876	FLSIIAKVLGSLF
AP00877	FLGSLIGAAIPAIKQLLGLKK
AP00878	FLPILASLAAKFGPKLFCLVTKKC
AP00879	GRLRNLIEKAGQNIRGKIQGIGRRIKDILKNLQPRPQV
AP00883	FLPLIASVAANLAPKIICKITKTC
AP00884	QLKVDLWGTRSGIQPEQHSSGKSDVRRWRSRY
AP00889	APPGARPPPGPPPPGPPPPGP
AP00890	IDWKKVDWKKVSKKTCKVMLKACKFLG
AP00891	IIGLVSKGTCVLVKTVCKKVLKQG
AP00892	PDITKLNIKKLTKATCKVISKGASMCKVLFDKKKQE
AP00893	DVKGMKKAIKGILDCVIEKGYDKLAAKLKKVIQQLWE
AP00894	GLLDFVTGVGKDIFAQLIKQI
AP00895	KRFKKFFKKLKNSVKKRAKKFFKKPRVIGVSIPF
AP00898	FLSGIVGMLGKLF
AP00900	FLSHIAGFLSNLF
AP00901	GWMSKIASGIGTFLSGVQQG
AP00902	LRPAVIRPKGK
AP00904	LGPALITRKPLKGKP
AP00906	FRPALIVRTKGTRL
AP00910	GLVSDLLSTVTGLLGNLGGGGLKKI
AP00911	FLSLIPHIVSGVAALAKHL
AP00915	QQCGRQAGNRRCANNLCCSQYGYCGRTNEYCCTSQGCQSQCRRCG
AP00918	ELCEKASKTWSGNCGNTGHCDNQCKSWEGAAHGACHVRNGKHMCFCYFNC
AP00920	NTCENLAGSYKGVCFGGCDRHCRTQEGAISGRCRDDFRCWCTKNC
AP00921	RVCMKGSAGFKGLCMRDQNCAQVCLQEGWGGGNCDGVMRQCKCIRQC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP00927	IYFIADKMGIQLAPAWYQDIVNWVSAGGTLTTGFAIIVGVTVPAWIAEAAAAFGIASA
AP00928	NKGCATCSIGAACLVDGPIPDFEIAGATGLFGLWG
AP00929	MAKEFGIPAAVAGTVINVVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVIA W
AP00930	KPAWCWYTLAMCGAGYDSGTCDYMYSHCFGVKHSSGGGGSYHC
AP00931	LAGYTGIASGTAKKVVDAIDKGAAAFVIISIISTVISAGALGAVSASADFIILTVKNYISRNLKAQAVIW
AP00939	ALWKTLLKGAGKVFGHVAKQFLGSQGQPES
AP00940	GLWSKIKEAAKTAGKMAMGFVNDMV
AP00942	GLWKSLLKNVGVAAGKAALNAVTDMVNQ
AP00949	GLWSTIKQKGKEAAIAAAKAAGQAVLNSASEAL
AP00959	ALWKTLLKKVGKVAGKAVLNAVTNMANQNEQ
AP00960	AVWKDFLKNIGKAAGKAVLNSVTDMVNE
AP00964	GLWSKIKEAAKAAGKAALNAVTGLVNQGDQPS
AP00965	SVLSTITDMAKAAGRAALNAITGLVNQ
AP00973	LLGMIPLAISAISALSKL
AP00982	GTCKAECPTWEGICINKAPCVKCCKAQPEKFTDGHCSKILRRCLCTKPC
AP00983	QNNICKTTSKHFKGLCFADSKCRKVCIQEDKFEDGHCSKLQRKCLCTKNC
AP00986	KSTCKAESNTFPGLCITKPPCRKACLSEKFTDGKCSKILRRCICYKPC
AP00987	SRWPSPGRPRPFPGRPKPIFRPRPCNCYAPPCPCDRW
AP00993	GIFSSRKCKTPSKTFKGICTRDSNCDTSCRYEGYPAGDCKGIRRRCMCSKPC
AP00994	GIFSNMYARTPAGYFRGPAGYAAN
AP00996	ISLEICAIFHDN
AP00998	ALPKKLKYLNLFNDGFNYMGVV
AP01001	NRWWQGVVPTVSYECRMNSWQHVFTCC
AP01003	FKSWSFCTPGCAKTGSFNSYCC
AP01005	YVSCLFRGARCRVYSGRSCCFGYYCRRDFPGSIFGTCSRRNF
AP01009	DYDWSLRGPPKCATYGQKCRTWSPRNCCWNLRCKAFRCRPR
AP01011	GLFGKLIKKFGRKAISYAVKKARGKH
AP01012	SWKSMAKKLKEYMEKLKQRA
AP01014	GLKDKFKSMGEKLKQYIQTWKAKF
AP01016	GFFGKMKEYFKKFGASFKRRFANLKKRL
AP01019	GETFDKLKEKLKTFYQKLVEKAEDLKGDLKAKLS
AP01020	QAFKTFTPDWNKIRNDAKRMQDNLEQMKKRFNLNL
AP01023	GTACGESCYVLPCFTVGCTCTSSQCFKN
AP01031	GVPICGETCTLGTCYTAGCSCSWPVCTRN
AP01034	GDPTFCGETCRVIPVCTYSAALGCTCDDRSDGLCKRN
AP01045	GTLPCGESCVWIPCISAVVGCSCKSKVCYKN
AP01056	SAIACGESCVYIPCFIPGCSCRNRVCYLN
AP01058	SISCGESCAMISFCFTEVIGCSCKNKVCYLN
AP01061	KIPCGESCVWIPCVTSIFNCKCKENKVCYHD
AP01063	KVCYRAIPCGESCVWIPCISAAIGCSCKN
AP01065	GSIPACGESCFKGKCYTPGCSCSKYPLCAKN
AP01068	GLPCGETTCFTGKCYTPGCSCSYPICKKIN
AP01070	GIPCGESCVYIIPCTVTALAQCKCKSKVCYN
AP01075	GLPTCGETCFGGTCNTPGCTCDSSWPICTHN
AP01076	DIFCGETCAFIPCITHVPGTCSCKSKVCYFN
AP01077	GGTIFDCGETCFLGTCYTPGCSCGNYGFCYGTN
AP01080	GVPCGESCVFIPCITGVIGCSCSSNVCYLN
AP01083	GSPIQCAETCFIGKCYTEELGCTCTAFLCMKN
AP01087	SISCGETCTTFNCWIPNCKCNHHDKVCYWN

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01093	GTPCAESCVYLPCFTGVIGCTCKDKVCYLN
AP01094	GNIPCGESCIFFPCFNPGCSCKDNLCYYN
AP01100	CGETCIWGRCYSENIGCHCGFGICTLN
AP01101	RGCYKICGETCLFIPCLTSVFGCSCKN
AP01102	CGETCVVDTRCYTKKCSCAWPVCMRN
AP01105	CGETCKVTKRCSGQGCSCLKGRSCYD
AP01107	CGETCIYIPCFTEAVGCKCKDKVCYKN
AP01111	GSLCGDTCFVLGCNDSSCSCNYPICVKD
AP01114	GLPVCGESCFGGSCYTPGCSCTWPICTRD
AP01129	GLGSLVGNALRIGAKLL
AP01130	GMASKAGSVLGKVAKVALKAAL
AP01131	MSWLNFLKYIAKYGKKAVSAAWKYKGKVLEWLNVGPTLEWVWQKLKKIAGL
AP01142	KPYCSCKWRCGIGEEEKGICHKFPIVTYVCCRRP
AP01148	IATQCRIRGGFCRVGSCRFPHIAIGKCATFISCCGRAY
AP01150	ELPKLPDDKVLIRSRSNCPKGKVWNGFDCKSPFAFS
AP01151	GTWDDIGQGIGRVAYWVGKALGNLSDVNQASRINRKKKH
AP01154	AIKLVQSPNGNFAASFVLDGTKWIFKSKYYDSSKGYWVGIYEVWDRK
AP01155	ESVFSKIGNAVGPAAYWILKGLGNMSDVNQADRINRKKH
AP01156	NKLAYNMGHYAGKATIFGLAAWALLA
AP01157	QRGSRGQRCGPGEVFNQCGSACPRVCGRPPAQACTLQCVSGCFCRRGYIRTQRGGCIPERQCHQR
AP01160	QDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
AP01162	QTCASRCPRPCNAGLCCSIYGYCGSGAAYCGAGNCRCQCRG
AP01165	MINRTDCNENSYLEIHNNEGRDTLCFANAGTMPVAIYGVNWVESGNNVVTLQFQRNLSDPRLETITLQ KWGSWNPGHIHEILSIRIY
AP01166	AVRIGPCDQVCPRIVPERHECCRAHGRSGYAYCSGGGMYCN
AP01167	LTTKLWSSWGYYLGKKARWNLKHPYVQF
AP01168	LVAYGIAQGTAEKVVSLINAGLTVGSIISILGGVTVGLSGVFTAVKAAIAKQGIKKAIQL
AP01169	NRWGDTVLSAASGAGTGIKACKSFGPWGMAICGVGGAAIGGYFGYTHN
AP01170	YSSKDCLKDIGKGIGAGTVAGAAGGGLAAGLGAIPGAFVGAHFGVIGGSAACIGGLLGN
AP01171	YSGKDCLKDMGGYALAGAGSGALWGAPAGGVGALPGAFVGAHVGAIAGGFACMGGMIGNKFN
AP01172	KRGPNCVGNFLGGLFAGAAAGVPLGPAGIVGGANLGMVGGALTCL
AP01174	KVSGGEAVAAIGICATASAAIGGLAGATLVTPYCVGTWGLIRSH
AP01175	GMSGYIQGIPDFLKGYLHGISAANKHKKGRLGY
AP01176	TTPACFTIGLGVGALFSAKFC
AP01177	FNRGGYNFGKSVRHVVDAIGSVAGILKSIR
AP01179	NGVYCNKQKCWVDWSRARSEIIDRGVKAYVNGFTKVLGGIGGR
AP01180	NPKVAHCASQIGRSTAWGAVSGA
AP01181	AYPGNGVHCGKYSCTVDKQTAIGNIGNNAA
AP01182	FTPSVSFSQNGGVVEAAAQRGYIYKKYPKGAKVPNKVKMLVNIRGKQTMRTCYLMSWTASSRTAKY YYYI
AP01183	ATYYGNGLYCNKEKCWVDWNQAKGEIGKIIVNGWVNHGPWAPRR
AP01185	ENDHRMPNNLNRPNNLSKGGAKCGAAIAGGLFGIPKGPLAWAAGLANVYSKCN
AP01186	KTYYGTNGVHCTKKSLWGKVRLKNVIPGTLCRKQSLPIKQDLKILLGWATGAFGKTFH
AP01187	MNFLKNGIAKWMTGAELQAYKKKYGCLPWEKISC
AP01188	MLAKIKAMIKKFPNPYTLAAKLTTYEINWYKQQYGRYPWERPVA
AP01189	APAGLVAKFGRPIVKKYYKQIMQFIGEGSAINKIIPWIARMWRT
AP01192	SDCNINSNTAADVILCFNQVGSCALCSPTLVGGPVP
AP01193	DIDITGCSACKYAAGQVCTIGCSAAGGFICGLLGITIPVAGLSCLGFVEIVCTVADEYSGCGDAVAKEA CNRAGLC
AP01194	CSTNTFSLSDYWGNNGAWCTLTHECMAWCK
AP01195	KRGSGWIATITDDCPNSVFVCC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01196	GETDPNTQLLNDLGNNMAWGAALGAPGGLGSAALGAAGGALQTVGQGLIDHGPVNVFIPVLIGPSW NGSGSGYNSATSSSGSGS
AP01198	LSCDEGMLAVGGLGAVGGPWGAAVGVLVGAALYCF
AP01199	KYYGNGVHCGKKTCYVDWGQATASIGKIIVNGWTQHGPWAHR
AP01201	KGGSGVIHTISHECNMNSWQFVFTCCS
AP01204	GKNGVFKTISHECHLNTWAFLATCCS
AP01205	STPVLASVAVSMELLPTASVLYSDVAGCFKYSAKHHC
AP01206	CTFTLPGGGGVCTLTSECIC
AP01208	GICRCICGRRICRCICGR
AP01213	EFRGSIVIQGTKEGKSRPSLDIDYKQRVYDKNGMTGDAYGGLNIRPGQPSRQHAGFEFGKEYKNGFIK GQSEVQRGPGGRLSPYFGINGGFRF
AP01215	FVPYNPPRPYQSKPFPSFPGHGPFNPKIQWPYPLPNPGH
AP01217	GFRDVLKGAAKQFVKTVAGHIANI
AP01219	GFKDWIKGAAKKLIKTVAANIANQ
AP01223	GFKDLLKGAAKALVKTVLF
AP01228	ASGRDIAMAIGTLSGQFVAGGIGAAAGGVAGGAIYDYASTHKPNPAMSPSGLGGTIKQKPEGIPSEAW NYAAGRLCNWSPNNLSDVCL
AP01229	GDVNWVDVGKTVATNGAGVIGGAFGAGLCGPVCAGAFAVGSSAAVAALYDAAGNSNSAKQKPEGL PPEAWNYAEGRMCNWSPNNLSDVCL
AP01230	DGNDGQAELIAIGSLAGTFISPGFGSIAGAYIGDKVHSWATTATVSPSMSPSGIGLSSQFGSGRGTSSASS SAGSGS
AP01231	GGAPATSANAAGAAAIVGALAGIPGGPLGVVVGAVSAGLTTGIGSTVGSGSASSSAGGGS
AP01232	MNLNGLPASTNVIDLRGKDMGTYIDANGACWAPDTPSIIMYPGGSGPSYSMSSSTSSANSGS
AP01233	QKKPPRPPQWAVGHFM
AP01234	FSKYERQKDKRPYSERKNQYTGPQFLYPPERIPPQKVIKWNEEGLPIYEIPGEGGHAEPAAA
AP01235	FNKLKQGSSKRTCAKCFRKIMPSVHELDERRRGANRWAAGFRKCVSSICRY
AP01238	NPLIPAIYIGATVGPSVWAYLVALVGAAAVTAANIRRASSDNHSCAGNRGWCRSKCFRHEYVDTYYSA VCGRYFCCRSR
AP01240	ALKAALLAILKIVRVIKK
AP01241	FASLLGKALKALAKQ
AP01242	GLLSFLPKVIGVIGHLIHPPS
AP01243	KGAPCAKKPCCGPLGHYKVDCSTIPDYPCCSKYGFCGSGPQYCG
AP01245	AVTCNTVVSSLAPCVPFFAGSAAQPTAACCNGVRSLNSAARTTPDRRTACNCIKSSASSIGLNYNKAA KLPSRCTVNVTVPISPSVNCAT
AP01246	FLPKTLRKFFCRIRGGRCAVLNCLGKEEQIGRCSNSGRKCCRKKK
AP01247	FMPIIGRLMSGSL
AP01249	GILDAIKAIAKAAG
AP01253	GLMSLFKGVLKTAGKHIFKNVGGSLLDQAKCKITGEC
AP01258	GLMDVFKGAAKNLLASALDKIRCKVTKC
AP01260	IIGHLIKTALGMLGL
AP01261	IIEKLVNTALGLLSGL
AP01262	GLADFLNKAVGKVVDFVKS
AP01263	FLPLVTMLLGKLF
AP01264	RIGVLLARLPKLFSLFKLMGKKV
AP01266	AVDLAKIANKVLSSLF
AP01269	GFLSILKKVLPKVMAHMK
AP01277	KSCCPNTTGRNIYNACRLTGAPRPTCAKLSGCKIISGSTCPSDYPK
AP01283	MRKEFHNVLSSGQLLADKRPARDYNRK
AP01285	AGDPLADPNSQIVRQIMSNAAWGPPLVPERFRGMAVGAAGGVTQTVLQGAAAHMPVNVPIPKVPMG PSWNGSKG
AP01287	IRNSLTCRFNFGICLPKRCPGRMRQIGTCF
AP01289	NSKRACYREGGECLQRCIGLFHKIGTCNFRFKCCKFQ

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01290	NEPVSCIRNGGICQYRCIGLRHKIGTCGSPFKCCK
AP01294	GLGGAKKNFIIAANKTAPQSVKKTFSCKLYNG
AP01296	FMPILSCSRFKRC
AP01298	GLFTLIKCAYQLIAPTVACN
AP01299	GLFTLIKGAAKLIGKTVPKKQARLGMNLWLVKLPTNVKT
AP01300	ATAVDFGPHGLLPIRPIRIRPLCGKDKS
AP01302	GFSPNLPGKGLRIS
AP01303	VIPFVASVAAEMMQHVYCAASKKC
AP01304	GLLSGILGAGKHIVCGLSGPCQSLNRKSSDVEYHLAKC
AP01305	FLPPSPWKETFRTS
AP01306	TSRCYIGYRRKVVCS
AP01307	GCSRWIIGIHGQICRD
AP01308	GLLSGTSVRGSI
AP01315	ARLKKCFNKVTGYCRKKCKVGERYEIGCLSGKLCCAN
AP01316	NPANPLNLKKHHGVFCDVCKALVEGGEKVGDDDLDAWLDVNIGTLCWTMLLPLHHECEEELKKVKK ELKKDIENKDSPDKACKDVDLC
AP01317	GAILCNLCKDTVKLVENLLTVDGAQAVRQYIDNLCGKASGFLGTLCEKILSFGVDELVKLIENHVDPV VVCEKIHAC
AP01318	IPVLCPVCTSLVGKLIDLVLGGAVDKVTDYLETLCAKADGLVETLCTKIVSYGIDKLIEKILEGGSAKLI CGLIHAC
AP01322	IPRPLDPCIAQNGRCFTGICRYPYFWIGTCRNGKSCCRRR
AP01323	LPVNEAQCRQVGGYCGLRICNFPSRFLGLCTRNHPCCSRVWV
AP01324	GPDSCNHDRGLCRVGNCNPGEYLAKYCFEPVILCCKPLSPTPTKT
AP01325	QPFIPRPIDTCRLRNGICFPGICRRPYYWIGTCNNGIGSCCARGWRS
AP01326	SKGKKANKDVELARG
AP01329	KQQLATEAESAGPIL
AP01331	IFGAILPLALGALKNLIK
AP01332	FIGAILPAIAGLVHGLINR
AP01339	FLSFPTTKTYFPHFDLSHGSAQVKGHGAK
AP01341	SVIGCEICEWLVATAEGFVNKTKPQIEQELLQICAKLGPYEQICDQLVLMELPDIIDQIIAKEPPAIVCSQ VKICNGSAMAVAA
AP01343	TESYFVFSVGM
AP01345	FFGTALKIAANVLPTAICKILKKC
AP01346	FFPLVLGALGSILPKIF
AP01347	FIITGLVRGLTKLF
AP01348	SLSRFLSFLKIVYPPAF
AP01353	FWGHIWNAVKRVGANALHGAVTGALS
AP01354	GFWKKVGSAAWGGVKAAAKGAAVGGLNALAKHIQ
AP01355	RESPSSRMECYEQAERYGYGGYGGGRYGGGYGSGRGQPVGQGVERSHDDNRNQPR
AP01358	VTCDLLSFEAKGFAANHSLCAAHCLAIGRRGGSCERGVCICRR
AP01363	ATCDLASGFGVGSSLCAAHCIARRYRGGYCNSKAVCVCRN
AP01365	AAKPMGITCDLLSLWKVGHAACAAHCLVLGDVGGYCTKEGLCVCKE
AP01366	ATCDLLSMWNVNHSACAAHCLLLGKSGGRCNDDAVCVCRK
AP01367	VTCNIGEWVCVAHCNSKSKKSGYCSRGVCYCTN
AP01368	ATCDLFSFRSKWVTPNHAACAAHCLLRGNRGGRCKGTICHCRK
AP01371	GVTITVKPPFPGCVFYECIANCRSRGYKNGGYCTINGCQCLR
AP01372	SKCKCSRKGPKIRYSDVKKLEMKPKYPHCEEKMVIITTKSVSRYRGQEHCLHPKLQSTKRFIKWYNA WNEKRRVYEE
AP01374	NLAKGKEESLDSDLYAELRCMCIKTTSGIHPKNIQSLEVIGKGTHCNQVEVIATLKDGRKICLDPDAPRI KKIVQKKLAGDES
AP01375	FDNPFGCPADEGKCFDHCNNKAYDIGYCGGSYRATCVCYRK
AP01378	AREASKSLIGTASCTCRRAWICRWGERHSGKCIDQKGSTYRLCCRR

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01379	ILENLLARSTNEDREGSIFDTGPIRRPKPRPRPRPEG
AP01380	YDLSKNCRLRGGICYIGKCPRRFFRSGSCSRGNVCCLRFG
AP01381	DDTPSSRCGSGGWGPCLPIVDLLCIVHVTVGCSGGFGCCRIG
AP01382	QKKCPGRCTLKCGKHERPTLPYNCGKYICCVPVKVK
AP01383	GLVSGLLNTAGGLLGDLLGSLGSLSGGES
AP01384	GMWGSLLKGVATVVKHVLPHALSSQQS
AP01386	LLGDLLGQTSKLVNDLTDTVGSIV
AP01388	GLLSGILNSAGGLLGNLIGSLSN
AP01396	GLMDSLKGLAATAGKTVLQGLLKTASCKLEKTC
AP01397	ILPFLAGLFSKIL
AP01398	YQEPVLGPVRGPFPIIV
AP01400	RPKHPIKHQGLPQEVLNENLLRF
AP01405	GLLGGLLGPLLGGGGGGGGLL
AP01407	HNKQEGRDHDKSKGHFHRVVIHHKGGKAH
AP01423	FLPAVLRVAAKIVPTVFCAISKKC
AP01425	GLLGSLFGAGKKVACALSGLC
AP01428	GFKGAFKNVMFGIAKSAGKSALNALACKIDKSC
AP01429	GLLDSFKNAMIGIAKSAGKTALNKIACKIDKTC
AP01432	FMGGLIKAATKIVPAAYCAITKKC
AP01434	FFGSVLKLIPKIL
AP01442	GLFLNTVKDVAKDVAKDVAGKLLESLKCKITGCKS
AP01445	FMGSALRIAAKVLPAALCQIFKKC
AP01447	FLPGLIAGIAKML
AP01448	FLPIALKALGSIFPKIL
AP01449	FLGAIAAALPHVINAVTNAL
AP01454	IIPLPLGYFAKKT
AP01455	FFPLALLCKVFKKC
AP01456	VGKTWIKVIRGIGKSKIKWQ
AP01457	GLKDIFKAGLGSLVKGIAAHVAN
AP01461	ILGKLLSTAAGLLSNL
AP01462	ILGAILPLVSGLLSNKL
AP01464	VNPSYRLDPESRPQCEAHCGQLGMRLGAIVIMGTATGCVCEPKEAATPESR
AP01465	VNWKKVLGKIIKVAK
AP01470	AQRCGDQARGAKCPNCLCCGKYGFCGSGDAYCGAGSCQSQCRGCR
AP01471	RPKPQQFFGLM
AP01472	QLYENKPRRPYIL
AP01474	YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY
AP01475	ECWMDGHCRLLCKDGEDSIIRCRNRKRCC
AP01476	ACDTATCVTHRLAGLLSRSGGVVKNNFVPTNVGSKAF
AP01477	HSDAVFTDNYTRLRKQMAVKKYLNSILN
AP01479	YRQSMNNFQGLRSFGCRFGTCTVQKLAHQIYQFTDKDKDNVAPRSKISPQGY
AP01484	QYGYGPMMGGYGPGMMGGYGPGMMGGYGPGMMGGYGPGMMGGYGMSPMYGGYGMYRPGLLG MLLG
AP01486	QWGYNSYGGYNSYGNYGGYGGGYNNGYGVNANLGVGGRGG
AP01492	QWGYGGPYGGYGGGYGGGPWGYGGGWRRRHWGGYGGGPWGGYGGGPWGGYYGK
AP01493	ASIIKTTIKVSKAVCKTLTCICTGSCSNCK
AP01494	GHHPHGHHPHGHHPHGHHHPH
AP01496	IPPFIKKVLTTVF
AP01497	FLPIVGRLISGIL
AP01499	FLPVLARLAVKFLPSIVCAATKKC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01509	FLPKMSTKLRVPYRRGTKDYH
AP01510	GILKKFMLHRGTKVYKMRTLSKRSH
AP01511	TITLSTCAILSKPLGNNGYLCTVTKECMPSSCN
AP01513	GKNPTLQCMGNRGFCRPSCKKGEQAYFYCRTYQICCLQSHVRISLTGVEDNTNWSYEKHWPRIP
AP01514	GVNMYIRQIYDTCWKLKGHCRNVCGKKEIFHIFCGTQFLCCIERKEMPVLFVK
AP01515	AACSDRAHGHICESFKSFCKDSGRNGVKLRANCKKTCGLC
AP01516	LNLKGIFKKVASLLT
AP01517	INLLKIAKGIIKSL
AP01520	SSFSPPRGPPGWGPPCVQQPCPKCPYDDYKCPTCDKFPECEECPHISIGCECGYFSCECPKPVCEPCE SPIAELIKKGGYKG
AP01521	RFRLPFRRPPIRIHPPPFYPPFRRFL
AP01522	TYMPVEEGEYIVNISYADQPKKNSPFTAKKQPGPKVDLSGVKAYGPG
AP01523	AVDFSSCARMDVPGLSKVAQGLCISSCKFQNCGTGHCEKRGGRPTCVCDRCGRGGGEWPSVPMPKG RSSRGRRHS
AP01528	RVCMKGSQHHSFPCISDRLCSNECVKEEGGWTAGYCHLRYCRCQKAC
AP01529	GAARKSIRLHRLYTWKATIYTR
AP01530	GSCSCSGTISPYGLRTCRATKTKPSHPTTKETHPQTLPT
AP01531	GKWGWIYITILFADVGGFKSSRHPEERRVQERRFKRITRGPD
AP01533	KRRGSVTTRYQFLMIHLLRPKKLFA
AP01539	SILSGNFGVGKKIVCGLSGLC
AP01540	AGANDLCQECEDIVHLLTKMTKEDAFQDTIRKFLEQECDILPLKLLVPRCRQVLDVYLPLVIDYFQGQI KPKAICSHVGLC
AP01541	AVLDILKDVGKGLLSHFMEKV
AP01542	AVLDFIKAAGKGLVTNIMEKVG
AP01543	KPWRFRRAIRRVRWRKVAPYIPFVVKTVGKK
AP01544	IFGAIAGLLKNIF
AP01545	FFGHLFKLATKIIPSLFQ
AP01547	GVIKSVLKGVAKTVALGML
AP01548	ADTLACRQSHQSCSFVACRAPSVDIGTCRGGKLKCCKWAPSS
AP01549	VLLFLFQAAPGSADAPFADTAACRSQGNFCRAGACPPTFAASGSCHGGLLNCCAK
AP01550	SAVGRHGRRFGLRKHRKH
AP01552	QIVDCWETWSRCTKWSQGGTGTLWKSCNDRCKELGRKRGQCEEKPSRCPLSKKAWTCICY
AP01553	RMCKTPSGKFKGYCVNNTNCKNVCRTEGFPTGSCDFHVAGRKCYCYKPCP
AP01555	TCRYWCKTPENQTYCCEDEREIPSKVGLKPGKCPPVRPVCPPTRGFFEPPKTCSNDGSCYGADKCCF DRCLGEHVCKPIQTRG
AP01556	GCFEDWSRCSPSTSRGTGVLWRDCDSYCKVCFKADRGECFDSPSLNCPQRLPNNKQCRCINARTAKD NRNPTCWA
AP01557	DHHHDHGHDDHEHEELTLEKIKEKIKDYADKTPVDQLTERVQAGRDYLLGKGARPSHLPARVDRHLS KLTAAEKQELADYLLTFLH
AP01558	LGAWLAGKVAGTVATYAWNRYV
AP01560	AKYTGKCTKSKNECKYKNDAGKDTFIKCPKFDNKKCTKDNNKCTVDTYNNAVDCD
AP01562	LSKFGGECSLKHNTCTYLKGGKNHVVNCGSAANKKCKSDRHHCEYDEHHKRVDCQTPV
AP01564	ATCDLLSKWNWNHTACAGHCIAKGFKGGYCNDKAVCVCRN
AP01565	DDMTMKPTPPPQYPLNLQGGGGGGSGDGFGFAVQGHQKVWTSDNGRHEIGLNGGYGQHLGGPYGN SEPSWKVGSTYTYRFPNF
AP01566	QRPYTQPLIYYPPPPTPPRIYRA
AP01569	SNDSLWYGVGQEMGKQANCITNHPVKHMIIPGYCSKILG
AP01570	GNAACVIGCIGSCVISEGIGSLVGTAFTLG
AP01571	IFGSIYHRKCVVKNRCETVSGHKTCKDLTCCRAVIFRHERPEVCRPQT
AP01572	WNPFKKIANRNCYPKTTCETAGGKKTCKDFSCCQIVLFGKKTRAKCTVVTS
AP01573	GWFKKTFHKVSHAVKSGIHAGQRGCSALGF
AP01574	SWFSRTVHNVGNAVRKGIHAGQGVCSGLGL

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01575	NLPIVERPVCKDSTRIRITDNMFCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGE GCDRDGKYGFYTHVFRLKKWIQKVIDQFGE
AP01576	RVPPYLGRDCKHWCRDNNQALYCCGPPGITYPPFIRKHPGKCPSVRSTCTGVRSSRPKFCPHDDACE FRSKCCYDACVKHHVCKTVEFY
AP01577	FLLFPLMCKIQGKC
AP01578	GIHDILKYGKPS
AP01579	FVLPLVMCKILRKC
AP01580	AQEPVKGPVSTKPGSCPIILIRCAMLNPPNRCLKDTDCPGIKKCCEGSCGMACFVPQ
AP01583	GWANTLKNVAGGLCKITGAA
AP01585	KSCCRSTQARNIYNAPRFAGGSRPLCALGSGCKIVDDKKTPPND
AP01586	KSCCRSTTARNIYNGCRVPGTARPVCAKKSGCKIQEAKKCEPPYD
AP01587	AQCGAQGGGATCPGGLCCSQWGWCGSTPKYCGAGCQSNCK
AP01589	KDRPKKPGLCPPRPQKPCVKECKNDDSCPGQQKCCNYGCKDECRDPIFVG
AP01590	DHYICAKKGGTCNFSPCPLFNRIEGTCYSGKAKCCIR
AP01591	KCWNLRGSCREKCIKNEKLYIFCTSGKLCCLKPKFQPNMLQR
AP01592	GIRNTVCFMQRGHCRLFMCRSGERKGDICSDPWNRCCVSSSIKNR
AP01593	CKQSCSFGPFTFVCDGNTK
AP01595	CANSCSYGPLTWSCDGNTK
AP01599	WNPFRKLYRKECNDVTSCDTVSGVKTCTKKNCCHRKFFGKTILKAPECTVIS
AP01600	RARAPHKAWYNCMTDAGISGAIAGAVAGCAATIEIGCVEGAIAGIGPSGIASMIAALWTCRSKY
AP01601	YVPPVQKPHPNGPKFPTFP
AP01603	SSSGWLCTLTIECGTIICACR
AP01604	DAPGHPGKHYLQVNVPSDVRTIGVAGGGVQQCFRVTPGAWNDTRALVSNGAQVEVWGYTVADCAN RTTANQKYYDKAAAPSDSSTYFWFTLKNLRV
AP01606	GLGKAQCAALWLQCASGGTIGCGGGAVACQNYRQFCR
AP01607	ADRGWIKTLTKDCPNVISSICAGTIITACKNCA
AP01609	KCKWWNISCDLGNNGHVCTLSHECQVSCN
AP01612	SASIVKTTIKASKKLCRGFTLTCGCHFTGKK
AP01613	LPRDTSRCVGYHGYCIRSKVCPKPFAAFGTCSWRQKTCCVDTTSDFHTCQDKGGHCVSPKIRCLEEQ LGLCPLKRWTCCKEI
AP01614	WRSLGRTLLRLSHALKPLARRSGW
AP01615	SASVLKTSIKVSKKYCKGVTLTCGCNITGGK
AP01616	SLGPAIKATRQVCPKATRFVTVSCKKSDCQ
AP01618	GTTVVNSTFSIVLGNKGYICTVTVECMRNCSK
AP01619	AANFGPSVFTPEVHETWQKFLNVVVAALGKQYH
AP01620	VDKPPYLPRPPPPRRIYNNR
AP01621	CAWYNISCRLGNKGAYCTLTVECMPSCN
AP01622	GLGSLLGKAFKIGLKTVGKMMGGAPREQ
AP01623	GFKLKGMARISCLPNGQWSNFPPKCIRECAMVSS
AP01624	HAEHKVKIGVEQKYGQFPQGTEVTYTCSGNYFLM
AP01625	LQDAALGWGRRCPQCPRCPSCPSCPRCPRCPRCKCNPK
AP01632	ATPATPTVAQFVIQGSTICLVC
AP01633	RRWVRRVRRWVRRVVRVVRRWVRR
AP01634	INWKKIFEKVKNLV
AP01637	INWKKIASIGKEVLKAL
AP01638	INWKKIAEVGGKILSSL
AP01641	IDWLKLGKMVMDVL
AP01642	LCLDQKPEMEPFRKDAQQALEPSRQRRWLHRRCLSGRGFCRAICSIFEEPVRGNIDCYFGYNCCRRMF SHYRTS
AP01644	GAFGNFLKGVAKKAGLKILSIAQCKLSGTC
AP01647	RCTCTTIISSSSTF

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01648	GKLNLFLSRLEILKLFVGAL
AP01650	YKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCCKKGSG
AP01651	LVATGMAAGVAKTIVNAVSAGMDIATALSLFSGAFTAAGGIMALIKKYAQKKLWKQLIAA
AP01652	LIDHLGAPRWAVDTILGAIAVGNLASWVLALVPGPGWAVKAGLATAAAIVKHQGKAAAAAW
AP01658	NALSMPRNKCNRALMCFG
AP01660	WNSNRRFRVGRPPVVGRPGCVCFRAPCPCSNY
AP01663	RRTCRCRFGRCFRRESYSGSCNINGRIFSLCCR
AP01667	RRICRCRIGRCLGLEVYFGVCFLHGRLARRCCR
AP01676	DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA
AP01678	ITCQQVTSELGPCVPYLTGQGIP
AP01679	GRILSFIKGLAEHL
AP01680	ILGIITSLLKSLGKK
AP01681	KDLHTVVSAILQAL
AP01683	NEMGGPLVVEARTCESQSHKFKGTCLSDTNCANVCHSERFSGGKCRGFRRRCFCTTHC
AP01686	KTCENLSGTFKGPCIPDGNCNKHCRNNEHLLSGRCRDDFRCWCTNRC
AP01695	GFGSLLGKALRLGANVL
AP01696	SPAGCRFCCGCCPNMRGCGVCCRF
AP01699	GRGREFMSNLKEKLSGVKEKMKNS
AP01700	VKLIQIRIWIQYVTVLQMFSMKTKQ
AP01701	MKTFSVAVAVAIVLAFICTQESSALPVTGVEELVELVSSDDPVADHQELPVELGERLFNIRKKRASPKC TPYCYPTRDGVFCGVRCDF
AP01708	SFLTTFKDLAIKAAKSAGQSVLSTLSCKLSNTC
AP01710	GIFSTVFKAGKGIVCGLTGLC
AP01713	SRSGRGSGKGGRGGSRGSSGSRGSKGPSGSRGSSGSRGSKGSRGGRSGRGSTIAGNGNRNNGGTRTA
AP01715	PSCVCSGFETSGIHFC
AP01718	FKVQNQHGQVVKIFHH
AP01719	GILDTFKGVAKGVAKDLAVHMLENLKCKMTGC
AP01724	GTPGFQTPDARVISRFGFN
AP01729	GSQLVYREWVGHSNVIKGPP
AP01739	GIGGVLLGAGKATLKGLAKVLAEKYAN
AP01743	GIGGALLSVGKLALKGLANVLADKFAN
AP01745	ERILDLRKTKKSCKNGEVLGCVSGHGPPGCSENECGMGPRPKACFFDCHYGCWCTGKLYRRKRDRK CVPKHECLL
AP01746	FLGGILNTITGLL
AP01747	SFPFFPPGICKRLKRC
AP01748	SFHVFPPWMCKSLKKC
AP01749	LVQRGRFGRFLKKVRRFIPKVIIAAQIGSRFG
AP01752	VTCELLMFGGVVGDSACAANCLSMGKAGGSCNGGLCDCRKTTFKELWDKRFG
AP01754	GGYYCPFFQDKCHRHCRSFGRKAGYCGGFLKKTCICV
AP01756	PDPGQPWQVKAGRPPCYSIPCRKHDECRVGSCSRCNNGLWGDRTCR
AP01757	SPRVSRRYGRPFGGRPFVGGQFGGRPGCVCIRSPCPCANYG
AP01758	IPAMEPAARVKRSPGYGGCSPRWACGGYG
AP01759	RMRRSKSGKGSGGSKGSGSKGSKGSKGSKGSGSKGSGSKGGSRPGGGSSIAGGGSKGKGGTQTA
AP01760	GSGRGSCRSQCMRRHEDEPWRVQECVSQCRRRRGGGD
AP01762	SPPNQPSIMTFDYAKTNK
AP01763	SPPSEQLGKSFNF
AP01765	APPPGYAMESDSFS
AP01766	FPPPGESAVDMSFFYALSNP
AP01768	QLGELIQQGGQKIVEKIQKIGQRIRDFFSNLRPRQEA
AP01769	KSLRPRCWIKIKFRCKSLKF

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01771	ILPLLLGKVVCAITKKC
AP01772	KICERASGTWKGICIHSNDCNNQCVKWENAGSGSCHYQFPNYMCFCYFDC
AP01775	GEFLKCGESCVQGECYTPGCSCDWPICKKN
AP01778	DSMGAVKLAKLLIDKMKCEVTKAC
AP01783	FLPGVLRLVTKVGPAVVCAITRNC
AP01786	GKLQAFLAKMKEIAAQTL
AP01788	QEAQSVACTSYYCSKFCGSAGCSLYGCYLLHPGKICYCLHCSR
AP01789	HSHACTSYWCGKFCGTASCTHYLCRVLHPGKMCACVHCSR
AP01790	HPHVCTSYYCSKFCGTAGCTRYGCRNLHRGKLCFCLHCSR
AP01793	GWINEEKIQKKIDEKIGNNILGGMAKAVVHKLAKGEFQCVANIDTMGNCETHCQKTSGEKGFCHGTK CKCGKPLSY
AP01794	FVDLKKIANIINSIFGK
AP01795	QIINNPITCMTNGAICWGPCPTAFRQIGNCGHFKVRCCKIR
AP01796	ASFPWSCPSLSGVCRKVCLPTELFFGPLGCGKGFLCGVSHFL
AP01797	GLWNSIKIAGKKLFVNVLDKIRCKVAGGCKTSPDVE
AP01798	SPRPDDKKNQGSASVDVQNERGEGTKVDARVRQELWRSDDGRTRAQAYGHWDRTYGGRNHGERSY GGGMRIEHTWGN
AP01799	KRGFGKKLRKRLKKFRNSIKKRLKNFNVVIPIPLPG
AP01800	KRGLWESLKRKATKLGDDIRNTLRNFKIKFPVPRQG
AP01801	RTKRRIKLIKNGVKKVKDILKNNNIIILPGSNEK
AP01802	RPWAGNGSVHRYTVLSPRLKTQ
AP01803	LMCTHPLDCSN
AP01804	GIRCPKSWKCKAFKQRVLKRLLAMLRQHAF
AP01815	DFGCARGMIFVCMRRCARMYPGSTGYCQGFRCMCDTMIPIRRPPFIMG
AP01824	FLPKLFAKITKKNMAHIR
AP01842	FIFPKKNIINSLFGR
AP01849	TSRCIFYRRKKCS
AP01850	SFLSTFKELAINAAKNAGQSILHTLSCKLDKTC
AP01855	GLFSKFVGKGIKNFLIKGVKHIGKEVGMDVIRVGIDVAGCKIKGVC
AP01860	ATNIPFKVHFRCKAAFC
AP01876	GIFGKILGVGKKTLCELSGMC
AP01883	GILGNIVGMGKKVVCGLSGLC
AP01886	VVKCSYRLGSPDSQCN
AP01891	RFIYMKGFGKPRFGKR
AP01892	IPWKLPATFRPVERPFSKPFCRKD
AP01893	AAPRGGKGFFCKLFKDC
AP01895	FLGSLLGLVGKVVPTLFCKISKKC
AP01896	GLMSTLKDFGKTAAKEIAQSLLSTASCKLAKTC
AP01898	RRSRRGRGGGRRGGSGGRGGRGGGGRSGAGSSIAGVGSRGGGGGRHYA
AP01899	FLKPLFNAALKLLP
AP01900	FLPVLAGVLSRA
AP01901	GLASFLGKALKAGLKIGSHLLGGAPQQ
AP01906	GIGSLLAKAAKLGANLL
AP01911	SALVGCWTKSYPPNPCFGRG
AP01914	AAFRGCWTKNYSPKPCL
AP01915	GGSVPCGESCVFIPCITSLAGCSCKNKVCYYD
AP01916	GTRCGETCFVLPCWSAKFGCYCQKGFCYRN
AP01917	GLWDSIKNFGKTIALNVMDKIKCKIGGGCPP
AP01919	FTLKKSQLLLFFLGTINFSLCEEERNAEEERRDYPEEKDVEVEKR
AP01921	ILPIIGKILSTIF

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP01922	GMWSKILGHLIR
AP01923	GKWMSLLKHILK
AP01925	GLLDAIKDTAQNLFANVLDKIKCKFTKC
AP01927	GLFNVFKKVGKNVLKNVAGSLMDNLKCKVSGEC
AP01929	GIFALIKTAAKFVGKNLLKQAGKAGLEHLACKANNQC
AP01939	CVISAGWNHKIRCKLTGNC
AP01940	FKTWKRPPFQTSCWGIIKE
AP01941	CVHWQTNTARTSCIGP
AP01947	FLGPIIKIATGILPTAICKFLKKC
AP01948	SIRDKIKTIAIDLAKSAGTGVLKTLICKLDKSC
AP01952	FFPLLFGALSSHLPKLF
AP01953	FALGAVTKLLPSLLCMITRKC
AP01955	EYHLMNGANGYLTRVNGKTVYRVTKDPVSAVFGVISNCWGSAGAGFGPQH
AP01956	GFGMALKLLKKVL
AP01957	GTGLPMSERRKIMLMMR
AP01958	GLPRKILCAIAKKKGKCKGPLKLVCKC
AP01959	AILTTLANWARKFL
AP01963	ACQCPDAISGWTHTDYQCHGLENKMYRHVYAICMNGTQVYCRTEWGSSC
AP01964	IKLSPETKDNLKKVLKGAIKGAIAVAKMV
AP01965	LKIPGFVKDTLKKVAKGIFSAVAGAMTPS
AP01968	IKIPPIVKDTLKKVAKGVLSTIAGALST
AP01969	GPVGLLSSPGSLPPVGGAP
AP01970	EGPVGLADPDGPASAPLGAP
AP01971	VTSKSLCTPGCITGVLMCLTQNSCVSCNSCIRC
AP01972	STIVCVSLRICNWSLRFCPSFKVRCPM
AP01973	MLCKLSMFGAVLGVPACAIDCLPMGKTGGSCEGGVCGCRKLTFKILWDKKFG
AP01974	YGQSTHAVIYAQGYTYSSDWR
AP01975	KQIMTQFFNFARSPAVKD
AP01976	VTCDVLSFEAKGIAVNHSACALHCIALRKKGGSCQNGVCVCRN
AP01979	VAKCTEESGGKYFVFCCYKPTRICYMNEQKCESTCIGK
AP01981	GGKCTVDWGGQGGGRRLPSPLFCCYKPTRICYLNQETCETETCP
AP01982	ANKCIIDCMKVKTTCGDECKGAGFKTGGCALPPDIMKCCHNC
AP01986	GDACGETCFTGICFTAGCSCNPWPTCTRN
AP01987	GIPCAESCVWIPPCTITALMGCSCKNNVCYNN
AP01991	GEYCGESCYLIPCFTPGCYCVSRQCVNKN
AP01993	TNWKKIGKCYAGTLGSAVLGFGAMGPVGYWAGAGVGYASFC
AP01995	ECELAKVDGGYTPKNCAMAVGGGMLSGAIRGGMSGTVFGVGTGNLAGAFAGAHIGLVAGGLACIGG YLGSH
AP01996	EDGLHPRLCSC
AP01997	TPGGIDFISGGPHVAQDVLNAIKNFFK
AP02001	GMATKAGTALGKVAKAVIGAAL
AP02003	GFWTTAAEGLKKFAKAGLASILNPK
AP02006	GLLDALSGILGL
AP02007	GLLGTLGNLLNGLGL
AP02011	GLFDVIKKVASVIGLASP
AP02012	VKVGINGFGRIGRLVTRAAFHGKKVEVVAIND
AP02013	FIGKLISAASGLLSHL
AP02016	GKIPVKAIKQAGKVIGKGLRAINIAGTTHDVVSFFRPKKKKH
AP02018	HTPTPTPICKSRSHEYKGRCIQDMDCNAACVKESESYTGGFCNGRPPFKQCFCTKPCKRERAAATLR WPGL

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02020	FLSTLLKVAFKVVPTLFCPITKKC
AP02021	FFPIVGKRLYGLL
AP02022	FLPLFLPKIICVITKKC
AP02023	RYCLSQSHRFKGLCMSSSNCANVCQTENFPGGECKADGATRKCFCKKIC
AP02024	RHRHCFSQSHKFVGACLRESNCENVCKTEGFPSGECKWHGIVSKCHCKRIC
AP02025	DCYEDWSRCTPGTSFLTGILWKDCHSRCKELGHRGGRCVDSPSKHCPGVLKNNKQCHCY
AP02026	GFWGKLWEGVKSAI
AP02028	KRKCPKTPFDNTPGAWFAHLILGC
AP02029	DSIRDVSPTFNKIRRWFDGLFK
AP02030	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLV LRLR
AP02032	RWKPFKKELKVGRNIRDGIIKAGPAVAVIGQATSIARPTGK
AP02033	VSCGDVTSSIAPCLSYVMGRESSPSSSCCSGVRTLNGKASSSADRRTACSCLKNMASSFRNLNMGNAASI PSKCGVSVAFPISTSVDCSKIN
AP02035	DPQTECQQCQRRCRQQESGPRQQQYCQRRCKEICEEEEEYN
AP02036	RQRDPQQQYEQCQKHCQRRETEPRHMQTCQQRCERRYEKEKRKQQKRYEEQQREDEEKYEERMK EEDN
AP02037	KRDPQQREYEDCRRRCEQQEPRQQHQCQLRCREQQRQHGRGGDMMNPQRGGSGRYEEGEEEQS
AP02038	FLGLIFHGLVHAGKLIHGLIHRNRG
AP02040	KSCCRSTLGRNCYNLCRARGAQKLCAGVCRCKISSGLSCPKGFPK
AP02042	DRCSQQCQHHRDPDRKQQCMRECRRHQGRSD
AP02043	AVPLAVPLVYMGASVSPAVWNWLLVTFGAAAVAAAAVTVSDNDSHSCANNRGWCRSRCFSHEYIDS WHSDVCGSYDCCRPRY
AP02049	FFHHIFRGIVHVGKTIHKLVTGT
AP02052	TPALAVVTTVLPAAAVTTAKSV
AP02053	GLSQGVEPDIGQTYFEESRINQD
AP02064	GILDKLKEFGISAARGVAQSLLNTTASCKLAKTC
AP02065	TFPKCAPTRPPGPKPCDINNFKSKFWHIWRA
AP02066	SYFSAWAGPGCNNHNARYSKCGCSNIGHNVHGGYEFVYQGQTAAAYNTDNCKGVAQTRFSSSVNQA CSNFGWKSVFIQC
AP02068	YCERMMKRRSLTSPCKDVNTFIHGNKSNIKAICGANGSPYRENLRMSKSPFQVTTCKHTGGSPRPPCQ YRASAGFRHVVIACENGLPVHFDESFFSL
AP02075	SNFDCCLGYTDRILHPKFIVGFTRQLANEGCDINAIIFHTKKKLSVCANPKQTWVKYIVRLLSKKVKNM
AP02076	ASVATELRCQCLQTLQGIHPKNIQSVNVKSPGPHCAQTEVIATLKNGRKACLNPASPIVKKIIEKMLNS DKSN
AP02080	VPLSRTVRCTCISISNQPVNPRSLEKLEIIPASQFCPRVEIIATMKKKGEKRCLNPESKAIKNLLKAVSKE RSKRSP
AP02081	FPMFKRGRCLCIGPGVKAVKVADIEKASIMYPSNNCDKIEVIITLKENKGQRCLNPKSKQARLIIKKVE RKNF
AP02082	KPVSLSYRCPCRFFESHVARANVKHLKILNTPNCALQIVARLKNNNRQVCIDPKLKWIQEYLEKALNK
AP02083	VLEVYYTSLRCRCVQESSVFIPRRFIDRIQILPRGNGCPRKEIIVWKKNKSIVCVDPQAEWIQRMMEVL RKRSSSTLPVPVFKRKIP
AP02084	VGSEVSDKRTCVSLTTQRLPVSRIKTYTITEGSLRAVIFITKRGLKVCADPQATWVRDVVRSMDRKSN TRNNMIQTKPTGTQQSTNTAVTLTG
AP02085	KSMQVPFSRCCFSFAEQEIPLRAILCYRNTSSICSNEGLIFKLKRGKEACALDTVGWVQRHRKMLRHCP SKRK
AP02086	PDSVSIPITCCFNVINRKIPIQRLESYTRITNIQCPKEAVIFKTKRGKEVCADPKERWVRDSMKHLDQIF QNLKP
AP02087	GPASVPTTCCFNLANRKIPLQRLESYRRITSGKCPQKAVIFKTKLAKDICADPKKKWVQDSMKYLDQK SPTPKP
AP02088	QPDALNVPSTCCFTFSSKKISLQRLKSYVITTSRCPQKAVIFRTKLGKEICADPKEKWVQNYMKHLGR KAHTLKT
AP02089	ARGTNVGRECCLEYFKGAIPLRKLKTWYQTSEDCSRDAIVFVTVQGRAICSDPNNKRVKNAVKYLQS LERS
AP02090	AQVGTNKELCCLVYTSWQIPQKFIVDYSETSPQCPKPGVILLTKRGRQICADPNKKWVQKYISDLKLN A

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02091	GTNDAEDCCLSVTQKPIPGYIVRNFHYLLIKDGCRVPAVVFTTLRGRQLCAPPDQPWVERIIQRLQRT SAKMKRRSS
AP02094	GPYGANMEDSVCCRDYVRYRLPLRVVKHFYWTSDSCPRPGVVLLTFRDKEICADPRVPWVKMILNK LSQ
AP02096	KVHGSLARAGKVRGQTPKVAKQEKKKKKTGRAKRRMQYNRRFVNVVPTFGKKKGPNANS
AP02098	SGTSEKERESGRLLGVVKRLIVCFRSPFP
AP02104	MQFITDLIKKAVDFFKGLFGNK
AP02105	MAADIISTIGDLVKLIINTVKKFQK
AP02116	FFPTIAGLTKLFCAITKKC
AP02117	VLSIVACSSGCGSGKTAASCVETCGNRCFTNVGSLC
AP02118	FLPAALAGIGGILGKLF
AP02119	GFGCPGDAYQCSEHCRALGGGRTGGYCAGPWYLGHPTCTCSF
AP02120	YVPKIPKPQPNKPNFPSFPGHGPFNPHASRFPRSPKDNGKIVFDAKKEGGKTQWNVETQQKVWGNK HGSIHVSAGAGKQPGGKPQGQVGIGGSFSWGK
AP02121	GFGCPFNENECHAHCLSIGRKFGFCAGPLRATCTCGKQ
AP02122	CLRIGMRGRELMGGIGKTM
AP02123	NVTPATKPTPSKPGYCRVMDELILCPDPPLSKDLCKNDSDCPGAQKCCYRTCIMQCLPPIFRE
AP02127	IWSAIWSGIKGLL
AP02128	SLQPGAPNFPIPGQEKQEGWKFDPSLTRGEDGNTLGSINIHHTGPNHEVGANWDKVIRGPGKAKPTY SIHGSWRW
AP02130	GSGPTYCWNEANNPGGPNRCSNNKQCDGARTCSSSGFCQGTSRKPDPGPKGPTYCWDEAKNPGGP NRCSNSKQCDGARTCSSSGFCQGTAGHAAA
AP02135	FIHHIIGGLFSAGKAIHRLIRRRRR
AP02136	FIHHIIGWISHGVRAIHRAIHG
AP02137	FLHHIVGLIHHGLSLFGDRAD
AP02140	IWDAIFHGAKHFLHRLVNPGGKDAVKDVQQKQ
AP02141	LLRHVVKILEKYL
AP02142	GFLDIIKDTGKEFAVKILNNLKCKLAGGCPP
AP02146	QGWEAVAAAVASKIVGLWRNEKTELLGHECKFTVKPYLKRFQVYYKGRMWCPGWTAIRGEASTRS QSGVAGKTAKDFVRKAFQKGLISQQEANQWLSS
AP02148	FFLLFLQGAAGNSVLCRIRGGRCHVGSCHFPERHIGRCSGFQACCIRTWG
AP02149	QYEALTAAILTKLSKMWHSDTLNFLGHTCHVSRTPTVKRFKLYWKGKFWCPGWAPFSGTSRTKSRS GSAREATKSFVDQALQRRLITQQEADLWLKG
AP02150	YEALVTSILGKLTGLWHNDSVDFMGHICYFRRRPKIRRFKLYHEGKFWCPGWAPFEGRCKYCVVF
AP02153	GGWLDIVKAIVVPAARETIKTQEITLLDHYCTLSRSPYIKSLELHYRAEVTCPGWTIIRGRGSNHRNPT NSGKDALKDFMTQAVAAGLVTKEEAAPWLN
AP02154	YVDREINLFDHYCIISRSPHISRWELKWQATVTCPGWTPVKGKVRGYSNPLSAEREATRDFVQRIVQR GLVTRDEASEWL
AP02158	EAEEDGDLQCLCVKTTSQVRPRHITSLEVIKAGPHCPTAQLIATLKNGRKICLDLQAPLYKKIIKKLLES
AP02160	TDTNVIGECFDEWSRCHRQTRWWTKILFQSCENRCKCKVQLMGNCIKVPFKCFLWKQKRFMCECY GPISGTKPWYCGWEL
AP02161	FTCAISCDIKVNGKPCKGSGEKKCSGGWSCKFNVCVKV
AP02162	KIKIPWGKVKDFLVGGMKAV
AP02163	GFFGNTWKKIKGKADKIMLKKAVKIMVKKEGISKEEAQAKVDAMSKKQIRLYLLKYYGKKALQKASE KL
AP02165	ITSFIGCTPGCGKTGSFNSFCC
AP02169	AKISGPEETSELPEVVSEERVPATATEPMADLRHGVTREPISPASKDSLRDKFKEKLDKWFHRPNLLS KRD
AP02171	PGGPGSAPPATCRYWCRTPQGQAYCCEGVDEPEGPVGVKIGSCPRVRPQCPPVRTFGPPSPCSNDFK CFGSDKCCYDICLEQHVCKPLSFFG
AP02172	FFGSLLSLGSKLLPSVFKLFQRKKE
AP02173	QLPICGETCVLGGCYTPNCRCQYPICVR
AP02174	FLPFLIPALTSLISSL
AP02178	LRVRRTLQCSCRRVCRNTCSCIRLSRSTYAS

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02179	LNWGAILKHIIK
AP02180	LDVKKIICVACKIKPNPACKKICPK
AP02182	DRCTKRYGRCKRDCLESEKQIDICSLPRKICCTEKLYEEDDMF
AP02183	GNEGGGHGGHGGYGGYHHHGGGGGGYGGYHGGGGS
AP02184	APAHRSSTFPKWVTKTERGRQPLRS
AP02185	GPVSAVLTELRCTCLRVTLRVNPKTIGKLQVFPAGPQCSKVEVVASLKNGKQVCLDPEAPFLKKVIQK ILDSGNKKN
AP02187	PPSTACCTQLYRKPLSDKLLRKVIQVELQEADGDCHLQAFVLHLAQRSICIHPQNP
AP02193	YSKSLPLSVLNP
AP02196	KCNTATCATQRLANFLVHSSNNFGAILSSTNVGSNTY
AP02197	PAAAAQAVAGLAPVAAEQ
AP02198	KAYSMPRCKGGFRAVMCWL
AP02199	KAYSTPRCKGLFRALMCWL
AP02202	RKCNFLCKLKEKLRTVITSHIDKVLRPQG
AP02203	WNDTGKDADGSEY
AP02205	MVFAYAPTCARCKSIGARYCGYGYLNRKGVSCDGQTTINSCEDCKRKFGRCSDGFITECFL
AP02208	SPIEPKGEILHRFRRSFCDYNLCVVSCKDSGFIGGYCSELDLCSCTIGWQ
AP02213	GILGKLWEGFKSIV
AP02214	IFGAIWKGISSLL
AP02215	FLSTIWNGIKSLL
AP02216	FLGALWNVAKSVF
AP02217	FLSTLWNAAKSIF
AP02222	FIPLVSGLFSRLL
AP02223	VIPIVSGLLSSLL
AP02225	YDTGIQGWTCGSRGLCRKHCYAQEHTVGYHGCPRRYRCCALRF
AP02226	FCHLCEDLIKDGKEAGDVALDVWLDEEIGSRCKDFGVLASECFKELKVAEHDIWEAIDQEIPEDKTCK EAKLC
AP02227	NGIECEMCKMSVKIVVPMLGEDTESIKKAVDAECKKEFHSIPFGTQECKKFIDTKLDPIIHELENGTAP SDVCTKLGMC
AP02229	GKCSVLKKVACAAAIAGAVAACGGIDLPCVLAALKAAEGCASCFCEDHCHGVCKDLHLC
AP02230	PKRKAEGDAKGDKAKVKDEPQRRSARLSAKPAPPKPEPKPKKAPAKKGEKVPKGKKGKADAGKEG NNPAENGDAKTDQAQKAEGAGDAK
AP02231	RAIGGGLSSVGGGSSTIKY
AP02232	AATAKKGAKKADAPAKPKKATKPKSPKKAAKKAGAKKGVKRAGKKGAKKTTKAKK
AP02236	GYGDGCYSEDDLSVCCKKKFKVIGKCFKSVRECQNSGCKYH
AP02237	FFRLLFHGVHHGGGYLNAA
AP02238	GWKKWFTKGERLSQRHFA
AP02239	GFLGILFHGVHHGRKKALHMNSERRS
AP02241	KYALMKKIAELIPNLKSRQVK
AP02242	TWATIGKTIVQSVKKCRTFTCGCSLGSCSNCN
AP02244	FQSHSLPTPADERNLLQQIDCGTSCSARCRLSSRPRLCKRACGTCCARCNCVPSGTAGNLDECPCYAN MTTHGNKRKCP
AP02245	LEYKGECFTKDNTCKYKIDGKTYLAKCPSAANTKCEKDGNKCTYDSYNRKVKCDFRH
AP02247	MAGFLKVVQLLAKYGSKAVQWAWANKGKILDWLNAGQAIDWVVSKIKQILGIK
AP02248	FKKKKRNIGTFVFFAIALFCTVMFAYLLLTNQYVPIDYNVPRYA
AP02249	GLLSLLSLLGKLL
AP02250	MKTILRFVAGYDIASHKKKTGGYPWERGKA
AP02251	MWGRILAFVAKYGTKAVQWAWKNKWFLLSLGEAVFDYIRSIWGG
AP02253	MGAIAKLVAKFGWPFIKKFYKQIMQFIGQGWTIDQIEKWLKRH
AP02258	CARLNCVPKGTSGNTETCPCYASLHSCRKYG
AP02259	QYEALVASILGKLSGLWHSDTVDFMGHTCHIRRRPKFRKFKLYHEGKFWCPGWTHLEGNSRTKSRSG SARDAIKDFVYKALQNKLITENNAAAWLKG

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02260	RILTMTKRVKMPQLYKQIVCRLFKTC
AP02262	FLGGLLASLLGKI
AP02263	YPELQQDLIARLL
AP02264	FLSGILKLAFKIPSVLCAVLKNC
AP02265	AKAWGIPPHVIPQIVPVRIRPLCGNV
AP02266	GFWDSVKEGLKNAAVTILNKIKCKISECPPA
AP02267	FIPGLRRLFATVVPTVVCAINKLPPG
AP02268	GLLDSVKEGLKKVAGQLLDTLKCKISGCTPA
AP02270	GFFDRIKALTKNVTLELLNTITCKLPVTPP
AP02273	FITGLIGGLMKAL
AP02281	GVLDTLKNVAIGVAKGAGTGVLKALLCQLDKSC
AP02283	SLFGTFAKMALKGASKLIPHLLPSRQQ
AP02285	KLGFENFLVKALKTVMHVPTSPLL
AP02287	GLKEVAHSAKKFAKGFISGLTGS
AP02288	GWASSIGSILGKFAKGGAQAFLQPK
AP02289	GWLPTFGKILRKAMQLGPKLIQPI
AP02292	GLLSNVAGLLKQFAKGGVNAVLNPK
AP02293	GFMSKVANFAKKFAKGGVNAIMNQK
AP02294	FIGALLRPALKLLAGK
AP02296	ILPIRSLIKKLL
AP02297	FLPLKKLRFGLL
AP02301	RISKKKGKGSWIKNGLIKGIKGLGKEISLDVIRTGIDIAGCKIKGEC
AP02303	SLWENFKNAGKQFILNILDKIRCRVAGGCRT
AP02304	FLAGLIGGLAKML
AP02306	PPCRGIFCRRVGSSSAIARPGKTLSTFITV
AP02307	GKCNVLCQLKQKLRSIGSGSHIGSVVLPRG
AP02308	GNGVVLTLTHECNLATWTKKLKCC
AP02311	ITIPPIVKDTLKKFFKGGIAGVMGKSQ
AP02317	ITIPPIVKNTLKKFIKGAVSALMS
AP02318	IKIPSFFRNILKKVGKEAVSLIAGALKQS
AP02319	GIFPIFAKLLGKVIKVASSLISKGRTE
AP02321	TNYGNGVGVPDAIMAGIIKLIFIFNIRQGYNFGKKAT
AP02322	MLWSASMRIFASAFSTRGLGTRMLMYCSLPSRCWRK
AP02323	KRKKHRCRVYNNGMPTGMYRWC
AP02324	VFHAYSARGNYYGNCPANWPSCRNNYKSAGGK
AP02328	GSVIKCGESCLLGKCYTPGCTCSRPICKKD
AP02329	SKWQHQQDSCRKQLQGVNLTPCEKHIMEKIQGRGDDDDDDDDD
AP02332	CETPSKHFNGLCIRSSNCASVCHGEHFTDGRCQGVRRRCMCLKPC
AP02335	ILSYLWNGIKSIF
AP02341	LVKDNPLDISPKQVQALCTDLVIRCMCCC
AP02342	DICTCCAGTKGCNTTSANGAFICEGQSDPKKPKACPLNCDPHIAYA
AP02343	IQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLLKNGERIEKVEHSDLSFSKDWSFYLLYYTEF TPTEKDEYACRVNHVTLSQPKIVKWDRDM
AP02346	GLEESPGHPGQPGPPGPPGAPGP
AP02347	KVTKSVKSIPVKI
AP02348	TTPLCVGVIIGLTTSIKICK
AP02351	QKIAEKFSGTRRG
AP02352	YPGPQAKEDSEGPSQGPASREK
AP02353	LPVNSPMNKGDTEVMKCIVEVISDTLSKPSPMPVSKECFETLRGDERILSILRHQNLLKELQDLALQGA KERTHQQ

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02354	KINNPVSCLRKGGRCWNRCIGNTRQIGSCGVPFLKCCKRK
AP02358	FRFGSFLKKVWKSKLAKKLRSKGKQLLKDYANKVLNGPEEEAAAPAE
AP02360	MVALLKSLERRRLMITISTMLQFGLFLIALIGLVIKLIELSNKK
AP02365	VIVKAIATLSKKLL
AP02370	FCKSLPLPLSVK
AP02371	GHLGRPYIGGGGGFNRGGGFHRGGGFHRGGGFHSGGGFHRGGGFHSGGSFGYR
AP02373	RRRRRRPPCEDVNGQCQPRGNPCLRLRGACPRGSRCCMPTVAAH
AP02374	GLVGTLLGHIGKAILG
AP02376	GLRRLFADQLVGRRNI
AP02384	AWLDKLKSLGKVVGKVALGVAQNYLNPQQ
AP02385	TKPTLLGLPLGAGPAAGPGKR
AP02386	KLSPSLGPVSKGKLLAGQR
AP02387	RLGTALPALLKTLLAGLNG
AP02388	RPDFCLEPPYTGPCKARMIRYFYNAKAGLCQPFVYGGCRAKRNNFKSSEDCMRTCGGA
AP02389	GKGLEVIKWKLKHVIQL
AP02390	LFAKINGLKVGPLKIQIV
AP02391	FSLFFPYAALKWLRKLLKK
AP02392	VKLEILGSKGGAKI
AP02393	VSKIKKYLKYKDRI
AP02394	DWTCWSCLVCAACSVELLNLVTAATGASTAS
AP02395	DWTFANWSCLVCDDCSVNLTV
AP02396	LASTLGISTAAAKKAIDIIDAASTIASIISLIGIVTGAGAISYAIVATAKTMIKKYGKKYAAAW
AP02397	CWSCMGHSCWSCMGHSCWSCAGHSCWSCMGHSCWSCMGHSCWSCAGHCCGSCWHGGM
AP02399	ESISVAGGTWNYGYGVGQAYSHYKHDYNNHGAKVVNSNNGVKDYKNAGPGVWAKASIGTVWDPAT FYYNPTGFYSN
AP02400	FAVWGCADYRGYCRAACFAFEYSLGPKGCTEGYVCCVPNTF
AP02401	VAPIAKYLATALAKWALKQGFAKLKS
AP02402	IGGALGNALNGLGTWANMMNGGGFVNQWQVYANKGKINQYRPY
AP02405	GGYKNFYGSALRKGFYEGEAGRAIRR
AP02406	TVKCGMNGKMPCKHGAFYTDTCDKNVFYRCVWGRPVKKACGRGLVWNPRGFCDYA
AP02407	SDYLNNNPLFPRYDIGNVELSTAYRSFANQKAPGRLNQNWALTADYTYR
AP02409	AIFIFIRWLLKLGHHGRAPP
AP02414	FLKGVINLASKIPSMLCAVLKTC
AP02415	GLFDSITQGLKDTAVKLLDKIKCKLSACPPA
AP02416	FIVPSIFLLKKAFCIALKKNC
AP02417	MKFFTLLAALMALFAICNNFSMVSASRDSRPVQPRVQPPPPPPKQKPSIYDTPIRRPGGQKTMYA
AP02420	GVLSAFKNALPGIMKIIV
AP02429	KTKQQFLIKAQTQLFKVFGYTL
AP02430	RLFRHAFKAVLRL
AP02432	APVPFSCTRGCLTHLV
AP02433	GDINGEFTTSPACVYSVMVVSKASSAKCAAGASAVSGAILSAIRC
AP02435	FWGAVWKILSKVLPHIPGTVKWLQEKV
AP02436	GTDSGRFCSSICGQRCSKAGMKDRCMKFCGICCGKCKCVPSGTYGNKHECPCYRDMKNSKGKPKCP
AP02437	FFGRLKSVWSAVKHGWKAAKSR
AP02439	SCTTCVCTCSCCTT
AP02440	QNCPTRRGLCVTSGLTACRNHCRSCHRGDVGCVRCSNAQCTGFLGTTCTCINPCPRC
AP02442	QVLEGLAAAVTGKLAGLWRNGEVELLGHYCSYSVTPTIRRWQLYFRGRMWCPGWTSIRGEAMTRSN SGVQGDTTRDFVTKALNAGLISQQEAQAWLDG
AP02443	NIFDDIFGKVTETLVDFGTTDIAGNPCNYRLSPRLIKFELYFVGLVWCPGWTTIQGESLTRSRTRVVN KAVEDFAKKAVAAGIMTQEDADPLLNA

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02444	GWLDRIIGTAVDSVAEFGTTNIVDQICNTRVMPTIKKFELYFRGRVWCPGWTTIQGESLTRSRTRVVN KAVEDFARKAVAAGLMTQEDANPLLNA
AP02446	DTFDYKKFGYRYDSLELEGRSISRIDELIQQRQEKDRTFAGFLLKGFGTSAS
AP02451	VVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGVIFTTKKGQQSCGDPKQEWVQRYMKNLDAKQ KKASPRARAVA
AP02452	TRGSDISKTCCFQYSHKPLPWTWVRSYEFTSNSCSQRAVIFTTKRGKKVCTHPRKKWVQKYISLLKTP KQL
AP02453	MDSFQKIEKIGEGTYGVVYKAKDKVSGRLVALKKIRLENESEGVPSTA
AP02455	AFGCPFDQGTCHSHCRSIRRRGRRCASFAKRTCTCYQK
AP02456	KKCKFFCKVKKKIKSIGFQIPIVSIPFK
AP02457	KKCGFFCKLKNKLKSTGSRSNIAAGTHGGTFRV
AP02458	FLPVLGKVIKLVGGLL
AP02461	FLPGLIKAAVGVGSTILCKITKKC
AP02466	SILSTLKDVGISAIKNAGSGVLKTLLCKLNKNCEK
AP02470	GFMDTAKNVAKNVAKNVAVTLLDKLRCKVTGGC
AP02476	SIMSTLKQFGISAIKGAAQNVLGVLSCKIAKTC
AP02477	GADRCRERCERRHRGDWQGKQRCLMECRRREQEED
AP02478	QSHISLCRWCCNCCKANKGCGFCCKF
AP02479	GMKCKFCCNCCNLNGCGVCCRF
AP02480	MTPLWRIMNSKPFGAYCQNNYECSTGLCRAGHCSTSHRATSETVNY
AP02494	FLFSLIPSVIAGLVSAIRN
AP02496	GSCGASIAEFNSSQILAKRAPPCRRPRLQNSEDVTHTTLP
AP02505	LTCNIDRSFCLAHCLLRGYKRGFCTVKKICVCRH
AP02506	GTCSFSSALCVVHCRVRGYPDGYCSRKGICTCRR
AP02507	FTCNSYACKAHCILQGHKSGSCARINLCKCQR
AP02509	HHHFGRIGHELHKGVKKVEKVTSDVNKVTNGVKQVANGIAKAKTVIEAGSIAGAVAAAAA
AP02510	HHLFGKVGREIERSAHKVGHKLEHVRHEVSKTAKKVDKVVGHIKTAKKVVAAAGAIAGVVAAA
AP02512	AGFRKRFNKLVKKVKHTIKETANVSKDVAIVAGSGVAVGAAMG
AP02514	GFGCPEDEYECHNHCKNSVGCRGGYCDAGTLRQRCTCYGCNQKGRSIQE
AP02515	FSTKTRNWFSEHFKKVKEKLKDTFA
AP02516	YSLKKTSMKIIPFTRL
AP02517	PPPVIKFNRPFLMWIVERDTRSILFMGKIVNPKAP
AP02518	ARDGYIVDEKGCKFACFIN
AP02519	MFAMKTKAALAIWCPGYSETQINATQAMKKRRKRKVTTNKCLEQVSQLQGLWRRFNRPLLKQQ
AP02520	LGSCVANKIKDEFFAMISISAIVKAAQKKAWKELAVTVLRFAKANGLKTNAIIVAGQLALWAVQCGLS
AP02521	MTPLWRVMGNKPFGAYCQDHVECSTGICKGGHCITSQPIKS
AP02523	KACPRNCDTDIAYMVCPSSGERIIRKVCTNCCAAQKGCKLFRSNGSIKCTGT
AP02524	GDCGGTCTWTKDCSICPSWSCWSWSC
AP02526	SNASVWECCSTGSWVPFTCC
AP02527	FLPLILPSIVTALSSFLKQG
AP02528	VSCDFEEANEDAVCQEHCLPKGYTYGICVSHTCSCIYIVELIKWYTNTYT
AP02529	GASPALWGCDSFLGYCRIACFAHEASVGQKDCAEGMICCLPNVF
AP02532	QGPGRQPDFQRCGQQLRNISPPQRCPSLRQAVQLTHQQQGQVGPQQVRQMYRVASNIPST
AP02533	SDKPDVKEVESFDKSKLKKVETQEKNPLPTKETIEQEKKG
AP02535	RRSKVRICSRGKNCVSRLGGGSIIGRPGGGSLIGRPGGGSVIGRPGGGSPPGGGSFNDEFIRDHSDGNRF A
AP02537	RRSKVRICSRGKNCVSRPGGGSFNDEFIRDHSDGNRFA
AP02539	RRGKDSGGPKMGRKDSKGCWRGRPGSGSRPGFGSGIAGASGVNHVGTLPASNSTTHPLDNCKISPQ
AP02543	GRADYNFGYGLGRGTRKFFNGIGRWVRKTF
AP02544	FCTCNVKGFNAKNKRGIIYP
AP02545	KDGYIIEHRGCKYSCFFGTNSWCNTECTLKKGSSGYCAWPACWCYGLPDNVKIFDSNNLKC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
AP02546	KDGYLVGNDGCKYSCFTRPGTYCANECSRVKGKDGYCYAWMACYCYSMPNWVKTWNRATNRCGR
AP02547	DNGYLLDKYTGCKVWCVINNESCNSECKIRGGYYGYCYFWKLACFCQGARKSELWNYNTNKCNGKL
AP02548	ATCDLLSPFKVGHAACAAHCIARGKRGGWCDKRAVCNCRK
AP02551	ILGEIWKGIKDIL
AP02553	ACGILHDNCVYVPAQNPCCRGLQCRYGKCLVQV
AP02554	RCLPAGKTCVRGPMRVPCCGSCSQNKCT
AP02557	ECIPTKHDCTNDRKNCCPGHECKCYNTQIGGSKKEQCGCKKSLLAKAKNFGGKVITIFKA
AP02559	TDDRCERMCQHYHDRREKKQCMKGCRYGESD
AP02561	GFREKHFQRFVKYAVPESTLRTVLQTVVHKVGKTQFGCPAYQGYCDDHCQDIEKKEGFCHGFKCKC GIPMGF
AP02562	GWMSEKKVQGILDKKLPEGIIRNAAKAIVHKMAKNQFGCFANVDVKGDCKRHCKAEDKEGICHGTK CKCGVPISYL
CAMPSQ5	FLSLLPSIVSGAVSLAKKLG
CAMPSQ10	FLPIPRPILLGLL
CAMPSQ11	FLIIRRPIVLGLL
CAMPSQ12	GLHKVMREVLGYERNSYKKFFLR
CAMPSQ16	INWKGIAAMKKLL
CAMPSQ27	KQATVGDINTERPGILDLKGKAKWDAWNGLKGTSKEDAMKAYINKVEELKKKYGI
CAMPSQ51	SDEKASPDRHHRFSLSRYAKLANRLSKWIGNRGNRLANPKLLETFKSV
CAMPSQ81	FLPLIGKILGTILGK
CAMPSQ83	FLPVILPVIGKLLNGILGK
CAMPSQ84	ISDYSIAMDKIRQQDFVNWLLAQKGKKSDWKHNITQ
CAMPSQ88	FLPLVRGAAKLIPSVVCAISKRC
CAMPSQ112	ACHAHCQSVGRRGGYCGNFRMTCYCY
CAMPSQ113	NCIQQCVSKGAQGGYCTNEKCTCY
CAMPSQ136	AGFAAQAAASLAPVAAQQL
CAMPSQ154	FLSLIPHAINAVSAIAKHFG
CAMPSQ169	AALRGALRAVARVGKAILPHVAIANPYVRTPYVHNNP
CAMPSQ171	KWKLFKKIGIGKFLHSAKKF
CAMPSQ175	INLKAIAALAKKLLG
CAMPSQ177	APIIRRIPYYPEVESDLRIVDCKRSEGFCQEYCNYLETQVGYCSKKKDACCLH
CAMPSQ190	FALALKALKKALKKLKKALKKAL
CAMPSQ276	DVQCGEGHFCHDXQTCCRASQGGXACCPYSQGVCCADQRHCCPVGF
CAMPSQ290	MTCGQVQGNLAQCIGFLQKGG
CAMPSQ299	GIGGKPVQTAFVDNDGIYD
CAMPSQ306	XTCESPSHKFKGPCATNRNCES
CAMPSQ311	QCVGTITLDQSDDLFDLNCNELQSVR
CAMPSQ331	MLTLKKSMLLLFFLGLVSVSLADDKREDEAEEGEDKRAAEEERNVEKRCYSAAKYPGFQEFINRKYKS SRFG
CAMPSQ353	GWKSVFRKAKKVGKTVGGLALDHYLG
CAMPSQ366	FIFHIIKGLFHAGKMIHGLVTRRRH
CAMPSQ463	RNNWQTNVGGAVGSAMIGATVGGTICGPACAVAGAHYLPILWTAVTAATGGFGKIRK
CAMPSQ474	GRSKKLGKKIEKAGKRVFNAAQKGLPVAAGVQAL
CAMPSQ476	NRWTNAYSAALGCAVPGVKYGKKLGGVWGAVIGGVGGAAVCGLAGYVRKG
CAMPSQ478	XXKEIXHIFHDN
CAMPSQ484	KNIGNSVSCLRNKGVCMPGKCAPKMKQIGTCGMPQVKCCKRK
CAMPSQ489	NNEAQCEQAGGICSKDHCFHLHTRAFGHCQRGVPCCRTVYD
CAMPSQ538	ATCDILSFQSQWVTPNHAGCALHCVIKGYKGGQCKITVCHCRR
CAMPSQ569	PAQPFRFPKHPQGPQTRPPI
CAMPSQ592	RFRPPIRRPPIRPPFNPPFRPPVRPPFRPPFRPPFRPPIGPFPGRR
CAMPSQ593	IGPDTKKCVQRKNACHYFECPWLYYSVGTCYKGKGKCCQKRY

Table A.13: Server Annotation Set Continued...

Definition	Sequence
CAMPSQ595	AKRGGFWRKVGRKLGKGIRKIGKTIKSQLGKFRPRLQYRYQF
CAMPSQ618	QGYKSGHTGPYPRPLYGSRPIGLRPITRPDPSCAGCRILTLDDAIACCRRLGRCCSALKG
CAMPSQ626	GSPEFGWLKKIGKKIERVGQHTRDATIQTIGVAQQAANVAATLKG
CAMPSQ649	IIGPVLGMVGSALGGLLKKIG
CAMPSQ654	ANDPQCLYGNVAAKF
CAMPSQ673	RSTEDIIKSISGGGFLNAMNA
CAMPSQ675	GFGCPFNQGACHRHCRSIRRRGGYCAGLIKQTCTCYRN
CAMPSQ677	LDTIKCLQGNNNCHIQKCPWFLLQVSTCYKGKGRCCQKRRWFARSHVYHV
CAMPSQ726	LKLLKKLLKLLKKLGGGK
CAMPSQ764	MSNLKWFSGGDDRRKKAEVIITELLDDLEMDLGNESLRKVLGSYLKKLKNEGTSVPLVLSRMNIEISNA IKKDGVSLNENQSKKLKELMSISNIRYGY
CAMPSQ793	GFFKKAWRKVKHAGRRVLDTAKGVGRHYVNNWLNRYRZ
CAMPSQ803	GLFSVLGSVAKHVVPRVVPVIAEHLG
CAMPSQ815	NIYWIADQFGIHLATGTARKLLDAVASGASLGTAFAAILGVTLPAWALAAAGALGATAA
CAMPSQ817	RNCESLSHRFKGPCTRDSNC
CAMPSQ818	KKWGWLAWVEPAGEFLKGFGKGAIKEGNKDKWKNI
CAMPSQ843	GLLDFAKHVIGIASKLG
CAMPSQ850	QAEESNLQSLVSQYFQTVADYGKDLVEKAKGSELQTQAKAYFEKTQEELTPFFKKAGTDLLNFLSSFI DPKKQPAT
CAMPSQ851	XXVPYPRPFPRPPIGPRPLPFPGGGRPFQS
CAMPSQ852	KTKLTEEEKNRLNFLKKISQRYQKFALPQYLKTVYQHQK
CAMPSQ882	ACLPNSCVSKGCCCGBSGYWCRQCGIKYTC
CAMPSQ888	DIDFSTCARMDVPILKKAAQGLCITSCSMQNCGTGSCKKRSGRPTCVCYRCANGGGDIPLGALIKRG
CAMPSQ898	FTCDVLGFEIAGTKLNSAACGAHCLALGRTGGYCNSKSVCVCR
CAMPSQ902	DEKPKLILPTPAPPNLPQLVGGGGGNRKDGFGVSVDAHQKVWTSDNGGHSIGVSPGYSQHLPGPYGN SRPDYRIGAGYSYNF
CAMPSQ907	SLQPGAPSFPMPGSQLPTSVSGNVEKQGRNTIATIDAQHKTDRYDVRGTWTKVVDGPGRSKPNFRIG GSYRW
CAMPSQ932	YLAFRCGRYSPCLDDGPNVNLYSCCSFYNCHKCLARLENCPKGLHYNAYLKVCDWPSKAGCTSVNKE CHLWKT
CAMPSQ941	QLINSPVTCMSYGGSCQRSCNGGFRLGGHCGHPKIRCCRRK
CAMPSQ956	QNICPRVNRIVTPCVAYGLGRAPIAPCCRALNDLRFVNTRNLRRAACRCLVGVVNRNPGLRRNPRFQ NIPRDCRNTFVRPFWWRPRIQCGRIN
CAMPSQ957	AITCGQVSSALGPCAAYAKGSGTSPSAGCCSGVKRLAGLARSTADKQATCRCLKSVAGAYNAGRAAG IPSRCGVSVPYTISASVDCSKIH
CAMPSQ961	AISCGAVTSDLSPCLTYLTGGPGPSPQCCGGVKKLLAAANTTPDRQAACNCLKSAAGSITKLNTNNAA ALPGKCGVDIPYKISTSTNCNTVKF
CAMPSQ976	ILAWKWAWWAWRX
CAMPSQ977	KNLRRITRKIIHIIKKYG
CAMPSQ981	RLFDKIRQVIRKF
CAMPSQ994	ISRLAGLLRKGGEKIGEKLKKIGQKIKNFFQKLVPQPE
CAMPSQ1028	GLFDKLKSLVSDF
CAMPSQ1030	KWKLFKKIGIGAVLKVLTTGLPALIS
CAMPSQ1065	VXLFPKKPFLXV
CAMPSQ1072	FHPSLWVLIPQYIQLIRKILKSG
CAMPSQ1083	GGAGHVPEYFVGIGTPISFYG
CAMPSQ1092	SCNCVCGFCCSCSP
CAMPSQ1098	PIRNWWIRIWEWLNGIRKRLRQRSPFYVRGHLNVTSTPQP
CAMPSQ1110	FIGGIISFIKKLF
CAMPSQ1113	LFGFLIPLLPHIIGAIPQVIGAIR
CAMPSQ1114	LFGFLIKLIPSLFGALSNIGRNRNQ
CAMPSQ1117	WFYQGMNIAIYANIGGVANIIGYTEAAVATLLGAVVAVAPVVP
CAMPSQ1118	VIEPKCYKYEGKKCPPDINPVCGTDKRTYYNECALCVFIRQSTKKADKAIKIKKWGKC

Table A.13: Server Annotation Set Continued...

Definition	Sequence
CAMPSQ1119	NWRKILGKIAKVAAGLLGSMLAGYQV
CAMPSQ1120	WWRELLKKLAFTAAGHLGSVLAAKQSGW
CAMPSQ1125	TTKNYGNGVCNSVNWCQCGNVWASCNLATGCAAWLCKLA
CAMPSQ1131	FSDSQLCRNNHGHCRRLCFHMESWAGSCMNGRLRCCRFSTKQPFSNPKHSVLHTAEQDPSPSLGGT
CAMPSQ1139	RPEIKKKNVFSKPGYCPEYRVPCPFVLIPKCRRDKGCKDALKCCFFYCQMRCVDPWESPE
CAMPSQ1140	GGVKGEEKRVCPPDYVRCIRQDDPQCYSDNDCGDQEICCFWQCGFKCVLPVKDNSEEQIPQSKV
CAMPSQ1143	YTAKQCLQAIGSCGIAGTGAGAAGGPAGAFVGAXVVXI
CAMPSQ1144	DLHIPPPDNKINWPQLSGGGGGSPKTGYDININAQQK
CAMPSQ1148	ADDRNPLEQCFRETDYEEFLEIARNNLKATSNPKHVVIVGAGMAGLSAAYVLSGGGHQVTV
CAMPSQ1175	IIGAIAAALPHVINAIKNTFG
CAMPSQ1180	FALLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
CAMPSQ1184	ALYKKFKKKLLKSLKRLG
CAMPSQ1194	SNMIEGVFAKGFKKASHLFKGIG
CAMPSQ2438	DSKYLFVFLIFNVIVIDLCQGFLWGLIPGAISAVTSLIKKGRRRRELGSQYDYLQDFRKRELDLDDLLSK FPDY
CAMPSQ2521	SRYARMRDSRPWSDRKNNYSGPQFTYPPEKAPPEKLIKWNNEGSPIFEMPAEGGHIEP
CAMPSQ2522	FFLPSLIGGLISAIK
CAMPSQ2523	FGLIPSMMGGLVSAFK
CAMPSQ2530	CPQTPRCTNYAEKGQCPPN
CAMPSQ2658	ISCKAGRVGCFASCQVQNCATGYCRGSTCVCSRCGKGTTPFNKFKIWNQLRVLVQKMVDEERA
CAMPSQ2720	IKELLPHLSGIIDSVANAIK
CAMPSQ2721	CTTCECCCSCS
CAMPSQ2728	TPATTSSWTCITAGVTVSASLCPTTKCTSRC
CAMPSQ2737	KFEPPLPPKKAHKKFWEDDGIYYPPNHNFP
CAMPSQ2798	SYKKIDCGGACAARC
CAMPSQ2811	IRPRPTARPPYINRPPNPFKPRW
CAMPSQ2866	LMDIFKVAVNKLLAAGMNKPRCKAAHC
CAMPSQ2894	CADLRGKTFCRLFKSYCDKKGIRGRLMRDKCSYSCGCRG
CAMPSQ3193	SMWSGMWRRKLKKLRNALKKKLKGEK
CAMPSQ3295	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRVCRTGRSRWRDVCRNFMRRYQSRVTQGLVAGET AQQICEDLRLCIPSTGPL
CAMPSQ3379	MTRILPCLFLVLLAAAPLLANPANPLNLKKHHGVF
CAMPSQ3383	GRRKRQMEARFEPQNRNYRKRELDLEKLFANMPDY
CAMPSQ3392	DAECEICKFVIQQVEAFIESNHSQAEIQKELNKLCSSVPSIFTQTCLSIARMVPYIIKKLEEHNSPGQVCQ GLHLCKSS
CAMPSQ3704	GIWSSIKNLASKAWNSDIGQSLRNKAAGAINKFVADKIGVTPSQAASMTLDEIVDAMYYD
CAMPSQ3903	FLFSLIPSAISGLISAFKGRR
CAMPSQ3904	FIGAIARLLSKIFGKR
CAMPSQ4070	VRLEFKLQQTSCRKRDWKKP
CAMPSQ4092	TSYGNGVHCNKSKCWIDVSELETYKAGTVSNPKDILWSLKE
CAMPSQ4094	FVYGNGVTSILVQAQFLVNGQRRFFYTPDK
CAMPSQ4109	YDNVNLDEILANDRLLVPYIKCLLDEGKKAPDAKELKEHIRXAL
CAMPSQ4113	GVTFLGHQCLVESTRQPDGFYTAKMSCASWTHDNPIVGEGRSRVELEALKGSITNFVQTASNYKKFTI DEVEDWIASY
CAMPSQ4114	DTHISEKIIDCNDIG
CAMPSQ4127	TTPATPAISILSAYISTNTCPTTKCTRAC
CAMPSQ4129	ASGGTVGXYGAWMRSXSLVSXSTITTFS
CAMPSQ4131	KLTFIQSTAAGDLYYNTNTHKYVYQQTQNAFGAAANTIVNGWMGGAAGGFGLHH
CAMPSQ4142	SGPVPSGCLRCICVVESGXRMPNPV
CAMPSQ4150	GCLEFWWKCNPNDDKCCRPKLKCSKLFKLCNFSF
CAMPSQ4160	AVPAVRKTNETLD

Table A.13: Server Annotation Set Continued...

Definition	Sequence
CAMPSQ4161	RITRFPTGNDCKEVNTFIQANGNHVRTVCTGGGTRQTDNRDLYMSNNQFTVITCTLRSGERHPNCRY RGKESSRKIVVACEGEWPTHYEKGVIV
CAMPSQ4168	AVPDVAFNAYG
CAMPSQ4181	MISTSSILVLVVLLACFMAASAQWGYGGYGRGYGGYGGYGRGMYGGYGRGMYGGYGRGMYGGYG RGMYGGWGK
CAMPSQ4182	AIPIAYVGMAVAPQVFRWLVRAYGAAAVTAAGVTLRRVINRSRSNDNHSCYGNRGWCRSSCRSYERE YRGGNLGVCGSYKCCVT
CAMPSQ4190	GFLSALKKYLPIVLKHV
CAMPSQ4269	ARDAYIANDRNCVYTCALNPYCDSECKKNGADSGYCQWFGRFGNACWCKNLPDKVPIRIPGECR
CAMPSQ4270	FFGHLYRGITSVVKHVHGLLSGETPRQQEVMKEAMREAMKVQEAMDQEAFDRERALV
CAMPSQ4388	EPEPSYVGDCGSNGGSCVSSYCPYGNRLNYFCPLGRTCCRHAYV
CAMPSQ4394	LYKLVKVVLNM
CAMPSQ4401	GVCRXVCRRGVCRXVCRR
CAMPSQ4468	QKIRTRRSKARGGSRGSKMGRKDSKGGSRGRPGSGSRPGGGSSIAGASRGDRGGTRNA
CAMPSQ4667	MKCVMIFLVLTLVVLMAEPGEGFFRHLFRGAKAIFRGARQGWRAHKVVSRYRNRDVPETDNNQEEP YNQR
CAMPSQ4668	MKVLAACVFLLLLVLLHGSPAACSALRAQANIGLMPRPETGAQSHGLEAAAGLMPHPEIGAQSLEVPL RRSKRFNSHFPICSYCCNCCRNKGCGLCCRT
CAMPSQ4677	FSCDVLSFQSKWVSPNHSACAVRCLAQRRKGGKCKNGDCVCR
CAMPSQ4679	MRFLCLVFAVLLLVSLAAPGYGLVLKYCPKIGYCSNTCSKTQIWATSHGCKMYCCLPASWKWK
CAMPSQ4728	IAPIIVAGLGYLVKDAWDHSDQIISGFKKGWNGGRRK
CAMPSQ4922	ACQFWSCNSSCISRGYRQGYCWGIQYKYCQCQ
CAMPSQ4938	PVVDTTGNNPLQQQEEYYV
CAMPSQ5006	FDVMGIIKKIAGAL
CAMPSQ5007	FDIMGLIKKVAGAL
CAMPSQ5038	MRILYLLLSVLFVVLQGVAGQPYFSSPIHACRYQRGVCIPGPCRWPYYRVGSCGSGLKSCCVRNRWA
CAMPSQ5039	MRIVYLLFPFILLLVQGAAGSSLAPRNKEKCHREKGFCGFLKCSFPFIISGKCSRFFFCCKKIFG
CAMPSQ5040	MQILPLLFAVLLLMLQAEPGLSLARGLPQDCERRGGFCSHKSCPPGIGRIGLCSKEDFCCRSRWYS
CAMPSQ5045	MKTQFAIFLITLVLFQMFSQSDAIFKAIWSGIKSLFGKRGLSDLDDLDESFDGEVSQADIDFLKELMQ

Chapter B: Chapter 6 Regional Features

Shown are 1276 global and regional features introduced in Chapter 6. EFC features used with various subsets are listed in Appendix C. Regions are abbreviated as: FULL (whole-peptide), NT (N-termini), AS (Amphipathic Segment) and CT (C-termini) as discussed in Chapter 6. References for AAIndex features are available from the AAIndex website at: `http://www.genome.jp/aaindex`.

Table B.1: Chapter 6 Global and Regional Features

Number	**Region**	**ID**	**Description**
1	FULL	Length	Number of AA
2	FULL	IP	Isoelectric Point
3	FULL	MolWeight	Molecular Weight
4	FULL	Charge	Charge of Peptide
5	FULL	TinyMolPerc	Molar Percent of Tiny AA
6	FULL	SmallMolPerc	Molar Percent of Small AA
7	FULL	AliphaticMolPerc	Molar Percent of Aliphatic AA
8	FULL	AromaticMolPerc	Molar Percent of Aromatic AA
9	FULL	NonPolarMolPerc	Molar Percent of Non-polar AA
10	FULL	PolarMolPerc	Molar Percent of Polar AA
11	FULL	ChargedMolPerc	Molar Percent of Charged AA
12	FULL	BasicMolPerc	Molar Percent of Basic AA
13	FULL	AcidicMolPerc	Molar Percent of Acidic AA
14	FULL	GRAVY	Mean Hydrophobicity
15	FULL	Helix_Propensity	Propensity to from Helix [101]
16	FULL	Turn_Propensity	Propensity to from Turn [101]
17	FULL	Beta_Propensity	Propensity to from Beta-Sheet [101]
18	FULL	*in vitro* Aggregation	Propensity for Aggregation [101]
19	FULL	Length x *in vitro* aggregation	Interaction term from Chapter 4
20	FULL	GRAVY x *in vitro* aggregation	Interaction term from Chapter 4
21	FULL	ANDN920101	alpha-CH chemical shifts (Andersen et al 1992)
22	FULL	ARGP820101	Hydrophobicity index (Argos et al 1982)
23	FULL	ARGP820102	Signal sequence helical potential (Argos et al 1982)
24	FULL	AURR980101	Normalized positional residue frequency at helix termini N4t(Aurora-Rose 1998)
25	FULL	AURR980102	Normalized positional residue frequency at helix termini Nttt (Aurora-Rose 1998)
26	FULL	AURR980103	Normalized positional residue frequency at helix termini N" (Aurora-Rose 1998)
27	FULL	AURR980105	Normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
28	FULL	AURR980106	Normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
29	FULL	AURR980107	Normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
30	FULL	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose 1998)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
31	FULL	AURR980112	Normalized positional residue frequency at helix termini C4 (Aurora-Rose 1998)
32	FULL	AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose 1998)
33	FULL	AURR980117	Normalized positional residue frequency at helix termini C' (Aurora-Rose 1998)
34	FULL	AURR980118	Normalized positional residue frequency at helix termini C" (Aurora-Rose 1998)
35	FULL	AURR980119	Normalized positional residue frequency at helix termini C"' (Aurora-Rose 1998)
36	FULL	AURR980120	Normalized positional residue frequency at helix termini C4' (Aurora-Rose 1998)
37	FULL	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al 2005)
38	FULL	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx 1975)
39	FULL	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx 1975)
40	FULL	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx 1975)
41	FULL	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy 1988)
42	FULL	BIGC670101	Residue volume (Bigelow 1967)
43	FULL	BIOV880101	Information value for accessibility average fraction 35percent (Biou et al 1988)
44	FULL	BIOV880102	Information value for accessibility average fraction 23percent (Biou et al 1988)
45	FULL	BLAS910101	Scaled side chain hydrophobicity values (Black-Mould 1991)
46	FULL	BROC820101	Retention coefficient in TFA (Browne et al 1982)
47	FULL	BULH740101	Transfer free energy to surface (Bull-Breese 1974)
48	FULL	BULH740102	Apparent partial specific volume (Bull-Breese 1974)
49	FULL	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich 1979)
50	FULL	BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich 1979)
51	FULL	BURA740101	Normalized frequency of alpha-helix (Burgess et al 1974)
52	FULL	BURA740102	Normalized frequency of extended structure (Burgess et al 1974)
53	FULL	CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl 1992)
54	FULL	CHAM810101	Steric parameter (Charton 1981)
55	FULL	CHAM820101	Polarizability parameter (Charton-Charton 1982)
56	FULL	CHAM820102	Free energy of solution in water kcal-mole (Charton-Charton 1982)
57	FULL	CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton 1983)
58	FULL	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)
59	FULL	CHAM830103	The number of atoms in the side chain labelled 1-plus1 (Charton-Charton 1983)
60	FULL	CHAM830104	The number of atoms in the side chain labelled 2-plus1 (Charton-Charton 1983)
61	FULL	CHAM830105	The number of atoms in the side chain labelled 3-plus1 (Charton-Charton 1983)
62	FULL	CHAM830107	A parameter of charge transfer capability (Charton-Charton 1983)
63	FULL	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton 1983)
64	FULL	CHOC760101	Residue accessible surface area in tripeptide (Chothia 1976)
65	FULL	CHOC760102	Residue accessible surface area in folded protein (Chothia 1976)
66	FULL	CHOC760103	Proportion of residues 95percent buried (Chothia 1976)
67	FULL	CHOC760104	Proportion of residues 100percent buried (Chothia 1976)
68	FULL	CHOP780202	Normalized frequency of beta-sheet (Chou-Fasman 1978b)
69	FULL	CHOP780203	Normalized frequency of beta-turn (Chou-Fasman 1978b)
70	FULL	CHOP780204	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
71	FULL	CHOP780205	Normalized frequency of C-terminal helix (Chou-Fasman 1978b)
72	FULL	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman 1978b)
73	FULL	CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
74	FULL	CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman 1978b)
75	FULL	CHOP780209	Normalized frequency of C-terminal beta-sheet (Chou-Fasman 1978b)
76	FULL	CHOP780210	Normalized frequency of N-terminal non beta region (Chou-Fasman 1978b)
77	FULL	CHOP780211	Normalized frequency of C-terminal non beta region (Chou-Fasman 1978b)
78	FULL	CHOP780212	Frequency of the 1st residue in turn (Chou-Fasman 1978b)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
79	FULL	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman 1978b)
80	FULL	CHOP780214	Frequency of the 3rd residue in turn (Chou-Fasman 1978b)
81	FULL	CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman 1978b)
82	FULL	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al 1992)
83	FULL	CIDH920103	Normalized hydrophobicity scales for alphaplusbeta-proteins (Cid et al 1992)
84	FULL	CORJ870103	PRIFT index (Cornette et al 1987)
85	FULL	CORJ870108	TOTLS index (Cornette et al 1987)
86	FULL	CRAJ730101	Normalized frequency of middle helix (Crawford et al 1973)
87	FULL	CRAJ730102	Normalized frequency of beta-sheet (Crawford et al 1973)
88	FULL	CRAJ730103	Normalized frequency of turn (Crawford et al 1973)
89	FULL	DAWD720101	Size (Dawson 1972)
90	FULL	DAYM780101	Amino acid composition (Dayhoff et al 1978a)
91	FULL	DAYM780201	Relative mutability (Dayhoff et al 1978b)
92	FULL	DESM900101	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al 1990)
93	FULL	DIGM050101	Hydrostatic pressure asymmetry index PAI (Di Giulio 2005)
94	FULL	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg 1984)
95	FULL	EISD860101	Solvation free energy (Eisenberg-McLachlan 1986)
96	FULL	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan 1986)
97	FULL	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan 1986)
98	FULL	FASG760102	Melting point (Fasman 1976)
99	FULL	FASG760103	Optical rotation (Fasman 1976)
100	FULL	FASG760104	pK-N (Fasman 1976)
101	FULL	FASG760105	pK-C (Fasman 1976)
102	FULL	FAUJ880101	Graph shape index (Fauchere et al 1988)
103	FULL	FAUJ880104	STERIMOL length of the side chain (Fauchere et al 1988)
104	FULL	FAUJ880105	STERIMOL minimum width of the side chain (Fauchere et al 1988)
105	FULL	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al 1988)
106	FULL	FAUJ880107	Nmr chemical shift of alpha-carbon (Fauchere et al 1988)
107	FULL	FAUJ880108	Localized electrical effect (Fauchere et al 1988)
108	FULL	FAUJ880110	Number of full nonbonding orbitals (Fauchere et al 1988)
109	FULL	FAUJ880111	Positive charge (Fauchere et al 1988)
110	FULL	FAUJ880112	Negative charge (Fauchere et al 1988)
111	FULL	FAUJ880113	pK-a(RCOOH) (Fauchere et al 1988)
112	FULL	FINA770101	Helix-coil equilibrium constant (Finkelstein-Ptitsyn 1977)
113	FULL	FINA910101	Helix initiation parameter at posision i-minus1 (Finkelstein et al 1991)
114	FULL	FINA910102	Helix initiation parameter at posision ii-plus1i-plus2 (Finkelstein et al 1991)
115	FULL	FINA910103	Helix termination parameter at posision j-minus2j-minus1j (Finkelstein et al 1991)
116	FULL	FINA910104	Helix termination parameter at posision j-plus1 (Finkelstein et al 1991)
117	FULL	FODM020101	Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi 2002)
118	FULL	FUKS010101	Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
119	FULL	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
120	FULL	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
121	FULL	FUKS010111	Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
122	FULL	FUKS010112	Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa 2001)
123	FULL	GARJ730101	Partition coefficient (Garel et al 1973)
124	FULL	GEIM800101	Alpha-helix indices (Geisow-Roberts 1980)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
125	FULL	GEIM800102	Alpha-helix indices for alpha-proteins (Geisow-Roberts 1980)
126	FULL	GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts 1980)
127	FULL	GEIM800104	Alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
128	FULL	GEIM800105	Beta-strand indices (Geisow-Roberts 1980)
129	FULL	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts 1980)
130	FULL	GEIM800108	Aperiodic indices (Geisow-Roberts 1980)
131	FULL	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts 1980)
132	FULL	GEOR030101	Linker propensity from all dataset (George-Heringa 2003)
133	FULL	GEOR030104	Linker propensity from 3-linker dataset (George-Heringa 2003)
134	FULL	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa 2003)
135	FULL	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa 2003)
136	FULL	GEOR030108	Linker propensity from helical (annotated by DSSP) dataset (George-Heringa 2003)
137	FULL	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa 2003)
138	FULL	GRAR740101	Composition (Grantham 1974)
139	FULL	GRAR740102	Polarity (Grantham 1974)
140	FULL	GRAR740103	Volume (Grantham 1974)
141	FULL	GUYH850101	Partition energy (Guy 1985)
142	FULL	GUYH850105	Apparent partition energies calculated from Chothia index (Guy 1985)
143	FULL	HOPA770101	Hydration number (Hopfinger 1971) Cited by Charton-Charton (1982)
144	FULL	HOPT810101	Hydrophilicity value (Hopp-Woods 1981)
145	FULL	HUTJ700101	Heat capacity (Hutchens 1970)
146	FULL	HUTJ700102	Absolute entropy (Hutchens 1970)
147	FULL	ISOY800101	Normalized relative frequency of alpha-helix (Isogai et al 1980)
148	FULL	ISOY800102	Normalized relative frequency of extended structure (Isogai et al 1980)
149	FULL	ISOY800103	Normalized relative frequency of bend (Isogai et al 1980)
150	FULL	ISOY800106	Normalized relative frequency of helix end (Isogai et al 1980)
151	FULL	ISOY800107	Normalized relative frequency of double bend (Isogai et al 1980)
152	FULL	ISOY800108	Normalized relative frequency of coil (Isogai et al 1980)
153	FULL	JANJ790101	Ratio of buried and accessible molar fractions (Janin 1979)
154	FULL	JANJ790102	Transfer free energy (Janin 1979)
155	FULL	JOND920101	Relative frequency of occurrence (Jones et al 1992)
156	FULL	KANM800104	Average relative probability of inner beta-sheet (Kanehisa-Tsong 1980)
157	FULL	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz 1985)
158	FULL	KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
159	FULL	KHAG800101	The Kerr-constant increments (Khanarian-Moore 1980)
160	FULL	KLEP840101	Net charge (Klein et al 1984)
161	FULL	KOEP990101	Alpha-helix propensity derived from designed sequences (Koehl-Levitt 1999)
162	FULL	KOEP990102	Beta-sheet propensity derived from designed sequences (Koehl-Levitt 1999)
163	FULL	KRIW710101	Side chain interaction parameter (Krigbaum-Rubin 1971)
164	FULL	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya 1979)
165	FULL	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al 2000)
166	FULL	KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al 2000)
167	FULL	LAWE840101	Transfer free energy CHPwater (Lawson et al 1984)
168	FULL	LEVM760103	Side chain angle theta(AAR) (Levitt 1976)
169	FULL	LEVM760106	van der Waals parameter R0 (Levitt 1976)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
170	FULL	LEVM780102	Normalized frequency of beta-sheet with weights (Levitt 1978)
171	FULL	LEVM780103	Normalized frequency of reverse turn with weights (Levitt 1978)
172	FULL	LEWP710101	Frequency of occurrence in beta-bends (Lewis et al 1971)
173	FULL	LIFS790102	Conformational preference for parallel beta-strands (Lifson-Sander 1979)
174	FULL	LIFS790103	Conformational preference for antiparallel beta-strands (Lifson-Sander 1979)
175	FULL	MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga 1976)
176	FULL	MAXF760105	Normalized frequency of zeta L (Maxfield-Scheraga 1976)
177	FULL	MCMT640101	Refractivity (McMeekin et al 1964) Cited by Jones (1975)
178	FULL	MEEJ800101	Retention coefficient in HPLC pH74 (Meek 1980)
179	FULL	MEEJ810101	Retention coefficient in NaClO4 (Meek-Rossetti 1981)
180	FULL	MEIH800101	Average reduced distance for C-alpha (Meirovitch et al 1980)
181	FULL	MEIH800103	Average side chain orientation angle (Meirovitch et al 1980)
182	FULL	MITS020101	Amphiphilicity index (Mitaku et al 2002)
183	FULL	MIYS850101	Effective partition energy (Miyazawa-Jernigan 1985)
184	FULL	MONM990101	Turn propensity scale for transmembrane helices (Monne et al 1999)
185	FULL	NADH010101	Hydropathy scale based on self-information values in the two-state model (5percent accessibility) (Naderi-Manesh et al 2001)
186	FULL	NADH010103	Hydropathy scale based on self-information values in the two-state model (16percent accessibility) (Naderi-Manesh et al 2001)
187	FULL	NADH010106	Hydropathy scale based on self-information values in the two-state model (36percent accessibility) (Naderi-Manesh et al 2001)
188	FULL	NADH010107	Hydropathy scale based on self-information values in the two-state model (50percent accessibility) (Naderi-Manesh et al 2001)
189	FULL	NAGK730103	Normalized frequency of coil (Nagano 1973)
190	FULL	NAKH900102	SD of AA composition of total proteins (Nakashima et al 1990)
191	FULL	NAKH900103	AA composition of mt-proteins (Nakashima et al 1990)
192	FULL	NAKH900104	Normalized composition of mt-proteins (Nakashima et al 1990)
193	FULL	NAKH900109	AA composition of membrane proteins (Nakashima et al 1990)
194	FULL	NAKH900110	Normalized composition of membrane proteins (Nakashima et al 1990)
195	FULL	NAKH900111	Transmembrane regions of non-mt-proteins (Nakashima et al 1990)
196	FULL	NAKH900113	Ratio of average and computed composition (Nakashima et al 1990)
197	FULL	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa 1992)
198	FULL	NAKH920103	AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa 1992)
199	FULL	NISK800101	8 A contact number (Nishikawa-Ooi 1980)
200	FULL	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi 1977)
201	FULL	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi 1977)
202	FULL	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi 1977)
203	FULL	OOBM770104	Average non-bonded energy per residue (Oobatake-Ooi 1977)
204	FULL	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al 1985)
205	FULL	OOBM850103	Optimized transfer energy parameter (Oobatake et al 1985)
206	FULL	OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al 1985)
207	FULL	OOBM850105	Optimized side chain interaction parameter (Oobatake et al 1985)
208	FULL	PALJ810105	Normalized frequency of turn from LG (Palau et al 1981)
209	FULL	PALJ810108	Normalized frequency of alpha-helix in alpha-plusbeta class (Palau et al 1981)
210	FULL	PALJ810111	Normalized frequency of beta-sheet in alpha-plusbeta class (Palau et al 1981)
211	FULL	PALJ810113	Normalized frequency of turn in all-alpha class (Palau et al 1981)
212	FULL	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al 1981)
213	FULL	PALJ810115	Normalized frequency of turn in alphaplusbeta class (Palau et al 1981)
214	FULL	PALJ810116	Normalized frequency of turn in alpha-beta class (Palau et al 1981)
215	FULL	PARS000101	p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
216	FULL	PARS000102	p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
217	FULL	PLIV810101	Partition coefficient (Pliska et al 1981)
218	FULL	PONP800101	Surrounding hydrophobicity in folded form (Ponnuswamy et al 1980)
219	FULL	PONP800104	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al 1980)
220	FULL	PONP800105	Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al 1980)
221	FULL	PONP800106	Surrounding hydrophobicity in turn (Ponnuswamy et al 1980)
222	FULL	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy 1982)
223	FULL	PRAM820102	Slope in regression analysis x 10E1 (Prabhakaran-Ponnuswamy 1982)
224	FULL	PRAM900101	Hydrophobicity (Prabhakaran 1990)
225	FULL	LEVM780101	Normalized frequency of alpha-helix with weights (Levitt 1978)
226	FULL	PTIO830101	Helix-coil equilibrium constant (Ptitsyn-Finkelstein 1983)
227	FULL	QIAN880101	Weights for alpha-helix at the window position of -6 (Qian-Sejnowski 1988)
228	FULL	QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski 1988)
229	FULL	QIAN880103	Weights for alpha-helix at the window position of -4 (Qian-Sejnowski 1988)
230	FULL	QIAN880104	Weights for alpha-helix at the window position of -3 (Qian-Sejnowski 1988)
231	FULL	QIAN880107	Weights for alpha-helix at the window position of 0 (Qian-Sejnowski 1988)
232	FULL	QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski 1988)
233	FULL	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski 1988)
234	FULL	QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski 1988)
235	FULL	QIAN880116	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski 1988)
236	FULL	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski 1988)
237	FULL	QIAN880118	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski 1988)
238	FULL	QIAN880121	Weights for beta-sheet at the window position of 1 (Qian-Sejnowski 1988)
239	FULL	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski 1988)
240	FULL	QIAN880123	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski 1988)
241	FULL	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski 1988)
242	FULL	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski 1988)
243	FULL	QIAN880128	Weights for coil at the window position of -5 (Qian-Sejnowski 1988)
244	FULL	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski 1988)
245	FULL	QIAN880130	Weights for coil at the window position of -3 (Qian-Sejnowski 1988)
246	FULL	QIAN880131	Weights for coil at the window position of -2 (Qian-Sejnowski 1988)
247	FULL	QIAN880135	Weights for coil at the window position of 2 (Qian-Sejnowski 1988)
248	FULL	QIAN880136	Weights for coil at the window position of 3 (Qian-Sejnowski 1988)
249	FULL	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski 1988)
250	FULL	QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski 1988)
251	FULL	QIAN880139	Weights for coil at the window position of 6 (Qian-Sejnowski 1988)
252	FULL	RACS770103	Side chain orientational preference (Rackovsky-Scheraga 1977)
253	FULL	RACS820101	Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
254	FULL	RACS820102	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
255	FULL	RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga 1982)
256	FULL	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)
257	FULL	RACS820105	Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga 1982)
258	FULL	RACS820106	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga 1982)
259	FULL	RACS820107	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga 1982)
260	FULL	RACS820108	Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga 1982)
261	FULL	RACS820110	Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga 1982)
262	FULL	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga 1982)
263	FULL	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga 1982)
264	FULL	RACS820113	Value of theta(i) (Rackovsky-Scheraga 1982)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
265	FULL	RACS820114	Value of theta(i-1) (Rackovsky-Scheraga 1982)
266	FULL	RADA880103	Transfer free energy from vap to chx (Radzicka-Wolfenden 1988)
267	FULL	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden 1988)
268	FULL	RADA880106	Accessible surface area (Radzicka-Wolfenden 1988)
269	FULL	RICJ880101	Relative preference value at N" (Richardson-Richardson 1988)
270	FULL	RICJ880103	Relative preference value at N-cap (Richardson-Richardson 1988)
271	FULL	RICJ880104	Relative preference value at N1 (Richardson-Richardson 1988)
272	FULL	RICJ880105	Relative preference value at N2 (Richardson-Richardson 1988)
273	FULL	RICJ880107	Relative preference value at N4 (Richardson-Richardson 1988)
274	FULL	RICJ880108	Relative preference value at N5 (Richardson-Richardson 1988)
275	FULL	RICJ880109	Relative preference value at Mid (Richardson-Richardson 1988)
276	FULL	RICJ880110	Relative preference value at C5 (Richardson-Richardson 1988)
277	FULL	RICJ880111	Relative preference value at C4 (Richardson-Richardson 1988)
278	FULL	RICJ880112	Relative preference value at C3 (Richardson-Richardson 1988)
279	FULL	RICJ880113	Relative preference value at C2 (Richardson-Richardson 1988)
280	FULL	RICJ880114	Relative preference value at C1 (Richardson-Richardson 1988)
281	FULL	RICJ880116	Relative preference value at C' (Richardson-Richardson 1988)
282	FULL	RICJ880117	Relative preference value at C" (Richardson-Richardson 1988)
283	FULL	ROBB760107	Information measure for extended without H-bond (Robson-Suzuki 1976)
284	FULL	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki 1976)
285	FULL	ROBB790101	Hydration free energy (Robson-Osguthorpe 1979)
286	FULL	ROSM880102	Side chain hydropathy corrected for solvation (Roseman 1988)
287	FULL	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman 1988)
288	FULL	SNEP660101	Principal component I (Sneath 1966)
289	FULL	SNEP660102	Principal component II (Sneath 1966)
290	FULL	SNEP660103	Principal component III (Sneath 1966)
291	FULL	SNEP660104	Principal component IV (Sneath 1966)
292	FULL	SUEM840102	Zimm-Bragg parameter sigma x 10E4 (Sueki et al 1984)
293	FULL	SUYM030101	Linker propensity index (Suyama-Ohara 2003)
294	FULL	SWER830101	Optimal matching hydrophobicity (Sweet-Eisenberg 1983)
295	FULL	TAKK010101	Side-chain contribution to protein stability (kJ-mol) (Takano-Yutani 2001)
296	FULL	TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga 1977)
297	FULL	TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga 1977)
298	FULL	TANS770107	Normalized frequency of left-handed helix (Tanaka-Scheraga 1977)
299	FULL	TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga 1977)
300	FULL	VASM830101	Relative population of conformational state A (Vasquez et al 1983)
301	FULL	VASM830102	Relative population of conformational state C (Vasquez et al 1983)
302	FULL	VASM830103	Relative population of conformational state E (Vasquez et al 1983)
303	FULL	VELV850101	Electron-ion interaction potential (Veljkovic et al 1985)
304	FULL	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al 1994)
305	FULL	WARP780101	Average interactions per side chain atom (Warme-Morgan 1978)
306	FULL	WEBA780101	RF value in high salt chromatography (Weber-Lacey 1978)
307	FULL	WERD780102	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
308	FULL	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga 1978)
309	FULL	WERD780104	Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga 1978)
310	FULL	WILM950101	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
311	FULL	WILM950102	Hydrophobicity coefficient in RP-HPLC C8 with 01percentTFA-MeCN-H2O (Wilce et al 1995)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
312	FULL	WILM950103	Hydrophobicity coefficient in RP-HPLC C4 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
313	FULL	WILM950104	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
314	FULL	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White 1996)
315	FULL	WOLS870102	Principal property value z2 (Wold et al 1987)
316	FULL	WOLS870103	Principal property value z3 (Wold et al 1987)
317	FULL	YUTK870101	Unfolding Gibbs energy in water pH70 (Yutani et al 1987)
318	FULL	YUTK870103	Activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
319	FULL	ZIMJ680101	Hydrophobicity (Zimmerman et al 1968)
320	FULL	BBU_All	SASA Backbone (Upper-bound) All Residues [128]
321	FULL	BBU_Polar	SASA Backbone (Upper-bound) Polar Residues [128]
322	FULL	BBU_Apolar	SASA Backbone (Upper-bound) Apolar Residues [128]
323	FULL	SCL_All	SASA Side-Chain (Lower-bound) All Residues [128]
324	FULL	SCL_Polar	SASA Side-Chain (Lower-bound) Polar Residues [128]
325	FULL	SCL_Apolar	SASA Side-Chain (Lower-bound) Apolar Residues [128]
326	NT	BBU_All	SASA Backbone (Upper-bound) All Residues [128]
327	NT	BBU_Polar	SASA Backbone (Upper-bound) Polar Residues [128]
328	NT	BBU_Apolar	SASA Backbone (Upper-bound) Apolar Residues [128]
329	NT	SCL_All	SASA Side-Chain (Lower-bound) All Residues [128]
330	NT	SCL_Polar	SASA Side-Chain (Lower-bound) Polar Residues [128]
331	NT	SCL_Apolar	SASA Side-Chain (Lower-bound) Apolar Residues [128]
332	CT	BBU_All	SASA Backbone (Upper-bound) All Residues [128]
333	CT	BBU_Polar	SASA Backbone (Upper-bound) Polar Residues [128]
334	CT	BBU_Apolar	SASA Backbone (Upper-bound) Apolar Residues [128]
335	CT	SCL_All	SASA Side-Chain (Lower-bound) All Residues [128]
336	CT	SCL_Polar	SASA Side-Chain (Lower-bound) Polar Residues [128]
337	CT	SCL_Apolar	SASA Side-Chain (Lower-bound) Apolar Residues [128]
338	NT	IP	Isoelectric Point
339	NT	MolWeight	Molecular Weight
340	NT	Charge	Charge of Peptide
341	NT	TinyMolPerc	Molar Percent of Tiny AA
342	NT	SmallMolPerc	Molar Percent of Small AA
343	NT	AliphaticMolPerc	Molar Percent of Aliphatic AA
344	NT	AromaticMolPerc	Molar Percent of Aromatic AA
345	NT	NonPolarMolPerc	Molar Percent of Non-polar AA
346	NT	PolarMolPerc	Molar Percent of Polar AA
347	NT	ChargedMolPerc	Molar Percent of Charged AA
348	NT	BasicMolPerc	Molar Percent of Basic AA
349	NT	AcidicMolPerc	Molar Percent of Acidic AA
350	NT	GRAVY	Mean Hydrophobicity
351	NT	ANDN920101	alpha-CH chemical shifts (Andersen et al 1992)
352	NT	ARGP820101	Hydrophobicity index (Argos et al 1982)
353	NT	ARGP820102	Signal sequence helical potential (Argos et al 1982)
354	NT	AURR980101	Normalized positional residue frequency at helix termini N4t(Aurora-Rose 1998)
355	NT	AURR980102	Normalized positional residue frequency at helix termini N”’ (Aurora-Rose 1998)
356	NT	AURR980103	Normalized positional residue frequency at helix termini N” (Aurora-Rose 1998)
357	NT	AURR980105	Normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
358	NT	AURR980106	Normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
359	NT	AURR980107	Normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
360	NT	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose 1998)
361	NT	AURR980112	Normalized positional residue frequency at helix termini C4 (Aurora-Rose 1998)
362	NT	AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose 1998)
363	NT	AURR980117	Normalized positional residue frequency at helix termini C' (Aurora-Rose 1998)
364	NT	AURR980118	Normalized positional residue frequency at helix termini C" (Aurora-Rose 1998)
365	NT	AURR980119	Normalized positional residue frequency at helix termini C"' (Aurora-Rose 1998)
366	NT	AURR980120	Normalized positional residue frequency at helix termini C4' (Aurora-Rose 1998)
367	NT	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al 2005)
368	NT	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx 1975)
369	NT	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx 1975)
370	NT	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx 1975)
371	NT	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy 1988)
372	NT	BIGC670101	Residue volume (Bigelow 1967)
373	NT	BIOV880101	Information value for accessibility average fraction 35percent (Biou et al 1988)
374	NT	BIOV880102	Information value for accessibility average fraction 23percent (Biou et al 1988)
375	NT	BLAS910101	Scaled side chain hydrophobicity values (Black-Mould 1991)
376	NT	BROC820101	Retention coefficient in TFA (Browne et al 1982)
377	NT	BULH740101	Transfer free energy to surface (Bull-Breese 1974)
378	NT	BULH740102	Apparent partial specific volume (Bull-Breese 1974)
379	NT	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich 1979)
380	NT	BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich 1979)
381	NT	BURA740101	Normalized frequency of alpha-helix (Burgess et al 1974)
382	NT	BURA740102	Normalized frequency of extended structure (Burgess et al 1974)
383	NT	CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl 1992)
384	NT	CHAM810101	Steric parameter (Charton 1981)
385	NT	CHAM820101	Polarizability parameter (Charton-Charton 1982)
386	NT	CHAM820102	Free energy of solution in water kcal-mole (Charton-Charton 1982)
387	NT	CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton 1983)
388	NT	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)
389	NT	CHAM830103	The number of atoms in the side chain labelled 1-plus1 (Charton-Charton 1983)
390	NT	CHAM830104	The number of atoms in the side chain labelled 2-plus1 (Charton-Charton 1983)
391	NT	CHAM830105	The number of atoms in the side chain labelled 3-plus1 (Charton-Charton 1983)
392	NT	CHAM830107	A parameter of charge transfer capability (Charton-Charton 1983)
393	NT	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton 1983)
394	NT	CHOC760101	Residue accessible surface area in tripeptide (Chothia 1976)
395	NT	CHOC760102	Residue accessible surface area in folded protein (Chothia 1976)
396	NT	CHOC760103	Proportion of residues 95percent buried (Chothia 1976)
397	NT	CHOC760104	Proportion of residues 100percent buried (Chothia 1976)
398	NT	CHOP780202	Normalized frequency of beta-sheet (Chou-Fasman 1978b)
399	NT	CHOP780203	Normalized frequency of beta-turn (Chou-Fasman 1978b)
400	NT	CHOP780204	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
401	NT	CHOP780205	Normalized frequency of C-terminal helix (Chou-Fasman 1978b)
402	NT	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman 1978b)
403	NT	CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
404	NT	CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman 1978b)
405	NT	CHOP780209	Normalized frequency of C-terminal beta-sheet (Chou-Fasman 1978b)
406	NT	CHOP780210	Normalized frequency of N-terminal non beta region (Chou-Fasman 1978b)
407	NT	CHOP780211	Normalized frequency of C-terminal non beta region (Chou-Fasman 1978b)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
408	NT	CHOP780212	Frequency of the 1st residue in turn (Chou-Fasman 1978b)
409	NT	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman 1978b)
410	NT	CHOP780214	Frequency of the 3rd residue in turn (Chou-Fasman 1978b)
411	NT	CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman 1978b)
412	NT	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al 1992)
413	NT	CIDH920103	Normalized hydrophobicity scales for alphaplusbeta-proteins (Cid et al 1992)
414	NT	CORJ870103	PRIFT index (Cornette et al 1987)
415	NT	CORJ870108	TOTLS index (Cornette et al 1987)
416	NT	CRAJ730101	Normalized frequency of middle helix (Crawford et al 1973)
417	NT	CRAJ730102	Normalized frequency of beta-sheet (Crawford et al 1973)
418	NT	CRAJ730103	Normalized frequency of turn (Crawford et al 1973)
419	NT	DAWD720101	Size (Dawson 1972)
420	NT	DAYM780101	Amino acid composition (Dayhoff et al 1978a)
421	NT	DAYM780201	Relative mutability (Dayhoff et al 1978b)
422	NT	DESM900101	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al 1990)
423	NT	DIGM050101	Hydrostatic pressure asymmetry index PAI (Di Giulio 2005)
424	NT	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg 1984)
425	NT	EISD860101	Solvation free energy (Eisenberg-McLachlan 1986)
426	NT	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan 1986)
427	NT	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan 1986)
428	NT	FASG760102	Melting point (Fasman 1976)
429	NT	FASG760103	Optical rotation (Fasman 1976)
430	NT	FASG760104	pK-N (Fasman 1976)
431	NT	FASG760105	pK-C (Fasman 1976)
432	NT	FAUJ880101	Graph shape index (Fauchere et al 1988)
433	NT	FAUJ880104	STERIMOL length of the side chain (Fauchere et al 1988)
434	NT	FAUJ880105	STERIMOL minimum width of the side chain (Fauchere et al 1988)
435	NT	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al 1988)
436	NT	FAUJ880107	Nmr chemical shift of alpha-carbon (Fauchere et al 1988)
437	NT	FAUJ880108	Localized electrical effect (Fauchere et al 1988)
438	NT	FAUJ880110	Number of full nonbonding orbitals (Fauchere et al 1988)
439	NT	FAUJ880111	Positive charge (Fauchere et al 1988)
440	NT	FAUJ880112	Negative charge (Fauchere et al 1988)
441	NT	FAUJ880113	pK-a(RCOOH) (Fauchere et al 1988)
442	NT	FINA770101	Helix-coil equilibrium constant (Finkelstein-Ptitsyn 1977)
443	NT	FINA910101	Helix initiation parameter at posision i-minus1 (Finkelstein et al 1991)
444	NT	FINA910102	Helix initiation parameter at posision ii-plus1i-plus2 (Finkelstein et al 1991)
445	NT	FINA910103	Helix termination parameter at posision j-minus2j-minus1j (Finkelstein et al 1991)
446	NT	FINA910104	Helix termination parameter at posision j-plus1 (Finkelstein et al 1991)
447	NT	FODM020101	Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi 2002)
448	NT	FUKS010101	Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
449	NT	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
450	NT	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
451	NT	FUKS010111	Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
452	NT	FUKS010112	Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa 2001)
453	NT	GARJ730101	Partition coefficient (Garel et al 1973)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
454	NT	GEIM800101	Alpha-helix indices (Geisow-Roberts 1980)
455	NT	GEIM800102	Alpha-helix indices for alpha-proteins (Geisow-Roberts 1980)
456	NT	GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts 1980)
457	NT	GEIM800104	Alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
458	NT	GEIM800105	Beta-strand indices (Geisow-Roberts 1980)
459	NT	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts 1980)
460	NT	GEIM800108	Aperiodic indices (Geisow-Roberts 1980)
461	NT	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts 1980)
462	NT	GEOR030101	Linker propensity from all dataset (George-Heringa 2003)
463	NT	GEOR030104	Linker propensity from 3-linker dataset (George-Heringa 2003)
464	NT	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa 2003)
465	NT	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa 2003)
466	NT	GEOR030108	Linker propensity from helical (annotated by DSSP) dataset (George-Heringa 2003)
467	NT	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa 2003)
468	NT	GRAR740101	Composition (Grantham 1974)
469	NT	GRAR740102	Polarity (Grantham 1974)
470	NT	GRAR740103	Volume (Grantham 1974)
471	NT	GUYH850101	Partition energy (Guy 1985)
472	NT	GUYH850105	Apparent partition energies calculated from Chothia index (Guy 1985)
473	NT	HOPA770101	Hydration number (Hopfinger 1971) Cited by Charton-Charton (1982)
474	NT	HOPT810101	Hydrophilicity value (Hopp-Woods 1981)
475	NT	HUTJ700101	Heat capacity (Hutchens 1970)
476	NT	HUTJ700102	Absolute entropy (Hutchens 1970)
477	NT	ISOY800101	Normalized relative frequency of alpha-helix (Isogai et al 1980)
478	NT	ISOY800102	Normalized relative frequency of extended structure (Isogai et al 1980)
479	NT	ISOY800103	Normalized relative frequency of bend (Isogai et al 1980)
480	NT	ISOY800106	Normalized relative frequency of helix end (Isogai et al 1980)
481	NT	ISOY800107	Normalized relative frequency of double bend (Isogai et al 1980)
482	NT	ISOY800108	Normalized relative frequency of coil (Isogai et al 1980)
483	NT	JANJ790101	Ratio of buried and accessible molar fractions (Janin 1979)
484	NT	JANJ790102	Transfer free energy (Janin 1979)
485	NT	JOND920101	Relative frequency of occurrence (Jones et al 1992)
486	NT	KANM800104	Average relative probability of inner beta-sheet (Kanehisa-Tsong 1980)
487	NT	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz 1985)
488	NT	KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
489	NT	KHAG800101	The Kerr-constant increments (Khanarian-Moore 1980)
490	NT	KLEP840101	Net charge (Klein et al 1984)
491	NT	KOEP990101	Alpha-helix propensity derived from designed sequences (Koehl-Levitt 1999)
492	NT	KOEP990102	Beta-sheet propensity derived from designed sequences (Koehl-Levitt 1999)
493	NT	KRIW710101	Side chain interaction parameter (Krigbaum-Rubin 1971)
494	NT	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya 1979)
495	NT	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al 2000)
496	NT	KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al 2000)
497	NT	LAWE840101	Transfer free energy CHPwater (Lawson et al 1984)
498	NT	LEVM760103	Side chain angle theta(AAR) (Levitt 1976)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
499	NT	LEVM760106	van der Waals parameter R0 (Levitt 1976)
500	NT	LEVM780102	Normalized frequency of beta-sheet with weights (Levitt 1978)
501	NT	LEVM780103	Normalized frequency of reverse turn with weights (Levitt 1978)
502	NT	LEWP710101	Frequency of occurrence in beta-bends (Lewis et al 1971)
503	NT	LIFS790102	Conformational preference for parallel beta-strands (Lifson-Sander 1979)
504	NT	LIFS790103	Conformational preference for antiparallel beta-strands (Lifson-Sander 1979)
505	NT	MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga 1976)
506	NT	MAXF760105	Normalized frequency of zeta L (Maxfield-Scheraga 1976)
507	NT	MCMT640101	Refractivity (McMeekin et al 1964) Cited by Jones (1975)
508	NT	MEEJ800101	Retention coefficient in HPLC pH74 (Meek 1980)
509	NT	MEEJ810101	Retention coefficient in NaClO4 (Meek-Rossetti 1981)
510	NT	MEIH800101	Average reduced distance for C-alpha (Meirovitch et al 1980)
511	NT	MEIH800103	Average side chain orientation angle (Meirovitch et al 1980)
512	NT	MITS020101	Amphiphilicity index (Mitaku et al 2002)
513	NT	MIYS850101	Effective partition energy (Miyazawa-Jernigan 1985)
514	NT	MONM990101	Turn propensity scale for transmembrane helices (Monne et al 1999)
515	NT	NADH010101	Hydropathy scale based on self-information values in the two-state model (5percent accessibility) (Naderi-Manesh et al 2001)
516	NT	NADH010103	Hydropathy scale based on self-information values in the two-state model (16percent accessibility) (Naderi-Manesh et al 2001)
517	NT	NADH010106	Hydropathy scale based on self-information values in the two-state model (36percent accessibility) (Naderi-Manesh et al 2001)
518	NT	NADH010107	Hydropathy scale based on self-information values in the two-state model (50percent accessibility) (Naderi-Manesh et al 2001)
519	NT	NAGK730103	Normalized frequency of coil (Nagano 1973)
520	NT	NAKH900102	SD of AA composition of total proteins (Nakashima et al 1990)
521	NT	NAKH900103	AA composition of mt-proteins (Nakashima et al 1990)
522	NT	NAKH900104	Normalized composition of mt-proteins (Nakashima et al 1990)
523	NT	NAKH900109	AA composition of membrane proteins (Nakashima et al 1990)
524	NT	NAKH900110	Normalized composition of membrane proteins (Nakashima et al 1990)
525	NT	NAKH900111	Transmembrane regions of non-mt-proteins (Nakashima et al 1990)
526	NT	NAKH900113	Ratio of average and computed composition (Nakashima et al 1990)
527	NT	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa 1992)
528	NT	NAKH920103	AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa 1992)
529	NT	NISK800101	8 A contact number (Nishikawa-Ooi 1980)
530	NT	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi 1977)
531	NT	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi 1977)
532	NT	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi 1977)
533	NT	OOBM770104	Average non-bonded energy per residue (Oobatake-Ooi 1977)
534	NT	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al 1985)
535	NT	OOBM850103	Optimized transfer energy parameter (Oobatake et al 1985)
536	NT	OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al 1985)
537	NT	OOBM850105	Optimized side chain interaction parameter (Oobatake et al 1985)
538	NT	PALJ810105	Normalized frequency of turn from LG (Palau et al 1981)
539	NT	PALJ810108	Normalized frequency of alpha-helix in alpha-plusbeta class (Palau et al 1981)
540	NT	PALJ810111	Normalized frequency of beta-sheet in alpha-plusbeta class (Palau et al 1981)
541	NT	PALJ810113	Normalized frequency of turn in all-alpha class (Palau et al 1981)
542	NT	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al 1981)
543	NT	PALJ810115	Normalized frequency of turn in alphaplusbeta class (Palau et al 1981)
544	NT	PALJ810116	Normalized frequency of turn in alpha-beta class (Palau et al 1981)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
545	NT	PARS000101	p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
546	NT	PARS000102	p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
547	NT	PLIV810101	Partition coefficient (Pliska et al 1981)
548	NT	PONP800101	Surrounding hydrophobicity in folded form (Ponnuswamy et al 1980)
549	NT	PONP800104	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al 1980)
550	NT	PONP800105	Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al 1980)
551	NT	PONP800106	Surrounding hydrophobicity in turn (Ponnuswamy et al 1980)
552	NT	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy 1982)
553	NT	PRAM820102	Slope in regression analysis x 10E1 (Prabhakaran-Ponnuswamy 1982)
554	NT	PRAM900101	Hydrophobicity (Prabhakaran 1990)
555	NT	LEVM780101	Normalized frequency of alpha-helix with weights (Levitt 1978)
556	NT	PTIO830101	Helix-coil equilibrium constant (Ptitsyn-Finkelstein 1983)
557	NT	QIAN880101	Weights for alpha-helix at the window position of -6 (Qian-Sejnowski 1988)
558	NT	QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski 1988)
559	NT	QIAN880103	Weights for alpha-helix at the window position of -4 (Qian-Sejnowski 1988)
560	NT	QIAN880104	Weights for alpha-helix at the window position of -3 (Qian-Sejnowski 1988)
561	NT	QIAN880107	Weights for alpha-helix at the window position of 0 (Qian-Sejnowski 1988)
562	NT	QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski 1988)
563	NT	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski 1988)
564	NT	QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski 1988)
565	NT	QIAN880116	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski 1988)
566	NT	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski 1988)
567	NT	QIAN880118	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski 1988)
568	NT	QIAN880121	Weights for beta-sheet at the window position of 1 (Qian-Sejnowski 1988)
569	NT	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski 1988)
570	NT	QIAN880123	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski 1988)
571	NT	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski 1988)
572	NT	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski 1988)
573	NT	QIAN880128	Weights for coil at the window position of -5 (Qian-Sejnowski 1988)
574	NT	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski 1988)
575	NT	QIAN880130	Weights for coil at the window position of -3 (Qian-Sejnowski 1988)
576	NT	QIAN880131	Weights for coil at the window position of -2 (Qian-Sejnowski 1988)
577	NT	QIAN880135	Weights for coil at the window position of 2 (Qian-Sejnowski 1988)
578	NT	QIAN880136	Weights for coil at the window position of 3 (Qian-Sejnowski 1988)
579	NT	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski 1988)
580	NT	QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski 1988)
581	NT	QIAN880139	Weights for coil at the window position of 6 (Qian-Sejnowski 1988)
582	NT	RACS770103	Side chain orientational preference (Rackovsky-Scheraga 1977)
583	NT	RACS820101	Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
584	NT	RACS820102	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
585	NT	RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga 1982)
586	NT	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)
587	NT	RACS820105	Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga 1982)
588	NT	RACS820106	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga 1982)
589	NT	RACS820107	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga 1982)
590	NT	RACS820108	Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga 1982)
591	NT	RACS820110	Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga 1982)
592	NT	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga 1982)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
593	NT	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga 1982)
594	NT	RACS820113	Value of theta(i) (Rackovsky-Scheraga 1982)
595	NT	RACS820114	Value of theta(i-1) (Rackovsky-Scheraga 1982)
596	NT	RADA880103	Transfer free energy from vap to chx (Radzicka-Wolfenden 1988)
597	NT	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden 1988)
598	NT	RADA880106	Accessible surface area (Radzicka-Wolfenden 1988)
599	NT	RICJ880101	Relative preference value at N" (Richardson-Richardson 1988)
600	NT	RICJ880103	Relative preference value at N-cap (Richardson-Richardson 1988)
601	NT	RICJ880104	Relative preference value at N1 (Richardson-Richardson 1988)
602	NT	RICJ880105	Relative preference value at N2 (Richardson-Richardson 1988)
603	NT	RICJ880107	Relative preference value at N4 (Richardson-Richardson 1988)
604	NT	RICJ880108	Relative preference value at N5 (Richardson-Richardson 1988)
605	NT	RICJ880109	Relative preference value at Mid (Richardson-Richardson 1988)
606	NT	RICJ880110	Relative preference value at C5 (Richardson-Richardson 1988)
607	NT	RICJ880111	Relative preference value at C4 (Richardson-Richardson 1988)
608	NT	RICJ880112	Relative preference value at C3 (Richardson-Richardson 1988)
609	NT	RICJ880113	Relative preference value at C2 (Richardson-Richardson 1988)
610	NT	RICJ880114	Relative preference value at C1 (Richardson-Richardson 1988)
611	NT	RICJ880116	Relative preference value at C' (Richardson-Richardson 1988)
612	NT	RICJ880117	Relative preference value at C" (Richardson-Richardson 1988)
613	NT	ROBB760107	Information measure for extended without H-bond (Robson-Suzuki 1976)
614	NT	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki 1976)
615	NT	ROBB790101	Hydration free energy (Robson-Osguthorpe 1979)
616	NT	ROSM880102	Side chain hydropathy corrected for solvation (Roseman 1988)
617	NT	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman 1988)
618	NT	SNEP660101	Principal component I (Sneath 1966)
619	NT	SNEP660102	Principal component II (Sneath 1966)
620	NT	SNEP660103	Principal component III (Sneath 1966)
621	NT	SNEP660104	Principal component IV (Sneath 1966)
622	NT	SUEM840102	Zimm-Bragg parameter sigma x 10E4 (Sueki et al 1984)
623	NT	SUYM030101	Linker propensity index (Suyama-Ohara 2003)
624	NT	SWER830101	Optimal matching hydrophobicity (Sweet-Eisenberg 1983)
625	NT	TAKK010101	Side-chain contribution to protein stability (kJ-mol) (Takano-Yutani 2001)
626	NT	TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga 1977)
627	NT	TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga 1977)
628	NT	TANS770107	Normalized frequency of left-handed helix (Tanaka-Scheraga 1977)
629	NT	TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga 1977)
630	NT	VASM830101	Relative population of conformational state A (Vasquez et al 1983)
631	NT	VASM830102	Relative population of conformational state C (Vasquez et al 1983)
632	NT	VASM830103	Relative population of conformational state E (Vasquez et al 1983)
633	NT	VELV850101	Electron-ion interaction potential (Veljkovic et al 1985)
634	NT	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al 1994)
635	NT	WARP780101	Average interactions per side chain atom (Warme-Morgan 1978)
636	NT	WEBA780101	RF value in high salt chromatography (Weber-Lacey 1978)
637	NT	WERD780102	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
638	NT	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga 1978)
639	NT	WERD780104	Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga 1978)
640	NT	WILM950101	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-MeCN-H2O (Wilce et al 1995)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
641	NT	WILM950102	Hydrophobicity coefficient in RP-HPLC C8 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
642	NT	WILM950103	Hydrophobicity coefficient in RP-HPLC C4 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
643	NT	WILM950104	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
644	NT	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White 1996)
645	NT	WOLS870102	Principal property value z2 (Wold et al 1987)
646	NT	WOLS870103	Principal property value z3 (Wold et al 1987)
647	NT	YUTK870101	Unfolding Gibbs energy in water pH70 (Yutani et al 1987)
648	NT	YUTK870103	Activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
649	NT	ZIMJ680101	Hydrophobicity (Zimmerman et al 1968)
650	AS	FaceLen	Length of Hydrophobic Face (see Chapter 6)
651	AS	SeqZ	Number of Charged Residues (see Chapter 6)
652	AS	MAM	Mean Amphipathic Moment (see Chapter 6)
653	AS	IP	Isoelectric Point
654	AS	MolWeight	Molecular Weight
655	AS	Charge	Charge of Peptide
656	AS	TinyMolPerc	Molar Percent of Tiny AA
657	AS	SmallMolPerc	Molar Percent of Small AA
658	AS	AliphaticMolPerc	Molar Percent of Aliphatic AA
659	AS	AromaticMolPerc	Molar Percent of Aromatic AA
660	AS	NonPolarMolPerc	Molar Percent of Non-polar AA
661	AS	PolarMolPerc	Molar Percent of Polar AA
662	AS	ChargedMolPerc	Molar Percent of Charged AA
663	AS	BasicMolPerc	Molar Percent of Basic AA
664	AS	AcidicMolPerc	Molar Percent of Acidic AA
665	AS	GRAVY	Mean Hydrophobicity
666	AS	ANDN920101	alpha-CH chemical shifts (Andersen et al 1992)
667	AS	ARGP820101	Hydrophobicity index (Argos et al 1982)
668	AS	ARGP820102	Signal sequence helical potential (Argos et al 1982)
669	AS	AURR980101	Normalized positional residue frequency at helix termini N4t(Aurora-Rose 1998)
670	AS	AURR980102	Normalized positional residue frequency at helix termini N”’ (Aurora-Rose 1998)
671	AS	AURR980103	Normalized positional residue frequency at helix termini N” (Aurora-Rose 1998)
672	AS	AURR980105	Normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
673	AS	AURR980106	Normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
674	AS	AURR980107	Normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
675	AS	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose 1998)
676	AS	AURR980112	Normalized positional residue frequency at helix termini C4 (Aurora-Rose 1998)
677	AS	AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose 1998)
678	AS	AURR980117	Normalized positional residue frequency at helix termini C’ (Aurora-Rose 1998)
679	AS	AURR980118	Normalized positional residue frequency at helix termini C” (Aurora-Rose 1998)
680	AS	AURR980119	Normalized positional residue frequency at helix termini C”’ (Aurora-Rose 1998)
681	AS	AURR980120	Normalized positional residue frequency at helix termini C4’ (Aurora-Rose 1998)
682	AS	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al 2005)
683	AS	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx 1975)
684	AS	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx 1975)
685	AS	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx 1975)
686	AS	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy 1988)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
687	AS	BIGC670101	Residue volume (Bigelow 1967)
688	AS	BIOV880101	Information value for accessibility average fraction 35percent (Biou et al 1988)
689	AS	BIOV880102	Information value for accessibility average fraction 23percent (Biou et al 1988)
690	AS	BLAS910101	Scaled side chain hydrophobicity values (Black-Mould 1991)
691	AS	BROC820101	Retention coefficient in TFA (Browne et al 1982)
692	AS	BULH740101	Transfer free energy to surface (Bull-Breese 1974)
693	AS	BULH740102	Apparent partial specific volume (Bull-Breese 1974)
694	AS	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich 1979)
695	AS	BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich 1979)
696	AS	BURA740101	Normalized frequency of alpha-helix (Burgess et al 1974)
697	AS	BURA740102	Normalized frequency of extended structure (Burgess et al 1974)
698	AS	CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl 1992)
699	AS	CHAM810101	Steric parameter (Charton 1981)
700	AS	CHAM820101	Polarizability parameter (Charton-Charton 1982)
701	AS	CHAM820102	Free energy of solution in water kcal-mole (Charton-Charton 1982)
702	AS	CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton 1983)
703	AS	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)
704	AS	CHAM830103	The number of atoms in the side chain labelled 1-plus1 (Charton-Charton 1983)
705	AS	CHAM830104	The number of atoms in the side chain labelled 2-plus1 (Charton-Charton 1983)
706	AS	CHAM830105	The number of atoms in the side chain labelled 3-plus1 (Charton-Charton 1983)
707	AS	CHAM830107	A parameter of charge transfer capability (Charton-Charton 1983)
708	AS	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton 1983)
709	AS	CHOC760101	Residue accessible surface area in tripeptide (Chothia 1976)
710	AS	CHOC760102	Residue accessible surface area in folded protein (Chothia 1976)
711	AS	CHOC760103	Proportion of residues 95percent buried (Chothia 1976)
712	AS	CHOC760104	Proportion of residues 100percent buried (Chothia 1976)
713	AS	CHOP780202	Normalized frequency of beta-sheet (Chou-Fasman 1978b)
714	AS	CHOP780203	Normalized frequency of beta-turn (Chou-Fasman 1978b)
715	AS	CHOP780204	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
716	AS	CHOP780205	Normalized frequency of C-terminal helix (Chou-Fasman 1978b)
717	AS	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman 1978b)
718	AS	CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
719	AS	CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman 1978b)
720	AS	CHOP780209	Normalized frequency of C-terminal beta-sheet (Chou-Fasman 1978b)
721	AS	CHOP780210	Normalized frequency of N-terminal non beta region (Chou-Fasman 1978b)
722	AS	CHOP780211	Normalized frequency of C-terminal non beta region (Chou-Fasman 1978b)
723	AS	CHOP780212	Frequency of the 1st residue in turn (Chou-Fasman 1978b)
724	AS	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman 1978b)
725	AS	CHOP780214	Frequency of the 3rd residue in turn (Chou-Fasman 1978b)
726	AS	CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman 1978b)
727	AS	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al 1992)
728	AS	CIDH920103	Normalized hydrophobicity scales for alphaplusbeta-proteins (Cid et al 1992)
729	AS	CORJ870103	PRIFT index (Cornette et al 1987)
730	AS	CORJ870108	TOTLS index (Cornette et al 1987)
731	AS	CRAJ730101	Normalized frequency of middle helix (Crawford et al 1973)
732	AS	CRAJ730102	Normalized frequency of beta-sheet (Crawford et al 1973)
733	AS	CRAJ730103	Normalized frequency of turn (Crawford et al 1973)
734	AS	DAWD720101	Size (Dawson 1972)
735	AS	DAYM780101	Amino acid composition (Dayhoff et al 1978a)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
736	AS	DAYM780201	Relative mutability (Dayhoff et al 1978b)
737	AS	DESM900101	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al 1990)
738	AS	DIGM050101	Hydrostatic pressure asymmetry index PAI (Di Giulio 2005)
739	AS	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg 1984)
740	AS	EISD860101	Solvation free energy (Eisenberg-McLachlan 1986)
741	AS	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan 1986)
742	AS	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan 1986)
743	AS	FASG760102	Melting point (Fasman 1976)
744	AS	FASG760103	Optical rotation (Fasman 1976)
745	AS	FASG760104	pK-N (Fasman 1976)
746	AS	FASG760105	pK-C (Fasman 1976)
747	AS	FAUJ880101	Graph shape index (Fauchere et al 1988)
748	AS	FAUJ880104	STERIMOL length of the side chain (Fauchere et al 1988)
749	AS	FAUJ880105	STERIMOL minimum width of the side chain (Fauchere et al 1988)
750	AS	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al 1988)
751	AS	FAUJ880107	Nmr chemical shift of alpha-carbon (Fauchere et al 1988)
752	AS	FAUJ880108	Localized electrical effect (Fauchere et al 1988)
753	AS	FAUJ880110	Number of full nonbonding orbitals (Fauchere et al 1988)
754	AS	FAUJ880111	Positive charge (Fauchere et al 1988)
755	AS	FAUJ880112	Negative charge (Fauchere et al 1988)
756	AS	FAUJ880113	pK-a(RCOOH) (Fauchere et al 1988)
757	AS	FINA770101	Helix-coil equilibrium constant (Finkelstein-Ptitsyn 1977)
758	AS	FINA910101	Helix initiation parameter at posision i-minus1 (Finkelstein et al 1991)
759	AS	FINA910102	Helix initiation parameter at posision ii-plus1i-plus2 (Finkelstein et al 1991)
760	AS	FINA910103	Helix termination parameter at posision j-minus2j-minus1j (Finkelstein et al 1991)
761	AS	FINA910104	Helix termination parameter at posision j-plus1 (Finkelstein et al 1991)
762	AS	FODM020101	Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi 2002)
763	AS	FUKS010101	Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
764	AS	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
765	AS	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
766	AS	FUKS010111	Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
767	AS	FUKS010112	Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa 2001)
768	AS	GARJ730101	Partition coefficient (Garel et al 1973)
769	AS	GEIM800101	Alpha-helix indices (Geisow-Roberts 1980)
770	AS	GEIM800102	Alpha-helix indices for alpha-proteins (Geisow-Roberts 1980)
771	AS	GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts 1980)
772	AS	GEIM800104	Alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
773	AS	GEIM800105	Beta-strand indices (Geisow-Roberts 1980)
774	AS	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts 1980)
775	AS	GEIM800108	Aperiodic indices (Geisow-Roberts 1980)
776	AS	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts 1980)
777	AS	GEOR030101	Linker propensity from all dataset (George-Heringa 2003)
778	AS	GEOR030104	Linker propensity from 3-linker dataset (George-Heringa 2003)
779	AS	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa 2003)
780	AS	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa 2003)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
781	AS	GEOR030108	Linker propensity from helical (annotated by DSSP) dataset (George-Heringa 2003)
782	AS	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa 2003)
783	AS	GRAR740101	Composition (Grantham 1974)
784	AS	GRAR740102	Polarity (Grantham 1974)
785	AS	GRAR740103	Volume (Grantham 1974)
786	AS	GUYH850101	Partition energy (Guy 1985)
787	AS	GUYH850105	Apparent partition energies calculated from Chothia index (Guy 1985)
788	AS	HOPA770101	Hydration number (Hopfinger 1971) Cited by Charton-Charton (1982)
789	AS	HOPT810101	Hydrophilicity value (Hopp-Woods 1981)
790	AS	HUTJ700101	Heat capacity (Hutchens 1970)
791	AS	HUTJ700102	Absolute entropy (Hutchens 1970)
792	AS	ISOY800101	Normalized relative frequency of alpha-helix (Isogai et al 1980)
793	AS	ISOY800102	Normalized relative frequency of extended structure (Isogai et al 1980)
794	AS	ISOY800103	Normalized relative frequency of bend (Isogai et al 1980)
795	AS	ISOY800106	Normalized relative frequency of helix end (Isogai et al 1980)
796	AS	ISOY800107	Normalized relative frequency of double bend (Isogai et al 1980)
797	AS	ISOY800108	Normalized relative frequency of coil (Isogai et al 1980)
798	AS	JANJ790101	Ratio of buried and accessible molar fractions (Janin 1979)
799	AS	JANJ790102	Transfer free energy (Janin 1979)
800	AS	JOND920101	Relative frequency of occurrence (Jones et al 1992)
801	AS	KANM800104	Average relative probability of inner beta-sheet (Kanehisa-Tsong 1980)
802	AS	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz 1985)
803	AS	KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
804	AS	KHAG800101	The Kerr-constant increments (Khanarian-Moore 1980)
805	AS	KLEP840101	Net charge (Klein et al 1984)
806	AS	KOEP990101	Alpha-helix propensity derived from designed sequences (Koehl-Levitt 1999)
807	AS	KOEP990102	Beta-sheet propensity derived from designed sequences (Koehl-Levitt 1999)
808	AS	KRIW710101	Side chain interaction parameter (Krigbaum-Rubin 1971)
809	AS	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya 1979)
810	AS	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al 2000)
811	AS	KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al 2000)
812	AS	LAWE840101	Transfer free energy CHPwater (Lawson et al 1984)
813	AS	LEVM760103	Side chain angle theta(AAR) (Levitt 1976)
814	AS	LEVM760106	van der Waals parameter R0 (Levitt 1976)
815	AS	LEVM780102	Normalized frequency of beta-sheet with weights (Levitt 1978)
816	AS	LEVM780103	Normalized frequency of reverse turn with weights (Levitt 1978)
817	AS	LEWP710101	Frequency of occurrence in beta-bends (Lewis et al 1971)
818	AS	LIFS790102	Conformational preference for parallel beta-strands (Lifson-Sander 1979)
819	AS	LIFS790103	Conformational preference for antiparallel beta-strands (Lifson-Sander 1979)
820	AS	MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga 1976)
821	AS	MAXF760105	Normalized frequency of zeta L (Maxfield-Scheraga 1976)
822	AS	MCMT640101	Refractivity (McMeekin et al 1964) Cited by Jones (1975)
823	AS	MEEJ800101	Retention coefficient in HPLC pH74 (Meek 1980)
824	AS	MEEJ810101	Retention coefficient in NaClO4 (Meek-Rossetti 1981)
825	AS	MEIH800101	Average reduced distance for C-alpha (Meirovitch et al 1980)
826	AS	MEIH800103	Average side chain orientation angle (Meirovitch et al 1980)
827	AS	MITS020101	Amphiphilicity index (Mitaku et al 2002)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
828	AS	MIYS850101	Effective partition energy (Miyazawa-Jernigan 1985)
829	AS	MONM990101	Turn propensity scale for transmembrane helices (Monne et al 1999)
830	AS	NADH010101	Hydropathy scale based on self-information values in the two-state model (5per-cent accessibility) (Naderi-Manesh et al 2001)
831	AS	NADH010103	Hydropathy scale based on self-information values in the two-state model (16per-cent accessibility) (Naderi-Manesh et al 2001)
832	AS	NADH010106	Hydropathy scale based on self-information values in the two-state model (36per-cent accessibility) (Naderi-Manesh et al 2001)
833	AS	NADH010107	Hydropathy scale based on self-information values in the two-state model (50per-cent accessibility) (Naderi-Manesh et al 2001)
834	AS	NAGK730103	Normalized frequency of coil (Nagano 1973)
835	AS	NAKH900102	SD of AA composition of total proteins (Nakashima et al 1990)
836	AS	NAKH900103	AA composition of mt-proteins (Nakashima et al 1990)
837	AS	NAKH900104	Normalized composition of mt-proteins (Nakashima et al 1990)
838	AS	NAKH900109	AA composition of membrane proteins (Nakashima et al 1990)
839	AS	NAKH900110	Normalized composition of membrane proteins (Nakashima et al 1990)
840	AS	NAKH900111	Transmembrane regions of non-mt-proteins (Nakashima et al 1990)
841	AS	NAKH900113	Ratio of average and computed composition (Nakashima et al 1990)
842	AS	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa 1992)
843	AS	NAKH920103	AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa 1992)
844	AS	NISK800101	8 A contact number (Nishikawa-Ooi 1980)
845	AS	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi 1977)
846	AS	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi 1977)
847	AS	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi 1977)
848	AS	OOBM770104	Average non-bonded energy per residue (Oobatake-Ooi 1977)
849	AS	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al 1985)
850	AS	OOBM850103	Optimized transfer energy parameter (Oobatake et al 1985)
851	AS	OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al 1985)
852	AS	OOBM850105	Optimized side chain interaction parameter (Oobatake et al 1985)
853	AS	PALJ810105	Normalized frequency of turn from LG (Palau et al 1981)
854	AS	PALJ810108	Normalized frequency of alpha-helix in alpha-plusbeta class (Palau et al 1981)
855	AS	PALJ810111	Normalized frequency of beta-sheet in alpha-plusbeta class (Palau et al 1981)
856	AS	PALJ810113	Normalized frequency of turn in all-alpha class (Palau et al 1981)
857	AS	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al 1981)
858	AS	PALJ810115	Normalized frequency of turn in alphaplusbeta class (Palau et al 1981)
859	AS	PALJ810116	Normalized frequency of turn in alpha-beta class (Palau et al 1981)
860	AS	PARS000101	p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
861	AS	PARS000102	p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
862	AS	PLIV810101	Partition coefficient (Pliska et al 1981)
863	AS	PONP800101	Surrounding hydrophobicity in folded form (Ponnuswamy et al 1980)
864	AS	PONP800104	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al 1980)
865	AS	PONP800105	Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al 1980)
866	AS	PONP800106	Surrounding hydrophobicity in turn (Ponnuswamy et al 1980)
867	AS	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy 1982)
868	AS	PRAM820102	Slope in regression analysis x 10E1 (Prabhakaran-Ponnuswamy 1982)
869	AS	PRAM900101	Hydrophobicity (Prabhakaran 1990)
870	AS	LEVM780101	Normalized frequency of alpha-helix with weights (Levitt 1978)
871	AS	PTIO830101	Helix-coil equilibrium constant (Ptitsyn-Finkelstein 1983)
872	AS	QIAN880101	Weights for alpha-helix at the window position of -6 (Qian-Sejnowski 1988)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
873	AS	QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski 1988)
874	AS	QIAN880103	Weights for alpha-helix at the window position of -4 (Qian-Sejnowski 1988)
875	AS	QIAN880104	Weights for alpha-helix at the window position of -3 (Qian-Sejnowski 1988)
876	AS	QIAN880107	Weights for alpha-helix at the window position of 0 (Qian-Sejnowski 1988)
877	AS	QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski 1988)
878	AS	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski 1988)
879	AS	QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski 1988)
880	AS	QIAN880116	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski 1988)
881	AS	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski 1988)
882	AS	QIAN880118	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski 1988)
883	AS	QIAN880121	Weights for beta-sheet at the window position of 1 (Qian-Sejnowski 1988)
884	AS	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski 1988)
885	AS	QIAN880123	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski 1988)
886	AS	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski 1988)
887	AS	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski 1988)
888	AS	QIAN880128	Weights for coil at the window position of -5 (Qian-Sejnowski 1988)
889	AS	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski 1988)
890	AS	QIAN880130	Weights for coil at the window position of -3 (Qian-Sejnowski 1988)
891	AS	QIAN880131	Weights for coil at the window position of -2 (Qian-Sejnowski 1988)
892	AS	QIAN880135	Weights for coil at the window position of 2 (Qian-Sejnowski 1988)
893	AS	QIAN880136	Weights for coil at the window position of 3 (Qian-Sejnowski 1988)
894	AS	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski 1988)
895	AS	QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski 1988)
896	AS	QIAN880139	Weights for coil at the window position of 6 (Qian-Sejnowski 1988)
897	AS	RACS770103	Side chain orientational preference (Rackovsky-Scheraga 1977)
898	AS	RACS820101	Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
899	AS	RACS820102	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
900	AS	RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga 1982)
901	AS	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)
902	AS	RACS820105	Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga 1982)
903	AS	RACS820106	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga 1982)
904	AS	RACS820107	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga 1982)
905	AS	RACS820108	Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga 1982)
906	AS	RACS820110	Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga 1982)
907	AS	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga 1982)
908	AS	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga 1982)
909	AS	RACS820113	Value of theta(i) (Rackovsky-Scheraga 1982)
910	AS	RACS820114	Value of theta(i-1) (Rackovsky-Scheraga 1982)
911	AS	RADA880103	Transfer free energy from vap to chx (Radzicka-Wolfenden 1988)
912	AS	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden 1988)
913	AS	RADA880106	Accessible surface area (Radzicka-Wolfenden 1988)
914	AS	RICJ880101	Relative preference value at N" (Richardson-Richardson 1988)
915	AS	RICJ880103	Relative preference value at N-cap (Richardson-Richardson 1988)
916	AS	RICJ880104	Relative preference value at N1 (Richardson-Richardson 1988)
917	AS	RICJ880105	Relative preference value at N2 (Richardson-Richardson 1988)
918	AS	RICJ880107	Relative preference value at N4 (Richardson-Richardson 1988)
919	AS	RICJ880108	Relative preference value at N5 (Richardson-Richardson 1988)
920	AS	RICJ880109	Relative preference value at Mid (Richardson-Richardson 1988)
921	AS	RICJ880110	Relative preference value at C5 (Richardson-Richardson 1988)
922	AS	RICJ880111	Relative preference value at C4 (Richardson-Richardson 1988)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
923	AS	RICJ880112	Relative preference value at C3 (Richardson-Richardson 1988)
924	AS	RICJ880113	Relative preference value at C2 (Richardson-Richardson 1988)
925	AS	RICJ880114	Relative preference value at C1 (Richardson-Richardson 1988)
926	AS	RICJ880116	Relative preference value at C' (Richardson-Richardson 1988)
927	AS	RICJ880117	Relative preference value at C" (Richardson-Richardson 1988)
928	AS	ROBB760107	Information measure for extended without H-bond (Robson-Suzuki 1976)
929	AS	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki 1976)
930	AS	ROBB790101	Hydration free energy (Robson-Osguthorpe 1979)
931	AS	ROSM880102	Side chain hydropathy corrected for solvation (Roseman 1988)
932	AS	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman 1988)
933	AS	SNEP660101	Principal component I (Sneath 1966)
934	AS	SNEP660102	Principal component II (Sneath 1966)
935	AS	SNEP660103	Principal component III (Sneath 1966)
936	AS	SNEP660104	Principal component IV (Sneath 1966)
937	AS	SUEM840102	Zimm-Bragg parameter sigma x 10E4 (Sueki et al 1984)
938	AS	SUYM030101	Linker propensity index (Suyama-Ohara 2003)
939	AS	SWER830101	Optimal matching hydrophobicity (Sweet-Eisenberg 1983)
940	AS	TAKK010101	Side-chain contribution to protein stability (kJ-mol) (Takano-Yutani 2001)
941	AS	TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga 1977)
942	AS	TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga 1977)
943	AS	TANS770107	Normalized frequency of left-handed helix (Tanaka-Scheraga 1977)
944	AS	TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga 1977)
945	AS	VASM830101	Relative population of conformational state A (Vasquez et al 1983)
946	AS	VASM830102	Relative population of conformational state C (Vasquez et al 1983)
947	AS	VASM830103	Relative population of conformational state E (Vasquez et al 1983)
948	AS	VELV850101	Electron-ion interaction potential (Veljkovic et al 1985)
949	AS	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al 1994)
950	AS	WARP780101	Average interactions per side chain atom (Warme-Morgan 1978)
951	AS	WEBA780101	RF value in high salt chromatography (Weber-Lacey 1978)
952	AS	WERD780102	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
953	AS	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga 1978)
954	AS	WERD780104	Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga 1978)
955	AS	WILM950101	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
956	AS	WILM950102	Hydrophobicity coefficient in RP-HPLC C8 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
957	AS	WILM950103	Hydrophobicity coefficient in RP-HPLC C4 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
958	AS	WILM950104	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
959	AS	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White 1996)
960	AS	WOLS870102	Principal property value z2 (Wold et al 1987)
961	AS	WOLS870103	Principal property value z3 (Wold et al 1987)
962	AS	YUTK870101	Unfolding Gibbs energy in water pH70 (Yutani et al 1987)
963	AS	YUTK870103	Activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
964	AS	ZIMJ680101	Hydrophobicity (Zimmerman et al 1968)
965	CT	IP	Isoelectric Point
966	CT	MolWeight	Molecular Weight
967	CT	Charge	Charge of Peptide

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
968	CT	TinyMolPerc	Molar Percent of Tiny AA
969	CT	SmallMolPerc	Molar Percent of Small AA
970	CT	AliphaticMolPerc	Molar Percent of Aliphatic AA
971	CT	AromaticMolPerc	Molar Percent of Aromatic AA
972	CT	NonPolarMolPerc	Molar Percent of Non-polar AA
973	CT	PolarMolPerc	Molar Percent of Polar AA
974	CT	ChargedMolPerc	Molar Percent of Charged AA
975	CT	BasicMolPerc	Molar Percent of Basic AA
976	CT	AcidicMolPerc	Molar Percent of Acidic AA
977	CT	GRAVY	Mean Hydrophobicity
978	CT	ANDN920101	alpha-CH chemical shifts (Andersen et al 1992)
979	CT	ARGP820101	Hydrophobicity index (Argos et al 1982)
980	CT	ARGP820102	Signal sequence helical potential (Argos et al 1982)
981	CT	AURR980101	Normalized positional residue frequency at helix termini N4t(Aurora-Rose 1998)
982	CT	AURR980102	Normalized positional residue frequency at helix termini N”’ (Aurora-Rose 1998)
983	CT	AURR980103	Normalized positional residue frequency at helix termini N” (Aurora-Rose 1998)
984	CT	AURR980105	Normalized positional residue frequency at helix termini Nc (Aurora-Rose 1998)
985	CT	AURR980106	Normalized positional residue frequency at helix termini N1 (Aurora-Rose 1998)
986	CT	AURR980107	Normalized positional residue frequency at helix termini N2 (Aurora-Rose 1998)
987	CT	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose 1998)
988	CT	AURR980112	Normalized positional residue frequency at helix termini C4 (Aurora-Rose 1998)
989	CT	AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose 1998)
990	CT	AURR980117	Normalized positional residue frequency at helix termini C’ (Aurora-Rose 1998)
991	CT	AURR980118	Normalized positional residue frequency at helix termini C” (Aurora-Rose 1998)
992	CT	AURR980119	Normalized positional residue frequency at helix termini C”’ (Aurora-Rose 1998)
993	CT	AURR980120	Normalized positional residue frequency at helix termini C4’ (Aurora-Rose 1998)
994	CT	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al 2005)
995	CT	BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx 1975)
996	CT	BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx 1975)
997	CT	BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx 1975)
998	CT	BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy 1988)
999	CT	BIGC670101	Residue volume (Bigelow 1967)
1000	CT	BIOV880101	Information value for accessibility average fraction 35percent (Biou et al 1988)
1001	CT	BIOV880102	Information value for accessibility average fraction 23percent (Biou et al 1988)
1002	CT	BLAS910101	Scaled side chain hydrophobicity values (Black-Mould 1991)
1003	CT	BROC820101	Retention coefficient in TFA (Browne et al 1982)
1004	CT	BULH740101	Transfer free energy to surface (Bull-Breese 1974)
1005	CT	BULH740102	Apparent partial specific volume (Bull-Breese 1974)
1006	CT	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich 1979)
1007	CT	BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich 1979)
1008	CT	BURA740101	Normalized frequency of alpha-helix (Burgess et al 1974)
1009	CT	BURA740102	Normalized frequency of extended structure (Burgess et al 1974)
1010	CT	CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl 1992)
1011	CT	CHAM810101	Steric parameter (Charton 1981)
1012	CT	CHAM820101	Polarizability parameter (Charton-Charton 1982)
1013	CT	CHAM820102	Free energy of solution in water kcal-mole (Charton-Charton 1982)
1014	CT	CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton 1983)
1015	CT	CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton 1983)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1016	CT	CHAM830103	The number of atoms in the side chain labelled 1-plus1 (Charton-Charton 1983)
1017	CT	CHAM830104	The number of atoms in the side chain labelled 2-plus1 (Charton-Charton 1983)
1018	CT	CHAM830105	The number of atoms in the side chain labelled 3-plus1 (Charton-Charton 1983)
1019	CT	CHAM830107	A parameter of charge transfer capability (Charton-Charton 1983)
1020	CT	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton 1983)
1021	CT	CHOC760101	Residue accessible surface area in tripeptide (Chothia 1976)
1022	CT	CHOC760102	Residue accessible surface area in folded protein (Chothia 1976)
1023	CT	CHOC760103	Proportion of residues 95percent buried (Chothia 1976)
1024	CT	CHOC760104	Proportion of residues 100percent buried (Chothia 1976)
1025	CT	CHOP780202	Normalized frequency of beta-sheet (Chou-Fasman 1978b)
1026	CT	CHOP780203	Normalized frequency of beta-turn (Chou-Fasman 1978b)
1027	CT	CHOP780204	Normalized frequency of N-terminal helix (Chou-Fasman 1978b)
1028	CT	CHOP780205	Normalized frequency of C-terminal helix (Chou-Fasman 1978b)
1029	CT	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman 1978b)
1030	CT	CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman 1978b)
1031	CT	CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman 1978b)
1032	CT	CHOP780209	Normalized frequency of C-terminal beta-sheet (Chou-Fasman 1978b)
1033	CT	CHOP780210	Normalized frequency of N-terminal non beta region (Chou-Fasman 1978b)
1034	CT	CHOP780211	Normalized frequency of C-terminal non beta region (Chou-Fasman 1978b)
1035	CT	CHOP780212	Frequency of the 1st residue in turn (Chou-Fasman 1978b)
1036	CT	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman 1978b)
1037	CT	CHOP780214	Frequency of the 3rd residue in turn (Chou-Fasman 1978b)
1038	CT	CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman 1978b)
1039	CT	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al 1992)
1040	CT	CIDH920103	Normalized hydrophobicity scales for alphaplusbeta-proteins (Cid et al 1992)
1041	CT	CORJ870103	PRIFT index (Cornette et al 1987)
1042	CT	CORJ870108	TOTLS index (Cornette et al 1987)
1043	CT	CRAJ730101	Normalized frequency of middle helix (Crawford et al 1973)
1044	CT	CRAJ730102	Normalized frequency of beta-sheet (Crawford et al 1973)
1045	CT	CRAJ730103	Normalized frequency of turn (Crawford et al 1973)
1046	CT	DAWD720101	Size (Dawson 1972)
1047	CT	DAYM780101	Amino acid composition (Dayhoff et al 1978a)
1048	CT	DAYM780201	Relative mutability (Dayhoff et al 1978b)
1049	CT	DESM900101	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al 1990)
1050	CT	DIGM050101	Hydrostatic pressure asymmetry index PAI (Di Giulio 2005)
1051	CT	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg 1984)
1052	CT	EISD860101	Solvation free energy (Eisenberg-McLachlan 1986)
1053	CT	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan 1986)
1054	CT	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan 1986)
1055	CT	FASG760102	Melting point (Fasman 1976)
1056	CT	FASG760103	Optical rotation (Fasman 1976)
1057	CT	FASG760104	pK-N (Fasman 1976)
1058	CT	FASG760105	pK-C (Fasman 1976)
1059	CT	FAUJ880101	Graph shape index (Fauchere et al 1988)
1060	CT	FAUJ880104	STERIMOL length of the side chain (Fauchere et al 1988)
1061	CT	FAUJ880105	STERIMOL minimum width of the side chain (Fauchere et al 1988)
1062	CT	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al 1988)
1063	CT	FAUJ880107	Nmr chemical shift of alpha-carbon (Fauchere et al 1988)
1064	CT	FAUJ880108	Localized electrical effect (Fauchere et al 1988)
1065	CT	FAUJ880110	Number of full nonbonding orbitals (Fauchere et al 1988)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1066	CT	FAUJ880111	Positive charge (Fauchere et al 1988)
1067	CT	FAUJ880112	Negative charge (Fauchere et al 1988)
1068	CT	FAUJ880113	pK-a(RCOOH) (Fauchere et al 1988)
1069	CT	FINA770101	Helix-coil equilibrium constant (Finkelstein-Ptitsyn 1977)
1070	CT	FINA910101	Helix initiation parameter at posision i-minus1 (Finkelstein et al 1991)
1071	CT	FINA910102	Helix initiation parameter at posision ii-plus1i-plus2 (Finkelstein et al 1991)
1072	CT	FINA910103	Helix termination parameter at posision j-minus2j-minus1j (Finkelstein et al 1991)
1073	CT	FINA910104	Helix termination parameter at posision j-plus1 (Finkelstein et al 1991)
1074	CT	FODM020101	Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi 2002)
1075	CT	FUKS010101	Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
1076	CT	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
1077	CT	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa 2001)
1078	CT	FUKS010111	Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa 2001)
1079	CT	FUKS010112	Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa 2001)
1080	CT	GARJ730101	Partition coefficient (Garel et al 1973)
1081	CT	GEIM800101	Alpha-helix indices (Geisow-Roberts 1980)
1082	CT	GEIM800102	Alpha-helix indices for alpha-proteins (Geisow-Roberts 1980)
1083	CT	GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts 1980)
1084	CT	GEIM800104	Alpha-helix indices for alpha-beta-proteins (Geisow-Roberts 1980)
1085	CT	GEIM800105	Beta-strand indices (Geisow-Roberts 1980)
1086	CT	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts 1980)
1087	CT	GEIM800108	Aperiodic indices (Geisow-Roberts 1980)
1088	CT	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts 1980)
1089	CT	GEOR030101	Linker propensity from all dataset (George-Heringa 2003)
1090	CT	GEOR030104	Linker propensity from 3-linker dataset (George-Heringa 2003)
1091	CT	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa 2003)
1092	CT	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa 2003)
1093	CT	GEOR030108	Linker propensity from helical (annotated by DSSP) dataset (George-Heringa 2003)
1094	CT	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa 2003)
1095	CT	GRAR740101	Composition (Grantham 1974)
1096	CT	GRAR740102	Polarity (Grantham 1974)
1097	CT	GRAR740103	Volume (Grantham 1974)
1098	CT	GUYH850101	Partition energy (Guy 1985)
1099	CT	GUYH850105	Apparent partition energies calculated from Chothia index (Guy 1985)
1100	CT	HOPA770101	Hydration number (Hopfinger 1971) Cited by Charton-Charton (1982)
1101	CT	HOPT810101	Hydrophilicity value (Hopp-Woods 1981)
1102	CT	HUTJ700101	Heat capacity (Hutchens 1970)
1103	CT	HUTJ700102	Absolute entropy (Hutchens 1970)
1104	CT	ISOY800101	Normalized relative frequency of alpha-helix (Isogai et al 1980)
1105	CT	ISOY800102	Normalized relative frequency of extended structure (Isogai et al 1980)
1106	CT	ISOY800103	Normalized relative frequency of bend (Isogai et al 1980)
1107	CT	ISOY800106	Normalized relative frequency of helix end (Isogai et al 1980)
1108	CT	ISOY800107	Normalized relative frequency of double bend (Isogai et al 1980)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1109	CT	ISOY800108	Normalized relative frequency of coil (Isogai et al 1980)
1110	CT	JANJ790101	Ratio of buried and accessible molar fractions (Janin 1979)
1111	CT	JANJ790102	Transfer free energy (Janin 1979)
1112	CT	JOND920101	Relative frequency of occurrence (Jones et al 1992)
1113	CT	KANM800104	Average relative probability of inner beta-sheet (Kanehisa-Tsong 1980)
1114	CT	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz 1985)
1115	CT	KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz 1985)
1116	CT	KHAG800101	The Kerr-constant increments (Khanarian-Moore 1980)
1117	CT	KLEP840101	Net charge (Klein et al 1984)
1118	CT	KOEP990101	Alpha-helix propensity derived from designed sequences (Koehl-Levitt 1999)
1119	CT	KOEP990102	Beta-sheet propensity derived from designed sequences (Koehl-Levitt 1999)
1120	CT	KRIW710101	Side chain interaction parameter (Krigbaum-Rubin 1971)
1121	CT	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya 1979)
1122	CT	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al 2000)
1123	CT	KUMS000103	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al 2000)
1124	CT	LAWE840101	Transfer free energy CHPwater (Lawson et al 1984)
1125	CT	LEVM760103	Side chain angle theta(AAR) (Levitt 1976)
1126	CT	LEVM760106	van der Waals parameter R0 (Levitt 1976)
1127	CT	LEVM780102	Normalized frequency of beta-sheet with weights (Levitt 1978)
1128	CT	LEVM780103	Normalized frequency of reverse turn with weights (Levitt 1978)
1129	CT	LEWP710101	Frequency of occurrence in beta-bends (Lewis et al 1971)
1130	CT	LIFS790102	Conformational preference for parallel beta-strands (Lifson-Sander 1979)
1131	CT	LIFS790103	Conformational preference for antiparallel beta-strands (Lifson-Sander 1979)
1132	CT	MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga 1976)
1133	CT	MAXF760105	Normalized frequency of zeta L (Maxfield-Scheraga 1976)
1134	CT	MCMT640101	Refractivity (McMeekin et al 1964) Cited by Jones (1975)
1135	CT	MEEJ800101	Retention coefficient in HPLC pH74 (Meek 1980)
1136	CT	MEEJ810101	Retention coefficient in NaClO4 (Meek-Rossetti 1981)
1137	CT	MEIH800101	Average reduced distance for C-alpha (Meirovitch et al 1980)
1138	CT	MEIH800103	Average side chain orientation angle (Meirovitch et al 1980)
1139	CT	MITS020101	Amphiphilicity index (Mitaku et al 2002)
1140	CT	MIYS850101	Effective partition energy (Miyazawa-Jernigan 1985)
1141	CT	MONM990101	Turn propensity scale for transmembrane helices (Monne et al 1999)
1142	CT	NADH010101	Hydropathy scale based on self-information values in the two-state model (5percent accessibility) (Naderi-Manesh et al 2001)
1143	CT	NADH010103	Hydropathy scale based on self-information values in the two-state model (16percent accessibility) (Naderi-Manesh et al 2001)
1144	CT	NADH010106	Hydropathy scale based on self-information values in the two-state model (36percent accessibility) (Naderi-Manesh et al 2001)
1145	CT	NADH010107	Hydropathy scale based on self-information values in the two-state model (50percent accessibility) (Naderi-Manesh et al 2001)
1146	CT	NAGK730103	Normalized frequency of coil (Nagano 1973)
1147	CT	NAKH900102	SD of AA composition of total proteins (Nakashima et al 1990)
1148	CT	NAKH900103	AA composition of mt-proteins (Nakashima et al 1990)
1149	CT	NAKH900104	Normalized composition of mt-proteins (Nakashima et al 1990)
1150	CT	NAKH900109	AA composition of membrane proteins (Nakashima et al 1990)
1151	CT	NAKH900110	Normalized composition of membrane proteins (Nakashima et al 1990)
1152	CT	NAKH900111	Transmembrane regions of non-mt-proteins (Nakashima et al 1990)
1153	CT	NAKH900113	Ratio of average and computed composition (Nakashima et al 1990)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1154	CT	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa 1992)
1155	CT	NAKH920103	AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa 1992)
1156	CT	NISK800101	8 A contact number (Nishikawa-Ooi 1980)
1157	CT	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi 1977)
1158	CT	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi 1977)
1159	CT	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi 1977)
1160	CT	OOBM770104	Average non-bonded energy per residue (Oobatake-Ooi 1977)
1161	CT	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al 1985)
1162	CT	OOBM850103	Optimized transfer energy parameter (Oobatake et al 1985)
1163	CT	OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al 1985)
1164	CT	OOBM850105	Optimized side chain interaction parameter (Oobatake et al 1985)
1165	CT	PALJ810105	Normalized frequency of turn from LG (Palau et al 1981)
1166	CT	PALJ810108	Normalized frequency of alpha-helix in alpha-plusbeta class (Palau et al 1981)
1167	CT	PALJ810111	Normalized frequency of beta-sheet in alpha-plusbeta class (Palau et al 1981)
1168	CT	PALJ810113	Normalized frequency of turn in all-alpha class (Palau et al 1981)
1169	CT	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al 1981)
1170	CT	PALJ810115	Normalized frequency of turn in alphaplusbeta class (Palau et al 1981)
1171	CT	PALJ810116	Normalized frequency of turn in alpha-beta class (Palau et al 1981)
1172	CT	PARS000101	p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
1173	CT	PARS000102	p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy 2000)
1174	CT	PLIV810101	Partition coefficient (Pliska et al 1981)
1175	CT	PONP800101	Surrounding hydrophobicity in folded form (Ponnuswamy et al 1980)
1176	CT	PONP800104	Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al 1980)
1177	CT	PONP800105	Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al 1980)
1178	CT	PONP800106	Surrounding hydrophobicity in turn (Ponnuswamy et al 1980)
1179	CT	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy 1982)
1180	CT	PRAM820102	Slope in regression analysis x 10E1 (Prabhakaran-Ponnuswamy 1982)
1181	CT	PRAM900101	Hydrophobicity (Prabhakaran 1990)
1182	CT	LEVM780101	Normalized frequency of alpha-helix with weights (Levitt 1978)
1183	CT	PTIO830101	Helix-coil equilibrium constant (Ptitsyn-Finkelstein 1983)
1184	CT	QIAN880101	Weights for alpha-helix at the window position of -6 (Qian-Sejnowski 1988)
1185	CT	QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski 1988)
1186	CT	QIAN880103	Weights for alpha-helix at the window position of -4 (Qian-Sejnowski 1988)
1187	CT	QIAN880104	Weights for alpha-helix at the window position of -3 (Qian-Sejnowski 1988)
1188	CT	QIAN880107	Weights for alpha-helix at the window position of 0 (Qian-Sejnowski 1988)
1189	CT	QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski 1988)
1190	CT	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski 1988)
1191	CT	QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski 1988)
1192	CT	QIAN880116	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski 1988)
1193	CT	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski 1988)
1194	CT	QIAN880118	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski 1988)
1195	CT	QIAN880121	Weights for beta-sheet at the window position of 1 (Qian-Sejnowski 1988)
1196	CT	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski 1988)
1197	CT	QIAN880123	Weights for beta-sheet at the window position of 3 (Qian-Sejnowski 1988)
1198	CT	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski 1988)
1199	CT	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski 1988)
1200	CT	QIAN880128	Weights for coil at the window position of -5 (Qian-Sejnowski 1988)
1201	CT	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski 1988)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1202	CT	QIAN880130	Weights for coil at the window position of -3 (Qian-Sejnowski 1988)
1203	CT	QIAN880131	Weights for coil at the window position of -2 (Qian-Sejnowski 1988)
1204	CT	QIAN880135	Weights for coil at the window position of 2 (Qian-Sejnowski 1988)
1205	CT	QIAN880136	Weights for coil at the window position of 3 (Qian-Sejnowski 1988)
1206	CT	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski 1988)
1207	CT	QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski 1988)
1208	CT	QIAN880139	Weights for coil at the window position of 6 (Qian-Sejnowski 1988)
1209	CT	RACS770103	Side chain orientational preference (Rackovsky-Scheraga 1977)
1210	CT	RACS820101	Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga 1982)
1211	CT	RACS820102	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga 1982)
1212	CT	RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga 1982)
1213	CT	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga 1982)
1214	CT	RACS820105	Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga 1982)
1215	CT	RACS820106	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga 1982)
1216	CT	RACS820107	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga 1982)
1217	CT	RACS820108	Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga 1982)
1218	CT	RACS820110	Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga 1982)
1219	CT	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga 1982)
1220	CT	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga 1982)
1221	CT	RACS820113	Value of theta(i) (Rackovsky-Scheraga 1982)
1222	CT	RACS820114	Value of theta(i-1) (Rackovsky-Scheraga 1982)
1223	CT	RADA880103	Transfer free energy from vap to chx (Radzicka-Wolfenden 1988)
1224	CT	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden 1988)
1225	CT	RADA880106	Accessible surface area (Radzicka-Wolfenden 1988)
1226	CT	RICJ880101	Relative preference value at N" (Richardson-Richardson 1988)
1227	CT	RICJ880103	Relative preference value at N-cap (Richardson-Richardson 1988)
1228	CT	RICJ880104	Relative preference value at N1 (Richardson-Richardson 1988)
1229	CT	RICJ880105	Relative preference value at N2 (Richardson-Richardson 1988)
1230	CT	RICJ880107	Relative preference value at N4 (Richardson-Richardson 1988)
1231	CT	RICJ880108	Relative preference value at N5 (Richardson-Richardson 1988)
1232	CT	RICJ880109	Relative preference value at Mid (Richardson-Richardson 1988)
1233	CT	RICJ880110	Relative preference value at C5 (Richardson-Richardson 1988)
1234	CT	RICJ880111	Relative preference value at C4 (Richardson-Richardson 1988)
1235	CT	RICJ880112	Relative preference value at C3 (Richardson-Richardson 1988)
1236	CT	RICJ880113	Relative preference value at C2 (Richardson-Richardson 1988)
1237	CT	RICJ880114	Relative preference value at C1 (Richardson-Richardson 1988)
1238	CT	RICJ880116	Relative preference value at C' (Richardson-Richardson 1988)
1239	CT	RICJ880117	Relative preference value at C" (Richardson-Richardson 1988)
1240	CT	ROBB760107	Information measure for extended without H-bond (Robson-Suzuki 1976)
1241	CT	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki 1976)
1242	CT	ROBB790101	Hydration free energy (Robson-Osguthorpe 1979)
1243	CT	ROSM880102	Side chain hydropathy corrected for solvation (Roseman 1988)
1244	CT	ROSM880103	Loss of Side chain hydropathy by helix formation (Roseman 1988)
1245	CT	SNEP660101	Principal component I (Sneath 1966)
1246	CT	SNEP660102	Principal component II (Sneath 1966)
1247	CT	SNEP660103	Principal component III (Sneath 1966)
1248	CT	SNEP660104	Principal component IV (Sneath 1966)
1249	CT	SUEM840102	Zimm-Bragg parameter sigma x 10E4 (Sueki et al 1984)
1250	CT	SUYM030101	Linker propensity index (Suyama-Ohara 2003)
1251	CT	SWER830101	Optimal matching hydrophobicity (Sweet-Eisenberg 1983)

Table B.1: Chapter 6 Global and Regional Features Continued...

Number	Region	ID	Description
1252	CT	TAKK010101	Side-chain contribution to protein stability (kJ-mol) (Takano-Yutani 2001)
1253	CT	TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga 1977)
1254	CT	TANS770106	Normalized frequency of chain reversal D (Tanaka-Scheraga 1977)
1255	CT	TANS770107	Normalized frequency of left-handed helix (Tanaka-Scheraga 1977)
1256	CT	TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga 1977)
1257	CT	VASM830101	Relative population of conformational state A (Vasquez et al 1983)
1258	CT	VASM830102	Relative population of conformational state C (Vasquez et al 1983)
1259	CT	VASM830103	Relative population of conformational state E (Vasquez et al 1983)
1260	CT	VELV850101	Electron-ion interaction potential (Veljkovic et al 1985)
1261	CT	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al 1994)
1262	CT	WARP780101	Average interactions per side chain atom (Warme-Morgan 1978)
1263	CT	WEBA780101	RF value in high salt chromatography (Weber-Lacey 1978)
1264	CT	WERD780102	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga 1978)
1265	CT	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga 1978)
1266	CT	WERD780104	Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga 1978)
1267	CT	WILM950101	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
1268	CT	WILM950102	Hydrophobicity coefficient in RP-HPLC C8 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
1269	CT	WILM950103	Hydrophobicity coefficient in RP-HPLC C4 with 01percentTFA-MeCN-H2O (Wilce et al 1995)
1270	CT	WILM950104	Hydrophobicity coefficient in RP-HPLC C18 with 01percentTFA-2-PrOH-MeCN-H2O (Wilce et al 1995)
1271	CT	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White 1996)
1272	CT	WOLS870102	Principal property value z2 (Wold et al 1987)
1273	CT	WOLS870103	Principal property value z3 (Wold et al 1987)
1274	CT	YUTK870101	Unfolding Gibbs energy in water pH70 (Yutani et al 1987)
1275	CT	YUTK870103	Activation Gibbs energy of unfolding pH70 (Yutani et al 1987)
1276	CT	ZIMJ680101	Hydrophobicity (Zimmerman et al 1968)

Chapter C: EFC Features

Shown are EFC features introduced in Chapters 5 and 7. The letters **A**,**C**,**G** and **T** refer to the GBMR4 alphabet described in Chapter 5 Table 5.1. EFC position numbers use 0 as the first AA in a sequence. For example, EFC feature EFC_9 below "matchesAtPosition (motif2 g c) 1" means the motif **GC** is found starting with the second AA in a sequence.

Table C.1: Chapter 5 EFC Features

Feature	**Definition**
EFC_1	(AND (matchesAtPosition (motif2 g t) 59) (matches (motif3 c a t))
EFC_2	(AND (matches (motif2 c a) (shiftPosition (motif4 t c t c) 78 ±3))
EFC_3	(AND (matches (motif2 c g) (matches (motif7 c g t a t a a))
EFC_4	(AND (matches (motif2 g a) (matches (motif5 g t c g c))
EFC_5	(AND (matches (motif4 a a a a) (matches (motif4 t a c a))
EFC_6	(AND (matches (motif5 t t c c c) (shiftPosition (motif3 g t t) 17 ±3))
EFC_7	(correlate (motif2 a t) (motif2 t c) 1 ±3)
EFC_8	(correlate (motif5 g a c g c) (motif2 a c) 17 ±3)
EFC_9	(matchesAtPosition (motif2 g c) 1)
EFC_10	(matchesAtPosition (motif2 g c) 18)
EFC_11	(matchesAtPosition (motif2 t c) 3)
EFC_12	(matchesAtPosition (motif2 t g) 13)
EFC_13	(matchesAtPosition (motif3 c t c) 2)
EFC_14	(matchesAtPosition (motif3 g g c) 27)
EFC_15	(matchesAtPosition (motif4 c g t c) 17)
EFC_16	(matchesAtPosition (motif4 t t a t) 12)
EFC_17	(matchesAtPosition (motif5 c a a g g) 33)
EFC_18	(matches (motif5 g t c g c)
EFC_19	(matches (motif5 t a t t t)
EFC_20	(matches (motif5 t g c c g)
EFC_21	(matches (motif5 t g g g a)
EFC_22	(matches (motif6 a a a t t c)
EFC_23	(matches (motif6 a c g g c g)
EFC_24	(matches (motif7 a c c c c t g)
EFC_25	(matches (motif7 a g c c a c t)
EFC_26	(matches (motif7 c g a g c a c)
EFC_27	(matches (motif7 g g g a c g g)
EFC_28	(matches (motif8 c g a c c g c c)
EFC_29	(matches (motif8 c t c a c a t c)
EFC_30	(matches (motif8 g a a g g a c a)
EFC_31	(shiftPosition (motif2 c g) 5 ±3)
EFC_32	(shiftPosition (motif2 g a) 11 ±3)
EFC_33	(shiftPosition (motif2 g c) 0 ±3)

Table C.1: Chapter 5 EFC Features Continued...

Feature	Definition
EFC_34	(shiftPosition (motif3 a g c) 8 ±3)
EFC_35	(shiftPosition (motif3 a g c) 9 ±3)
EFC_36	(shiftPosition (motif3 c g a) 44 ±3)
EFC_37	(shiftPosition (motif3 g g c) 10 ±3)
EFC_38	(shiftPosition (motif3 g g c) 11 ±3)
EFC_39	(shiftPosition (motif3 g g c) 13 ±3)
EFC_40	(shiftPosition (motif3 g t t) 16 ±3)
EFC_41	(shiftPosition (motif4 a g c c) 15 ±3)
EFC_42	(shiftPosition (motif4 c c a a) 5 ±3)
EFC_43	(shiftPosition (motif4 g g g g) 4 ±3)
EFC_44	(shiftPosition (motif5 c t t a t) 14 ±3)
EFC_45	(shiftPosition (motif5 g c c c a) 0 ±3)
EFC_46	(shiftPosition (motif6 t c a g t c) 5 ±3)
EFC_47	(shiftPosition (motif7 c a c g c c g) 7 ±3)
EFC_48	(AND (matches (motif3 a a a) (matches (motif3 c g t))
EFC_49	(AND (matches (motif3 a t t) (matches (motif3 a a a))
EFC_50	(AND (matches (motif3 t t g) (matches (motif4 g c g t))
EFC_51	(AND (matches (motif5 a g g a a) (matchesAtPosition (motif2 c t) 51))
EFC_52	(AND (shiftPosition (motif4 c c g c) 42 3) (shiftPosition (motif2 t g) 50 ±3))
EFC_53	(correlate (motif2 c c) (motif4 g c g a) 19 ±3)
EFC_54	(correlate (motif2 c g) (motif2 g a) 20 ±3)
EFC_55	(correlate (motif2 g a) (motif3 t a c) 22 ±3)
EFC_56	(correlate (motif2 t c) (motif4 g t c t) 8 ±3)
EFC_57	(correlate (motif2 t t) (motif2 g t) 37 ±3)
EFC_58	(matchesAtPosition (motif2 a g) 9)
EFC_59	(matchesAtPosition (motif2 c a) 58)
EFC_60	(matchesAtPosition (motif2 g c) 1)
EFC_61	(matchesAtPosition (motif2 g c) 8)
EFC_62	(matchesAtPosition (motif2 t c) 3)
EFC_63	(matchesAtPosition (motif3 c a c) 3)
EFC_64	(matchesAtPosition (motif3 c t c) 2)
EFC_65	(matchesAtPosition (motif3 g a c) 10)
EFC_66	(matchesAtPosition (motif3 t c c) 3)
EFC_67	(matchesAtPosition (motif4 a c g a) 23)
EFC_68	(matchesAtPosition (motif4 a g c c) 1)
EFC_69	(matchesAtPosition (motif4 c a t c) 10)
EFC_70	(matchesAtPosition (motif4 c t c g) 42)
EFC_71	(matchesAtPosition (motif4 c t g g) 41)
EFC_72	(matchesAtPosition (motif4 g c g t) 24)
EFC_73	(matchesAtPosition (motif4 t t a t) 11)
EFC_74	(matchesAtPosition (motif5 g c a a c) 18)
EFC_75	(matchesAtPosition (motif5 t a a c a) 80)
EFC_76	(matchesAtPosition (motif7 a a a a t c a) 11)
EFC_77	(matchesAtPosition (motif7 a a t a a a c) 52)
EFC_78	(matchesAtPosition (motif7 g c c c a g c) 2)
EFC_79	(matches (motif5 g g t c g)
EFC_80	(matches (motif5 t a t t t)
EFC_81	(matches (motif5 t t g c g)
EFC_82	(matches (motif6 a t a t t c)
EFC_83	(matches (motif6 c g g g g c)

Table C.1: Chapter 5 EFC Features Continued...

Feature	Definition
EFC_84	(matches (motif6 c t a t t t)
EFC_85	(matches (motif6 g a g g c g)
EFC_86	(matches (motif6 g g g g g c)
EFC_87	(matches (motif6 g g g g g g)
EFC_88	(matches (motif6 g t c a g t)
EFC_89	(matches (motif7 a a a c g t t)
EFC_90	(matches (motif7 a a t c c g t)
EFC_91	(matches (motif7 a g c g c g a)
EFC_92	(matches (motif7 a g g a a c g)
EFC_93	(matches (motif7 a t a t t a t)
EFC_94	(matches (motif7 c t c t a t g)
EFC_95	(matches (motif7 g a a g g a c)
EFC_96	(matches (motif7 g g a t a c t)
EFC_97	(matches (motif7 g g g g g a g)
EFC_98	(matches (motif7 t a t a t t a)
EFC_99	(matches (motif7 t t c a t t c)
EFC_100	(matches (motif8 c c a g t a c a)
EFC_101	(matches (motif8 g a c c a t g a)
EFC_102	(matches (motif8 t a c a t a t c)
EFC_103	(matches (motif8 t a t t c t a c)
EFC_104	(matches (motif8 t c a c a g c t)
EFC_105	(matches (motif8 t g c g t t c c)
EFC_106	(matches (motif8 t t a a a t a c)
EFC_107	(shiftPosition (motif3 c g a) 10 ±3)
EFC_108	(shiftPosition (motif3 g c a) 9 ±3)
EFC_109	(shiftPosition (motif3 g c c) 13 ±3)
EFC_110	(shiftPosition (motif3 g g c) 13 ±3)
EFC_111	(shiftPosition (motif3 g g g) 14 ±3)
EFC_112	(shiftPosition (motif4 a c g a) 10 ±3)
EFC_113	(shiftPosition (motif4 a g c c) 13 ±3)
EFC_114	(shiftPosition (motif4 a g c c) 16 ±3)
EFC_115	(shiftPosition (motif4 a g g c) 16 ±3)
EFC_116	(shiftPosition (motif4 a g t g) 60 ±3)
EFC_117	(shiftPosition (motif4 c a g t) 54 ±3)
EFC_118	(shiftPosition (motif4 c g g g) 9 ±3)
EFC_119	(shiftPosition (motif4 t a t g) 32 ±3)
EFC_120	(shiftPosition (motif4 t t t t) 37 ±3)
EFC_121	(shiftPosition (motif5 a c c a c) 42 ±3)
EFC_122	(shiftPosition (motif5 a c g t a) 0 ±3)
EFC_123	(shiftPosition (motif5 c a t t a) 50 ±3)
EFC_124	(shiftPosition (motif5 g c a a c) 6 ±3)
EFC_125	(shiftPosition (motif5 t a g c g) 51 ±3)
EFC_126	(shiftPosition (motif5 t c t a c) 12 ±3)
EFC_127	(shiftPosition (motif5 t g g a g) 39 ±3)
EFC_128	(shiftPosition (motif5 t t c c t) 22 ±3)
EFC_129	(shiftPosition (motif5 t t t a t) 13 ±3)
EFC_130	(shiftPosition (motif7 a c c t a c g) 3 ±3)
EFC_131	(shiftPosition (motif8 a c t a a c g c) 67 ±3)
EFC_132	(shiftPosition (motif8 a t t t a t a a) 9 ±3)
EFC_133	(AND (matches (motif3 c g a) (shiftPosition (motif2 a c) 8 ±3))

Table C.1: Chapter 5 EFC Features Continued...

Feature	Definition
EFC_134	(AND (matches (motif4 c a c c) (matches (motif2 t a))
EFC_135	(correlate (motif2 a g) (motif3 a t g) 70 ±3)
EFC_136	(correlate (motif2 g t) (motif2 a g) 5 ±3)
EFC_137	(correlate (motif2 t g) (motif4 c c a c) 13 ±3)
EFC_138	(correlate (motif4 a g g a) (motif2 g c) 30 ±3)
EFC_139	(correlate (motif5 c a a t g) (motif2 c c) 62 ±3)
EFC_140	(matchesAtPosition (motif2 c g) 2)
EFC_141	(matchesAtPosition (motif2 g a) 10)
EFC_142	(matchesAtPosition (motif2 g c) 18)
EFC_143	(matchesAtPosition (motif2 g g) 17)
EFC_144	(matchesAtPosition (motif2 t g) 9)
EFC_145	(matchesAtPosition (motif2 t t) 19)
EFC_146	(matchesAtPosition (motif3 a t g) 8)
EFC_147	(matchesAtPosition (motif3 c c a) 42)
EFC_148	(matchesAtPosition (motif3 c c a) 5)
EFC_149	(matchesAtPosition (motif3 c c a) 78)
EFC_150	(matchesAtPosition (motif3 c t c) 2)
EFC_151	(matchesAtPosition (motif3 g t c) 5)
EFC_152	(matchesAtPosition (motif3 t a t) 13)
EFC_153	(matchesAtPosition (motif3 t c c) 3)
EFC_154	(matchesAtPosition (motif3 t g a) 13)
EFC_155	(matchesAtPosition (motif3 t g t) 22)
EFC_156	(matchesAtPosition (motif3 t g t) 32)
EFC_157	(matchesAtPosition (motif4 a t t c) 11)
EFC_158	(matchesAtPosition (motif4 c c g g) 9)
EFC_159	(matchesAtPosition (motif4 c c t a) 12)
EFC_160	(matchesAtPosition (motif4 g g a g) 2)
EFC_161	(matchesAtPosition (motif4 t g c a) 5)
EFC_162	(matchesAtPosition (motif5 c t a t c) 4)
EFC_163	(matchesAtPosition (motif5 t t c c a) 4)
EFC_164	(matchesAtPosition (motif6 c c a c g a) 19)
EFC_165	(matchesAtPosition (motif6 g a a t a c) 52)
EFC_166	(matchesAtPosition (motif7 c g a c c a c) 3)
EFC_167	(matchesAtPosition (motif7 g a c g a a g) 12)
EFC_168	(matchesAtPosition (motif8 a c a a c c c g) 13)
EFC_169	(matches (motif5 g t c g c)
EFC_170	(matches (motif6 c g t c g c)
EFC_171	(matches (motif6 t c t a t t)
EFC_172	(matches (motif6 t c t c a t)
EFC_173	(matches (motif6 t g a c c g)
EFC_174	(matches (motif7 g c a a t c t)
EFC_175	(matches (motif7 g t a c a c a)
EFC_176	(matches (motif8 t c a c g g c c)
EFC_177	(shiftPosition (motif2 a t) 38 ±3)
EFC_178	(shiftPosition (motif2 c g) 73 ±3)
EFC_179	(shiftPosition (motif2 c g) 74 ±3)
EFC_180	(shiftPosition (motif2 c t) 62 ±3)
EFC_181	(shiftPosition (motif3 a g c) 3 ±3)
EFC_182	(shiftPosition (motif3 c a g) 8 ±3)
EFC_183	(shiftPosition (motif3 c g a) 12 ±3)

Table C.1: Chapter 5 EFC Features Continued...

Feature	Definition
EFC_184	(shiftPosition (motif3 c g a) 13 ±3)
EFC_185	(shiftPosition (motif3 c g a) 8 ±3)
EFC_186	(shiftPosition (motif3 c g t) 9 ±3)
EFC_187	(shiftPosition (motif3 g a c) 60 ±3)
EFC_188	(shiftPosition (motif3 g c c) 9 ±3)
EFC_189	(shiftPosition (motif3 g g c) 28 ±3)
EFC_190	(shiftPosition (motif3 g g c) 29 ±3)
EFC_191	(shiftPosition (motif3 t g t) 35 ±3)
EFC_192	(shiftPosition (motif4 a a a a) 33 ±3)
EFC_193	(shiftPosition (motif4 a g c c) 8 ±3)
EFC_194	(shiftPosition (motif4 c t c c) 4 ±3)
EFC_195	(shiftPosition (motif4 g c c a) 2 ±3)
EFC_196	(shiftPosition (motif4 g g a t) 61 ±3)
EFC_197	(shiftPosition (motif4 g g c c) 25 ±3)
EFC_198	(shiftPosition (motif4 g t c g) 21 ±3)
EFC_199	(shiftPosition (motif4 t t a t) 11 ±3)
EFC_200	(shiftPosition (motif5 t t a t t) 7 ±3)
EFC_201	(shiftPosition (motif6 g a c a c g) 0 ±3)
EFC_202	(shiftPosition (motif6 g c c t g a) 1 ±3)
EFC_203	(shiftPosition (motif6 t a t c a t) 10 ±3)
EFC_204	(shiftPosition (motif6 t g c c a a) 44 ±3)
EFC_205	(shiftPosition (motif7 a t a a t c c) 10 ±3)
EFC_206	(shiftPosition (motif7 c c t c c a a) 1 ±3)
EFC_207	(shiftPosition (motif7 g c t a g c c) 12 ±3)
EFC_208	(shiftPosition (motif8 c g t c a c a c) 8 ±3)

Table C.2: Chapter 7 Gram-Negative EFC Features

Feature	Definition
GN_EFC_1	(AND (matches (motif3 a a a) (matches (motif3 c g t))
GN_EFC_2	(AND (matches (motif3 a t t) (matches (motif3 a a a))
GN_EFC_3	(AND (matches (motif3 t t g) (matches (motif4 g c g t))
GN_EFC_4	(AND (matches (motif5 a g g a a) (matchesAtPosition (motif2 c t) @ 51))
GN_EFC_5	(AND (shiftPosition (motif4 c c g c) @ 42 3) (shiftPosition (motif2 t g) @ 50 ±3))
GN_EFC_6	(correlate (motif2 c c) (motif4 g c g a) @ 19 ±3)
GN_EFC_7	(correlate (motif2 c g) (motif2 g a) @ 20 ±3)
GN_EFC_8	(correlate (motif2 g a) (motif3 t a c) @ 22 ±3)
GN_EFC_9	(correlate (motif2 t c) (motif4 g t c t) @ 8 ±3)
GN_EFC_10	(correlate (motif2 t t) (motif2 g t) @ 37 ±3)
GN_EFC_11	(matchesAtPosition (motif2 a g) @ 9)
GN_EFC_12	(matchesAtPosition (motif2 c a) @ 58)
GN_EFC_13	(matchesAtPosition (motif2 g c) @ 1)
GN_EFC_14	(matchesAtPosition (motif2 g c) @ 8)
GN_EFC_15	(matchesAtPosition (motif2 t c) @ 3)
GN_EFC_16	(matchesAtPosition (motif3 c a c) @ 3)
GN_EFC_17	(matchesAtPosition (motif3 c t c) @ 2)
GN_EFC_18	(matchesAtPosition (motif3 g a c) @ 10)
GN_EFC_19	(matchesAtPosition (motif3 t c c) @ 3)
GN_EFC_20	(matchesAtPosition (motif4 a c g a) @ 23)

Table C.2: Chapter 7 Gram-Negative EFC Features Continued...

Feature	Definition
GN_EFC_21	(matchesAtPosition (motif4 a g c c) @ 1)
GN_EFC_22	(matchesAtPosition (motif4 c a t c) @ 10)
GN_EFC_23	(matchesAtPosition (motif4 c t c g) @ 42)
GN_EFC_24	(matchesAtPosition (motif4 c t g g) @ 41)
GN_EFC_25	(matchesAtPosition (motif4 g c g t) @ 24)
GN_EFC_26	(matchesAtPosition (motif4 t t a t) @ 11)
GN_EFC_27	(matchesAtPosition (motif5 g c a a c) @ 18)
GN_EFC_28	(matchesAtPosition (motif5 t a a c a) @ 80)
GN_EFC_29	(matchesAtPosition (motif7 a a a a t c a) @ 11)
GN_EFC_30	(matchesAtPosition (motif7 a a t a a a c) @ 52)
GN_EFC_31	(matchesAtPosition (motif7 g c c c a g c) @ 2)
GN_EFC_32	(matches (motif5 g g t c g)
GN_EFC_33	(matches (motif5 t a t t t)
GN_EFC_34	(matches (motif5 t t g c g)
GN_EFC_35	(matches (motif6 a t a t t c)
GN_EFC_36	(matches (motif6 c g g g g c)
GN_EFC_37	(matches (motif6 c t a t t t)
GN_EFC_38	(matches (motif6 g a g g c g)
GN_EFC_39	(matches (motif6 g g g g g c)
GN_EFC_40	(matches (motif6 g g g g g g)
GN_EFC_41	(matches (motif6 g t c a g t)
GN_EFC_42	(matches (motif7 a a a c g t t)
GN_EFC_43	(matches (motif7 a a t c c g t)
GN_EFC_44	(matches (motif7 a g c g c g a)
GN_EFC_45	(matches (motif7 a g g a a c g)
GN_EFC_46	(matches (motif7 a t a t t a t)
GN_EFC_47	(matches (motif7 c t c t a t g)
GN_EFC_48	(matches (motif7 g a a g g a c)
GN_EFC_49	(matches (motif7 g g a t a c t)
GN_EFC_50	(matches (motif7 g g g g g a g)
GN_EFC_51	(matches (motif7 t a t a t t a)
GN_EFC_52	(matches (motif7 t t c a t t c)
GN_EFC_53	(matches (motif8 c c a g t a c a)
GN_EFC_54	(matches (motif8 g a c c a t g a)
GN_EFC_55	(matches (motif8 t a c a t a t c)
GN_EFC_56	(matches (motif8 t a t t c t a c)
GN_EFC_57	(matches (motif8 t c a c a g c t)
GN_EFC_58	(matches (motif8 t g c g t t c c)
GN_EFC_59	(matches (motif8 t t a a a t a c)
GN_EFC_60	(shiftPosition (motif3 c g a) @ 10 ±3)
GN_EFC_61	(shiftPosition (motif3 g c a) @ 9 ±3)
GN_EFC_62	(shiftPosition (motif3 g c c) @ 13 ±3)
GN_EFC_63	(shiftPosition (motif3 g g c) @ 13 ±3)
GN_EFC_64	(shiftPosition (motif3 g g g) @ 14 ±3)
GN_EFC_65	(shiftPosition (motif4 a c g a) @ 10 ±3)
GN_EFC_66	(shiftPosition (motif4 a g c c) @ 13 ±3)
GN_EFC_67	(shiftPosition (motif4 a g c c) @ 16 ±3)
GN_EFC_68	(shiftPosition (motif4 a g g c) @ 16 ±3)
GN_EFC_69	(shiftPosition (motif4 a g t g) @ 60 ±3)
GN_EFC_70	(shiftPosition (motif4 c a g t) @ 54 ±3)

Table C.2: Chapter 7 Gram-Negative EFC Features Continued...

Feature	Definition
GN_EFC_71	(shiftPosition (motif4 c g g g) @ 9 ±3)
GN_EFC_72	(shiftPosition (motif4 t a t g) @ 32 ±3)
GN_EFC_73	(shiftPosition (motif4 t t t t) @ 37 ±3)
GN_EFC_74	(shiftPosition (motif5 a c c a c) @ 42 ±3)
GN_EFC_75	(shiftPosition (motif5 a c g t a) @ 0 ±3)
GN_EFC_76	(shiftPosition (motif5 c a t t a) @ 50 ±3)
GN_EFC_77	(shiftPosition (motif5 g c a a c) @ 6 ±3)
GN_EFC_78	(shiftPosition (motif5 t a g c g) @ 51 ±3)
GN_EFC_79	(shiftPosition (motif5 t c t a c) @ 12 ±3)
GN_EFC_80	(shiftPosition (motif5 t g g a g) @ 39 ±3)
GN_EFC_81	(shiftPosition (motif5 t t c c t) @ 22 ±3)
GN_EFC_82	(shiftPosition (motif5 t t t a t) @ 13 ±3)
GN_EFC_83	(shiftPosition (motif7 a c c t a c g) @ 3 ±3)
GN_EFC_84	(shiftPosition (motif8 a c t a a c g c) @ 67 ±3)
GN_EFC_85	(shiftPosition (motif8 a t t t a t a a) @ 9 ±3)

Table C.3: Chapter 7 Gram-Positive EFC Features

Feature	Definition
GP_EFC_1	(AND (matches (motif3 c g a) (shiftPosition (motif2 a c) @ 8 ±3))
GP_EFC_2	(AND (matches (motif4 c a c c) (matches (motif2 t a))
GP_EFC_3	(correlate (motif2 a g) (motif3 a t g) @ 70 ±3)
GP_EFC_4	(correlate (motif2 g t) (motif2 a g) @ 5 ±3)
GP_EFC_5	(correlate (motif2 t g) (motif4 c c a c) @ 13 ±3)
GP_EFC_6	(correlate (motif4 a g g a) (motif2 g c) @ 30 ±3)
GP_EFC_7	(correlate (motif5 c a a t g) (motif2 c c) @ 62 ±3)
GP_EFC_8	(matchesAtPosition (motif2 c g) @ 2)
GP_EFC_9	(matchesAtPosition (motif2 g a) @ 10)
GP_EFC_10	(matchesAtPosition (motif2 g c) @ 18)
GP_EFC_11	(matchesAtPosition (motif2 g g) @ 17)
GP_EFC_12	(matchesAtPosition (motif2 t g) @ 9)
GP_EFC_13	(matchesAtPosition (motif2 t t) @ 19)
GP_EFC_14	(matchesAtPosition (motif3 a t g) @ 8)
GP_EFC_15	(matchesAtPosition (motif3 c c a) @ 42)
GP_EFC_16	(matchesAtPosition (motif3 c c a) @ 5)
GP_EFC_17	(matchesAtPosition (motif3 c c a) @ 78)
GP_EFC_18	(matchesAtPosition (motif3 c t c) @ 2)
GP_EFC_19	(matchesAtPosition (motif3 g t c) @ 5)
GP_EFC_20	(matchesAtPosition (motif3 t a t) @ 13)
GP_EFC_21	(matchesAtPosition (motif3 t c c) @ 3)
GP_EFC_22	(matchesAtPosition (motif3 t g a) @ 13)
GP_EFC_23	(matchesAtPosition (motif3 t g t) @ 22)
GP_EFC_24	(matchesAtPosition (motif3 t g t) @ 32)
GP_EFC_25	(matchesAtPosition (motif4 a t t c) @ 11)
GP_EFC_26	(matchesAtPosition (motif4 c c g g) @ 9)
GP_EFC_27	(matchesAtPosition (motif4 c c t a) @ 12)
GP_EFC_28	(matchesAtPosition (motif4 g g a g) @ 2)
GP_EFC_29	(matchesAtPosition (motif4 t g c a) @ 5)
GP_EFC_30	(matchesAtPosition (motif5 c t a t c) @ 4)

Table C.3: Chapter 7 Gram-Positive EFC Features Continued...

Feature	Definition
GP_EFC_31	(matchesAtPosition (motif5 t t c c a) @ 4)
GP_EFC_32	(matchesAtPosition (motif6 c c a c g a) @ 19)
GP_EFC_33	(matchesAtPosition (motif6 g a a t a c) @ 52)
GP_EFC_34	(matchesAtPosition (motif7 c g a c c a c) @ 3)
GP_EFC_35	(matchesAtPosition (motif7 g a c g a a g) @ 12)
GP_EFC_36	(matchesAtPosition (motif8 a c a a c c c g) @ 13)
GP_EFC_37	(matches (motif5 g t c g c)
GP_EFC_38	(matches (motif6 c g t c g c)
GP_EFC_39	(matches (motif6 t c t a t t)
GP_EFC_40	(matches (motif6 t c t c a t)
GP_EFC_41	(matches (motif6 t g a c c g)
GP_EFC_42	(matches (motif7 g c a a t c t)
GP_EFC_43	(matches (motif7 g t a c a c a)
GP_EFC_44	(matches (motif8 t c a c g g c c)
GP_EFC_45	(shiftPosition (motif2 a t) @ 38 ±3)
GP_EFC_46	(shiftPosition (motif2 c g) @ 73 ±3)
GP_EFC_47	(shiftPosition (motif2 c g) @ 74 ±3)
GP_EFC_48	(shiftPosition (motif2 c t) @ 62 ±3)
GP_EFC_49	(shiftPosition (motif3 a g c) @ 3 ±3)
GP_EFC_50	(shiftPosition (motif3 c a g) @ 8 ±3)
GP_EFC_51	(shiftPosition (motif3 c g a) @ 12 ±3)
GP_EFC_52	(shiftPosition (motif3 c g a) @ 13 ±3)
GP_EFC_53	(shiftPosition (motif3 c g a) @ 8 ±3)
GP_EFC_54	(shiftPosition (motif3 c g t) @ 9 ±3)
GP_EFC_55	(shiftPosition (motif3 g a c) @ 60 ±3)
GP_EFC_56	(shiftPosition (motif3 g c c) @ 9 ±3)
GP_EFC_57	(shiftPosition (motif3 g g c) @ 28 ±3)
GP_EFC_58	(shiftPosition (motif3 g g c) @ 29 ±3)
GP_EFC_59	(shiftPosition (motif3 t g t) @ 35 ±3)
GP_EFC_60	(shiftPosition (motif4 a a a a) @ 33 ±3)
GP_EFC_61	(shiftPosition (motif4 a g c c) @ 8 ±3)
GP_EFC_62	(shiftPosition (motif4 c t c c) @ 4 ±3)
GP_EFC_63	(shiftPosition (motif4 g c c a) @ 2 ±3)
GP_EFC_64	(shiftPosition (motif4 g g a t) @ 61 ±3)
GP_EFC_65	(shiftPosition (motif4 g g c c) @ 25 ±3)
GP_EFC_66	(shiftPosition (motif4 g t c g) @ 21 ±3)
GP_EFC_67	(shiftPosition (motif4 t t a t) @ 11 ±3)
GP_EFC_68	(shiftPosition (motif5 t t a t t) @ 7 ±3)
GP_EFC_69	(shiftPosition (motif6 g a c a c g) @ 0 ±3)
GP_EFC_70	(shiftPosition (motif6 g c c t g a) @ 1 ±3)
GP_EFC_71	(shiftPosition (motif6 t a t c a t) @ 10 ±3)
GP_EFC_72	(shiftPosition (motif6 t g c c a a) @ 44 ±3)
GP_EFC_73	(shiftPosition (motif7 a t a a t c c) @ 10 ±3)
GP_EFC_74	(shiftPosition (motif7 c c t c c a a) @ 1 ±3)
GP_EFC_75	(shiftPosition (motif7 g c t a g c c) @ 12 ±3)
GP_EFC_76	(shiftPosition (motif8 c g t c a c a c) @ 8 ±3)

Table C.4: Chapter 7 Gram-Both EFC Features

Feature	**Definition**
GB_EFC_1	(AND (matchesAtPosition (motif2 g t) @ 59) (matches (motif3 c a t))
GB_EFC_2	(AND (matches (motif2 c a) (shiftPosition (motif4 t c t c) @ 78 ±3))
GB_EFC_3	(AND (matches (motif2 c g) (matches (motif7 c g t a t a a))
GB_EFC_4	(AND (matches (motif2 g a) (matches (motif5 g t c g c))
GB_EFC_5	(AND (matches (motif4 a a a a) (matches (motif4 t a c a))
GB_EFC_6	(AND (matches (motif5 t t c c c) (shiftPosition (motif3 g t t) @ 17 ±3))
GB_EFC_7	(correlate (motif2 a t) (motif2 t c) @ 1 ±3)
GB_EFC_8	(correlate (motif5 g a c g c) (motif2 a c) @ 17 ±3)
GB_EFC_9	(matchesAtPosition (motif2 g c) @ 1)
GB_EFC_10	(matchesAtPosition (motif2 g c) @ 18)
GB_EFC_11	(matchesAtPosition (motif2 t c) @ 3)
GB_EFC_12	(matchesAtPosition (motif2 t g) @ 13)
GB_EFC_13	(matchesAtPosition (motif3 c t c) @ 2)
GB_EFC_14	(matchesAtPosition (motif3 g g c) @ 27)
GB_EFC_15	(matchesAtPosition (motif4 c g t c) @ 17)
GB_EFC_16	(matchesAtPosition (motif4 t t a t) @ 12)
GB_EFC_17	(matchesAtPosition (motif5 c a a g g) @ 33)
GB_EFC_18	(matches (motif5 g t c g c)
GB_EFC_19	(matches (motif5 t a t t t)
GB_EFC_20	(matches (motif5 t g c c g)
GB_EFC_21	(matches (motif5 t g g g a)
GB_EFC_22	(matches (motif6 a a a t t c)
GB_EFC_23	(matches (motif6 a c g g c g)
GB_EFC_24	(matches (motif7 a c c c c t g)
GB_EFC_25	(matches (motif7 a g c c a c t)
GB_EFC_26	(matches (motif7 c g a g c a c)
GB_EFC_27	(matches (motif7 g g g a c g g)
GB_EFC_28	(matches (motif8 c g a c c g c c)
GB_EFC_29	(matches (motif8 c t c a c a t c)
GB_EFC_30	(matches (motif8 g a a g g a c a)
GB_EFC_31	(shiftPosition (motif2 c g) @ 5 ±3)
GB_EFC_32	(shiftPosition (motif2 g a) @ 11 ±3)
GB_EFC_33	(shiftPosition (motif2 g c) @ 0 ±3)
GB_EFC_34	(shiftPosition (motif3 a g c) @ 8 ±3)
GB_EFC_35	(shiftPosition (motif3 a g c) @ 9 ±3)
GB_EFC_36	(shiftPosition (motif3 c g a) @ 44 ±3)
GB_EFC_37	(shiftPosition (motif3 g g c) @ 10 ±3)
GB_EFC_38	(shiftPosition (motif3 g g c) @ 11 ±3)
GB_EFC_39	(shiftPosition (motif3 g g c) @ 13 ±3)
GB_EFC_40	(shiftPosition (motif3 g t t) @ 16 ±3)
GB_EFC_41	(shiftPosition (motif4 a g c c) @ 15 ±3)
GB_EFC_42	(shiftPosition (motif4 c c a a) @ 5 ±3)
GB_EFC_43	(shiftPosition (motif4 g g g g) @ 4 ±3)
GB_EFC_44	(shiftPosition (motif5 c t t a t) @ 14 ±3)
GB_EFC_45	(shiftPosition (motif5 g c c c a) @ 0 ±3)
GB_EFC_46	(shiftPosition (motif6 t c a g t c) @ 5 ±3)
GB_EFC_47	(shiftPosition (motif7 c a c g c c g) @ 7 ±3)

Bibliography

Bibliography

[1] C. T. Bergstrom and M. Feldgarden, "The ecology and evolution of antibiotic-resistant bacteria." Oxford University Press, 22 November 2007, vol. 1, pp. 125–139.

[2] D. Byarugaba, "Antimicrobial resistance in developing countries and responsible risk factors," *International J. of Antimicrobial Agents*, vol. 24, no. 2, pp. 105 – 110, 2004.

[3] World Health Organization, "Race against time to develop new antibiotics," *Bulletin of the World Health Organization*, vol. 89, pp. 88–89, 2011.

[4] M. Barber and J. Whitehead, "Bacteriophage types in penicillin-resistant Staphylococcal infection," *Br. Med. J.*, vol. 2, pp. 565–569, 1949.

[5] M. N. Swartz, "Hospital-acquired infections: diseases with increasingly limited therapies," *Proceedings of the National Academy of Sciences*, vol. 91, no. 7, pp. 2420–2427, 1994.

[6] "Antibiotic resistance threats in the United States, 2013," 2013.

[7] "Executive order no. 13676, 79 c.f.r 56931 (2014)."

[8] A. S. Kesselheim and K. Outterson, "Fighting antibiotic resistance: marrying new financial incentives to meeting public health goals," *Health Affairs*, vol. 29, no. 9, pp. 1689–1696, 2010.

[9] H. Pearson, "Antibiotic faces uncertain future," *Nature*, vol. 441, no. 7091, pp. 260–261, 2006.

[10] R. E. W. Hancock and G. Diamond, "The role of cationic antimicrobial peptides in innate host defenses," *Trends in Microbiology*, vol. 8, no. 9, pp. 402–410, 2000.

[11] H. G. Boman, "Antibacterial peptides: basic facts and emerging concepts," *Journal of Internal Medicine*, vol. 254, no. 3, pp. 197–215, 2003.

[12] R. E. Hancock, K. L. Brown, and N. Mookherjee, "Host defence peptides from invertebrates - emerging antimicrobial strategies," *Immunobiology*, vol. 211, no. 4, pp. 315–322, 2006.

[13] A. Tossi, L. Sandri, and A. Giangaspero, "Amphipathic, α-helical antimicrobial peptides," *Peptide Science*, vol. 55, no. 1, pp. 4–30, 2000.

[14] Y. Wang, F. C. Knoop, I. Remy-Jouet, C. Delarue, H. Vaudry, and J. M. Conlon, "Antimicrobial peptides of the brevinin-2 family isolated from gastric tissue of the frog, Rana esculenta," *Biochemical and Biophysical Research Communications*, vol. 253, no. 3, pp. 600–603, 1998.

[15] K. G. Meade, S. Cahalane, F. Narciandi, P. Cormican, A. T. Lloyd, and C. O'Farrelly, "Directed alteration of a novel bovine β-defensin to improve antimicrobial efficacy against methicillin-resistant Staphylococcus aureus (MRSA)," *International Journal of Antimicrobial Agents*, vol. 32, no. 5, pp. 392–397, 2008.

[16] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[17] Z. Wang and G. Wang, "APD: the antimicrobial peptide database," *Nucl. Acids Res.*, vol. 32, no. Sup.1, pp. D590–D592, 2004.

[18] G. Wang, *Antimicrobial Peptides: Discovery, Design and Novel Therapeutic Strategies*. Wallingford, England: CABI Bookshop, 2010.

[19] I. Zelezetsky, A. Pontillo, L. Puzzi, N. Antcheva, L. Segat, S. Pacor, S. Crovella, and A. Tossi, "Evolution of the primate cathelicidin," *J. of Biological Chemistry*, vol. 281, no. 29, pp. 19 861–19 871, 2006.

[20] B. Ramanathan, E. G. Davis, C. R. Ross, and F. Blecha, "Cathelicidins: microbicidal activity, mechanisms of action, and roles in innate immunity," *Microbes and Infection*, vol. 4, no. 3, pp. 361–372, 2002.

[21] Y. Shai, "Mode of action of membrane active antimicrobial peptides," *Peptide Science*, vol. 66, no. 4, pp. 236–248, 2002.

[22] H. Raghuraman and A. Chattopadhyay, "Interaction of melittin with membrane cholesterol: A fluorescence approach," *Biophysical Journal*, vol. 87, no. 4, pp. 2419–2432, 2004.

[23] I. Cornut, K. Büttner, J.-L. Dasseux, and J. Dufourcq, "The amphipathic α-helix concept: application to the de novo design of ideally amphipathic Leu, Lys peptides with hemolytic activity higher than that of melittin," *FEBS letters*, vol. 349, no. 1, pp. 29–33, 1994.

[24] A. A. Strmstedt, L. Ringstad, A. Schmidtchen, and M. Malmsten, "Interaction between amphiphilic peptides and phospholipid membranes," *Current Opinion in Colloid and Interface Science*, vol. 15, no. 6, pp. 467–478, 2010.

[25] R. F. Epand, G. Wang, B. Berno, and R. M. Epand, "Lipid segregation explains selective toxicity of a series of fragments derived from the human cathelicidin LL-37," *Antimicrob. Agents Chemother.*, vol. 53, no. 9, pp. 3705–3714, 2009.

[26] J. Yu, N. Mookherjee, K. Wee, D. M. Bowdish, J. Pistolic, Y. Li, L. Rehaume, and R. E. Hancock, "Host defense peptide LL-37, in synergy with inflammatory mediator il-1β, augments immune responses by multiple pathways," *The Journal of Immunology*, vol. 179, no. 11, pp. 7684–7691, 2007.

[27] W. C. Wimley and K. Hristova, "Antimicrobial peptides: successes, challenges and unanswered questions," *The Journal of membrane biology*, vol. 239, no. 1-2, pp. 27–34, 2011.

[28] J. Oreopoulos, R. F. Epand, R. M. Epand, and C. M. Yip, "Peptide-induced domain formation in supported lipid bilayers: direct evidence by combined atomic force and polarized total internal reflection fluorescence microscopy," *Biophysical journal*, vol. 98, no. 5, pp. 815–823, 2010.

[29] G. Gogoladze, M. Grigolava, B. Vishnepolsky, M. Chubinidze, P. Duroux, M.-P. Lefranc, and M. Pirtskhalava, "DBAASP: database of antimicrobial activity and structure of peptides," *FEMS Microbiology Letters*, vol. 357, no. 1, pp. 63–68, 2014.

[30] Y. Pouny, D. Rapaport, A. Mor, P. Nicolas, and Y. Shai, "Interaction of antimicrobial dermaseptin and its fluorescently labeled analogues with phospholipid membranes," *Biochemistry*, vol. 31, pp. 12 416–12 423, 1992.

[31] E. G and L. H., "Electrically gated ionic channels in lipid bilayers," *Q Rev Biophys*, vol. 10, pp. 1–34, 1977.

[32] A. K. Mahalka and P. K. Kinnunen, "Binding of amphipathic α-helical antimicrobial peptides to lipid membranes: Lessons from temporins B and L," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1788, no. 8, pp. 1600–1609, 2009.

[33] V. Teixeira, M. J. Feio, and M. Bastos, "Role of lipids in the interaction of antimicrobial peptides with membranes," *Progress in lipid research*, vol. 51, no. 2, pp. 149–177, 2012.

[34] C. D. Fjell, J. A. Hiss, R. E. Hancock, and G. Schneider, "Designing antimicrobial peptides: form follows function," *Nature reviews Drug discovery*, vol. 11, no. 1, pp. 37–51, 2012.

[35] E. F. Haney and R. E. Hancock, "Peptide design for antimicrobial and immunomodulatory applications," *Peptide Science*, vol. 100, no. 6, pp. 572–583, 2013.

[36] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chemistry*, vol. 11, no. 3, pp. 218–234, 2015.

[37] ——, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[38] Z. Liu, X. Xiao, W.-R. Qiu, and K.-C. Chou, "iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition," *Analytical biochemistry*, vol. 474, pp. 69–77, 2015.

[39] W. Chen, H. Lin, and K.-C. Chou, "Pseudo nucleotide composition or pseknc: an effective formulation for analyzing genomic sequences," *Molecular BioSystems*, 2015.

[40] P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen *et al.*, "Prediction of antimicrobial peptides based on sequence alignment and feature selection methods," *PloS one*, vol. 6, no. 4, p. e18476, 2011.

[41] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical biochemistry*, 2013.

[42] S. Lata, B. K. Sharma, and G. P. Raghava, "Analysis and prediction of antibacterial peptides," *BMC Bioinformatics*, vol. 23, no. 8, pp. 263–272, 2007.

[43] R. Ferre, M. N. Melo, A. D. Correia, L. Feliu, E. Bardají, M. Planas, and M. Castanho, "Synergistic effects of the membrane actions of cecropin-melittin antimicrobial hybrid peptide bp100," *Biophysical journal*, vol. 96, no. 5, pp. 1815–1827, 2009.

[44] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, and S. I. Thomas, "CAMP: a useful resource for research on antimicrobial peptides," *Nucl. Acids Res.*, vol. 38, no. Suppl 1, pp. D774–D780, 2009.

[45] W. F. Porto, F. C. Fernandes, and O. L. Franco, "An SVM model based on physicochemical properties to predict antimicrobial activity from protein sequences with cysteine knot motifs," *Lecture Notes in Computer Science*, vol. 6268, pp. 59–62, 2010.

[46] M. Torrent, D. Andreu, V. M. Nogués, and E. Boix, "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model," *PLoS One*, vol. 6, no. 2, p. e16968, 2011.

[47] F. C. Fernandes, D. J. Rigden, and O. L. Franco, "Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application," *Peptide Science*, vol. 98, no. 4, pp. 280–287, 2012.

[48] E. G. Randou, D. Veltri, and A. Shehu, "Systematic analysis of global features and model building for recognition of antimicrobial peptides," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on.* IEEE, 2013, pp. 1–6.

[49] ——, "Binary response models for recognition of antimicrobial peptides," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics.* ACM, 2013, p. 76.

[50] C. Fjell, R. Hancock, and A. Cherkasov, "AMPer: a database and an automated discovery tool for antimicrobial peptides," *Bioinformatics*, vol. 23, no. 9, pp. 1148–1155, 2007.

[51] C. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Pante, R. E. Hancock, and A. Cherkasov, "Identification of novel antibacterial peptides by chemoinformatics and machine learning," *J. Med. Chem.*, vol. 52, no. 7, pp. 2006–2015, 2009.

[52] D. Veltri and A. Shehu, "Physicochemical determinants of antimicrobial activity," in *Proceedings from the 5th International Conference on Bioinformatics and Computational Biology (BICoB2013).* ISCA, 2013.

[53] M. Bunge, "A general black box theory," *Philosophy of Science*, vol. 30, no. 4, pp. 346–358, 1963.

[54] X. Zhu, L. Zhang, J. Wang, Z. Ma, W. Xu, J. Li, and A. Shan, "Characterization of antimicrobial activity and mechanisms of low amphipathic peptides with different alpha-helical propensity," *Acta Biomaterialia*, vol. 18, pp. 155–167, 2015.

[55] R. M. Epand, "Detecting the presence of membrane domains using DSC," *Biophysical chemistry*, vol. 126, no. 1, pp. 197–200, 2007.

[56] R. M. Epand, S. Rotem, A. Mor, B. Berno, and R. F. Epand, "Bacterial membranes as predictors of antimicrobial potency," *Journal of the American Chemical Society*, vol. 130, no. 43, pp. 14 346–14 352, 2008.

[57] R. M. Epand and R. F. Epand, "Bacterial membrane lipids in the action of antimicrobial agents," *Journal of Peptide Science*, vol. 17, no. 5, pp. 298–305, 2011.

[58] M. N. Melo, R. Ferre, and M. A. Castanho, "Antimicrobial peptides: linking partition, activity and high membrane-bound concentrations," *Nature Reviews Microbiology*, vol. 7, no. 3, pp. 245–250, 2009.

[59] M. Mihajlovic and T. Lazaridis, "Antimicrobial peptides in toroidal and cylindrical pores," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1798, no. 8, pp. 1485–1493, 2010.

[60] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011.

[61] Y. Li, D.-Q. Wei, W.-N. Gao, H. Gao, B.-N. Liu, C.-J. Huang, W.-R. Xu, D.-K. Liu, H.-F. Chen, and K.-C. Chou, "Computational approach to drug design for oxazolidinones as antibacterial agents," *Medicinal Chemistry*, vol. 3, no. 6, pp. 576–582, 2007.

[62] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, 2013.

[63] Y. Xu, X. Wen, L.-S. Wen, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition," 2014.

[64] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 8, pp. 1–12, 2014.

[65] R. Xu, J. Zhou, B. Liu, Y. He, Q. Zou, X. Wang, and K.-C. Chou, "Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 8, pp. 1–11, 2014.

[66] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC," *Journal of theoretical biology*, vol. 377, pp. 47–56, 2015.

[67] B. M. Bishop, M. L. Juba, M. C. Devine, S. M. Barksdale, C. A. Rodriguez, M. C. Chung, P. S. Russo, K. A. Vliet, J. M. Schnur, and M. L. van Hoek, "Bioprospecting the american alligator (*Alligator mississippiensis*) host defense peptidome," *PLoS ONE*, vol. 10, no. 2, p. e0117394, 02 2015.

[68] M. L. Juba, P. S. Russo, M. Devine, S. Barksdale, C. Rodriguez, J. M. Schnur, M. L. van Hoek, and B. M. Bishop, "Large scale discovery and de novo-assisted sequencing of cationic antimicrobial peptides (CAMPs) by microparticle capture and electron-transfer dissociation (ETD) mass spectrometry," *Journal of proteome research*, 2015.

[69] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*, 2nd ed. Springer, 2009.

[70] M. Kuhn and K. Johnson, *Applied predictive modeling*, 1st ed. Springer, 2013.

[71] E. Alpaydin, *Introduction to machine learning*. MIT press, 2004.

[72] J. M. Andrews, "Determination of minimum inhibitory concentrations," *Journal of antimicrobial Chemotherapy*, vol. 48, no. Suppl 1, pp. 5–16, 2001.

[73] G. Wang, X. Li, and Z. Wang, "APD2: the updated antimicrobial peptide database and its application in peptide design," *Nucl. Acids Res.*, vol. 37, no. Sup.1, pp. D933–D937, 2009. [Online]. Available: http://aps.unmc.edu/AP

[74] K. Jensen, M. Styczynski, and G. Stephanopoulos, "Machine learning approaches to modeling the physiochemical properties of small peptides," 2005.

[75] P. Duckert, S. Brunak, and N. Blom, "Prediction of proprotein convertase cleavage sites," *Protein Engineering Design and Selection*, vol. 17, no. 1, pp. 107–112, 2004.

[76] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "Incremental genetic k-means algorithm and its application in gene expression data analysis," *BMC Bioinformatics*, vol. 5, no. 1, p. 172, 2004.

[77] U. Kamath, K. A. De Jong, and A. Shehu, "An evolutionary-based approach for feature generation: Eukaryotic promoter recognition," in *IEEE CEC*, A. E. Smith, Ed. IEEE Press, 2011, pp. 277–284.

[78] J. Selbig, T. Mevissen, and T. Lengauer, "Decision tree-based formation of consensus protein secondary structure prediction," *Bioinformatics*, vol. 15, no. 12, pp. 1039–1046, 1999.

[79] J. R. Taylor, *An introduction to error analysis: the study of uncertainties in physical measurements*. Univ. Science Books, 1997.

[80] B. W. Matthews *et al.*, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." *Biochimica et biophysica acta*, vol. 405, no. 2, p. 442, 1975.

[81] E. W. Steyerberg, F. E. Harrel, G. J. Borsboom, and e. al., "Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis," *J. of Clinical Epidimiology*, vol. 54, pp. 774–781, 2001.

[82] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[83] B. P. Burnham and D. R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer, 1998.

[84] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[85] T. M. Mitchell, *Machine learning.* McGraw-Hill Boston, MA., 1997.

[86] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal, The*, vol. 27, no. 3, pp. 379–423, July 1948.

[87] S. Lata, N. K. Mishra, and G. P. Raghava, "AntiBP2: improved version of antibacterial peptide prediction." *BMC Bioinformatics*, vol. 11, no. Suppl 1, pp. S1–S19, 2010.

[88] D. Veltri, "Physicochemical feature selection for cathelicidin antimicrobial peptides," Master's thesis, George Mason University, 2013.

[89] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215–242, 1958.

[90] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.

[91] S. Milborrow, "Notes on the earth package." [Online]. Available: http://www.milbo.org/doc/earth-notes.pdf

[92] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[93] ——, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[94] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[95] L. Käll, A. Krogh, and E. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction - the phobius web server," *Nucl. Acids Res.*, vol. 35, no. suppl 2, pp. W429–W432, 2007.

[96] M. Magrane and the UniProt consortium, "UniProt knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, no. bar009, pp. 1–13, 2011.

[97] W. J. Kent, "BLAT - the BLAST-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.

[98] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[99] P. Rice, I. Longden, A. Bleasby *et al.*, "EMBOSS: the european molecular biology open software suite," *Trends in genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[100] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org

[101] A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, and L. Serrano, "Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins," *Nat Biotechnology*, vol. 22, no. 10, pp. 1302–1306, 2004.

[102] O. Conchillo-Solé, N. de Groot, F. Avilés, J. Vendrell, X. Daura, and S. Ventura, "AGGRESCAN: a server for the prediction and evaluation of," *BMC bioinformatics*, vol. 8, no. 1, p. 65, 2007.

[103] P. Artimo, M. Jonnalagedda, K. Arnold *et al.*, "ExPASy: SIB bioinformatics resource portal," *Nucl. Acids Res.*, vol. 40, no. W1, pp. W597–W603, 2012.

[104] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105–132, 1982.

[105] K. A. Henzler Wildman, D.-K. Lee, and A. Ramamoorthy, "Mechanism of lipid bilayer disruption by the human antimicrobial peptide, LL-37," *Biochemistry*, vol. 42, no. 21, pp. 6545–6558, 2003.

[106] S. S. Qian, R. S. King, and C. J. Richardson, "Two statistical methods for the detection of environmental thresholds," *Ecological Modelling*, vol. 166, no. 1, pp. 87–97, 2003.

[107] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[108] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[109] M. Crawley, *The R Book*. New York, NY: Wiley & Sons, 2013.

[110] U. Kamath, J. Compton, R. Islamaj-Dogan, D. K. A., and A. Shehu, "An evolutionary algorithm approach for feature generation from sequence data and its application to dna splice-site prediction," *Trans Comp Biol and Bioinf*, 2012.

[111] U. Kamath, K. A. De Jong, and A. Shehu, "Effective automated feature construction and selection for classification of biological sequences," *PLoS ONE*, vol. 9, no. 7, p. e99982, 2014.

[112] M. Torrent, P. Di Tommaso, D. Pulido, M. V. Nogues, N. C., E. Boix, and D. Andreu, "AMPA: An automated web server for prediction of protein antimicrobial regions," *Bioinformatics*, vol. 28, no. 1, pp. 130–131, 2012.

[113] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[114] G. Wang, B. Mishra, K. Lau, T. Lushnikova, R. Golla, and X. Wang, "Antimicrobial peptides in 2014," *Pharmaceuticals*, vol. 8, no. 1, pp. 123–150, 2015.

[115] D. Veltri, U. Kamath, and A. Shehu, "A novel method to improve recognition of antimicrobial peptides through distal sequence-based features," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on.* IEEE, 2014, pp. 371–378.

[116] ——, "Improving recognition of antimicrobial peptides and target selectivity through machine learning and genetic programming," *Transactions on Computational Biology and Bioinformatics*, 2015.

[117] Waikato Machine Learning Group, "Weka," 2010. [Online]. Available: http://weka.org

[118] A. D. Solis and S. Rackovsky, "Optimized representations and maximal information in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 38, no. 2, pp. 149–164, 2000.

[119] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucl. Acids Res.*, vol. 28, no. 1, p. 374, 2000.

[120] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, p. 7881, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de

[121] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucl. Acids Res.*, vol. 28, no. 1, p. 374, 2000.

[122] M. Bachar and O. M. Becker, "Melittin at a membrane/water interface: Effects on water orientation and water penetration," *The Journal of Chemical Physics*, vol. 111, no. 18, pp. 8672–8685, 1999.

[123] M. A. Fox, J. E. Thwaite, D. O. Ulaeto, T. P. Atkins, and H. S. Atkins, "Design and characterization of novel hybrid antimicrobial peptides based on cecropin A, LL-37 and magainin II," *Peptides*, vol. 33, no. 2, pp. 197–205, 2012.

[124] X. Feng, C. Liu, J. Guo, X. Song, J. Li, W. Xu, and Z. Li, "Recombinant expression, purification, and antimicrobial activity of a novel hybrid antimicrobial peptide LFT33," *Applied microbiology and biotechnology*, vol. 95, no. 5, pp. 1191–1198, 2012.

[125] Y. Liu, K. M. Knapp, L. Yang, S. Molin, H. Franzyk, and A. Folkesson, "High in vitro antimicrobial activity of β-peptoid–peptide hybrid oligomers against planktonic and biofilm cultures of staphylococcus epidermidis," *International journal of antimicrobial agents*, vol. 41, no. 1, pp. 20–27, 2013.

[126] R. Wu, Q. Wang, Z. Zheng, L. Zhao, Y. Shang, X. Wei, X. Liao, and R. Zhang, "Design, characterization and expression of a novel hybrid peptides melittin (113)-LL37 (1730)," *Molecular Biology Reports*, vol. 41, no. 7, pp. 4163–4169, 2014.

[127] S. M. Derived from MDA:MARS by T. Hastie and R. Tibshirani., *earth: Multivariate Adaptive Regression Splines*, 2011. [Online]. Available: http://CRAN.R-project.org/package=earth

[128] T. P. Creamer, R. Srinivasan, and G. D. Rose, "Modeling unfolded states of proteins and peptides. ii. backbone solvent accessibility," *Biochemistry*, vol. 36, no. 10, pp. 2832–2835, 1997.

[129] R. Gautier, D. Douguet, B. Antonny, and D. G., "HELIQUEST: a web server to screen sequences with specific α-helical properties," *Bioinformatics*, vol. 24, no. 18, pp. 2101–2102, 2008. [Online]. Available: http://heliquest.ipmc.cnrs.fr

[130] M. Schiffer and A. B. Edmundson, "Use of helical wheels to represent the structures of proteins and to identify segments with helical potential," *Biophysical Journal*, vol. 7, no. 2, p. 121, 1967.

[131] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, "The hydrophobic moment detects periodicity in protein hydrophobicity," *Proc. Natl. Acad. Sci.*, vol. 81, no. 1, pp. 140–144, 1984.

[132] D. Eisenberg, "Three-dimensional structure of membrane and surface proteins," *Annual review of biochemistry*, vol. 53, no. 1, pp. 595–623, 1984.

[133] W. C. Wimley and S. H. White, "Experimentally determined hydrophobicity scale for proteins at membrane interfaces," *Nature Structural & Molecular Biology*, vol. 3, no. 10, pp. 842–848, 1996.

[134] L. J. McGuffin, K. Bryson, and D. T. Jones, "The psipred protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[135] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, "JPred4: a protein secondary structure prediction server," *Nucleic Acids Research*, vol. 43, no. W1, pp. W389–W394, 2015.

[136] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, P. Hnigschmid, A. Schafferhans, M. Roos, M. Bernhofer *et al.*, "Predictprotein- an open resource for online prediction of protein structural and functional features," *Nucleic acids research*, p. gku366, 2014.

[137] M. Zvelebil and B. Jeremy, *Understanding bioinformatics*. Garland Science, 2007.

[138] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and the UniProt Consortium, "Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, 2015.

[139] J. R. Quinlan, "C4.5: Programs for machine learning," 1993.

[140] R. Fan, P. Chen, and C. Lin, "Working set selection using the second order information for training SVM," *J. Mach. Learn. Res.*, vol. 6, no. 1532-4435, pp. 1889–1918, 2005.

[141] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[142] S. Cowan and T. Schirmer, "Crystal structures explain functional properties of two E. coli porins," *Nature*, vol. 358, no. 6389, pp. 727–733, 1992.

[143] M. L. Mangoni, R. F. Epand, Y. Rosenfeld, A. Peleg, D. Barra, R. M. Epand, and Y. Shai, "Lipopolysaccharide, a key molecule involved in the synergism between temporins in inhibiting bacterial growth and in endotoxin neutralization," *Journal of Biological Chemistry*, vol. 283, no. 34, pp. 22 907–22 917, 2008.

[144] P. D. Allison, "Comparing logit and probit coefficients across groups," *Sociological Methods & Research*, vol. 28, no. 2, pp. 186–208, 1999.

[145] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *European Sociological Review*, vol. 26, no. 1, pp. 67–82, 2010.

[146] A. R and R. GD., "Helix capping," *Protein Sci.*, vol. 7, no. 1, pp. 23–38, 1998.

[147] K. Yutani, K. Ogasahara, T. Tsujita, and Y. Sugino, "Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit," *PNAS*, vol. 84, no. 13, pp. 4441–4444, 1987.

[148] F. R. Maxfield and H. A. Scheraga, "Status of empirical methods for the prediction of protein backbone topography," *Biochemistry*, vol. 15, no. 23, pp. 5138–5153, 1976.

[149] P. Karplus and G. Schulz, "Prediction of chain flexibility in proteins," *Naturwissenschaften*, vol. 72, no. 4, pp. 212–213, 1985.

[150] J. S. Richardson and D. C. Richardson, "Amino acid preferences for specific locations at the ends of alpha helices," *Science*, vol. 240, no. 4859, pp. 1648–1652, 1988.

[151] S. Rackovsky and H. Scheraga, "Differential geometry and polymer conformation. 4. conformational and nucleation properties of individual amino acids," *Macromolecules*, vol. 15, no. 5, pp. 1340–1346, 1982.

[152] P. Klein, M. Kanehisa, and C. DeLisi, "Prediction of protein function from sequence properties: Discriminant analysis of a data base," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 787, no. 3, pp. 221–226, 1984.

[153] R. A. George and J. Heringa, "An analysis of protein domain linkers: their classification and role in protein folding," *Protein Engineering*, vol. 15, no. 11, pp. 871–879, 2002.

[154] G. Wang, "Structures of human host defense cathelicidin LL-37 and its smallest antimicrobial peptide KR-12 in lipid micelles," *Journal of Biological Chemistry*, vol. 283, no. 47, pp. 32 637–32 643, 2008.

[155] H. Naderi-Manesh, M. Sadeghi, S. Arab, and A. A. Moosavi Movahedi, "Prediction of protein surface accessibility with information theory," *Proteins: Structure, Function, and Bioinformatics*, vol. 42, no. 4, pp. 452–459, 2001.

[156] H. Nakashima, K. Nishikawa, and T. Ooi, "Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 8, no. 2, pp. 173–178, 1990.

[157] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[158] N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts, and T. Katayama, "BioRuby: bioinformatics software for the Ruby programming language," *Bioinformatics*, vol. 26, no. 20, pp. 2617–2619, 2010.

[159] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girn, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Khri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. Searle, "Ensembl 2014," vol. 42, no. D1, pp. D749–D755, 2014.

[160] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[161] J. A. St John, E. L. Braun, S. R. Isberg, L. G. Miles, A. Y. Chong, J. Gongora, P. Dalzell, C. Moran, B. Bed'Hom, A. Abzhanov *et al.*, "Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes," *Genome biology*, vol. 13, no. 1, p. 415, 2012.

[162] J. Slaninová, V. Mlsová, H. Kroupová, L. Alán, T. Tumová, L. Monincová, L. Borovičková, V. Fučík, and V. Čeřovskỳ, "Toxicity study of antimicrobial peptides from wild bee venom and their analogs toward mammalian normal and cancer cells," *Peptides*, vol. 33, no. 1, pp. 18–26, 2012.

[163] P. Bulet, J. Dimarcq, C. Hetru, M. Lagueux, M. Charlet, G. Hegy, A. Van Dorsselaer, and J. Hoffmann, "A novel inducible antibacterial peptide of Drosophila carries an O-glycosylated substitution." *Journal of Biological Chemistry*, vol. 268, no. 20, pp. 14 893–14 897, 1993.

[164] C. R. Bevier, A. Sonnevend, J. Kolodziejek, N. Nowotny, P. F. Nielsen, and J. M. Conlon, "Purification and characterization of antimicrobial peptides from the skin secretions of the mink frog (Rana septentrionalis)," *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, vol. 139, no. 1, pp. 31–38, 2004.

[165] M. Lamberty, D. Zachary, R. Lanot, C. Bordereau, A. Robert, J. A. Hoffmann, and P. Bulet, "Insect immunity constitutive expression of a cysteine-rich antifungal and a linear antibacterial peptide in a termite insect," *Journal of Biological Chemistry*, vol. 276, no. 6, pp. 4085–4092, 2001.

[166] Y. Lu, Y. Zhuang, and J. Liu, "Mining antimicrobial peptides from small open reading frames in ciona intestinalis," *Journal of Peptide Science*, vol. 20, no. 1, pp. 25–29, 2014.

[167] J. Conlon, A. Sonnevend, M. Patel, C. Davidson, P. Nielsen, T. Pal, and L. Rollins-Smith, "Isolation of peptides of the brevinin-1 family with potent candidacidal activity from the skin secretions of the frog Rana boylii," *The Journal of peptide research*, vol. 62, no. 5, pp. 207–213, 2003.

[168] R. Lai, H. Liu, W. H. Lee, and Y. Zhang, "A novel proline rich bombesin-related peptide (PR-bombesin) from toad bombina maxima," *Peptides*, vol. 23, no. 3, pp. 437–442, 2002.

[169] V. Rydengard, O. Shannon, K. Lundqvist, L. Kacprzyk, A. Chalupka, A.-K. Olsson, M. Morgelin, W. Jahnen-Dechent, M. Malmsten, and A. Schmidtchen, "Histidine-rich glycoprotein protects from systemic candida infection," *PLoS Pathog*, vol. 4, no. 8, pp. e1 000 116–e1 000 116, 2008.

[170] C. Clarke, W. Williams, and J. Teruya, "Ubiquinone biosynthesis in Saccharomyces cerevisiae. isolation and sequence of COQ3, the 3, 4-dihydroxy-5-hexaprenylbenzoate methyltransferase gene." *Journal of Biological Chemistry*, vol. 266, no. 25, pp. 16 636–16 644, 1991.

[171] I. H. Lee, Y. Cho, and R. I. Lehrer, "Styelins, broad-spectrum antimicrobial peptides from the solitary tunicate, Styela clava," *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, vol. 118, no. 3, pp. 515–521, 1997.

[172] H. Lee, J.-S. Hwang, J. Lee, J. I. Kim, and D. G. Lee, "Scolopendin 2, a cationic antimicrobial peptide from centipede, and its membrane-active mechanism," *BBA Biomembranes*, vol. 1848, no. 2, pp. 634–642, 2015.

[173] P. Narbonne, D. E. Simpson, and J. B. Gurdon, "Deficient induction response in a Xenopus nucleocytoplasmic hybrid (Figure S1)," *PLOS Biology*, 11 2011. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.1001197

[174] K. Roelants, B. G. Fry, L. Ye, B. Stijlemans, L. Brys, P. Kok, E. Clynen, L. Schoofs, P. Cornelis, and F. Bossuyt, "Origin and functional diversification of an amphibian defense peptide arsenal," *PLoS Genet*, vol. 9, no. 8, p. e1003662, 2013.

[175] M. R. Yeaman, K. D. Gank, A. S. Bayer, and E. P. Brass, "Synthetic peptides that exert antimicrobial activities in whole blood and blood-derived matrices," *Antimicrobial agents and chemotherapy*, vol. 46, no. 12, pp. 3883–3891, 2002.

[176] K. L. Borden and P. S. Freemont, "The ring finger domain: a recent example of a sequencestructure family," *Current opinion in structural biology*, vol. 6, no. 3, pp. 395–401, 1996.

[177] A. Klug, "Zinc finger peptides for the regulation of gene expression," *Journal of molecular biology*, vol. 293, no. 2, pp. 215–218, 1999.

[178] T. M. T. Hall, "Multiple modes of RNA recognition by zinc finger proteins," *Current opinion in structural biology*, vol. 15, no. 3, pp. 367–373, 2005.

[179] R. Gamsjaeger, C. K. Liew, F. E. Loughlin, M. Crossley, and J. P. Mackay, "Sticky fingers: zinc-fingers as protein-recognition motifs," *Trends in biochemical sciences*, vol. 32, no. 2, pp. 63–70, 2007.

[180] D. A. Rasko, D. R. Webster, J. W. Sahl, A. Bashir, N. Boisen, F. Scheutz, E. E. Paxinos, R. Sebra, C.-S. Chin, D. Iliopoulos *et al.*, "Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in germany," *New England Journal of Medicine*, vol. 365, no. 8, pp. 709–717, 2011.

[181] P. M. Griffin and R. V. Tauxe, "The epidemiology of infections caused by Escherichia coli O157: H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome," *Epidemiologic reviews*, vol. 13, no. 1, pp. 60–98, 1991.

[182] K. A. Davis, J. J. Stewart, H. K. Crouch, C. E. Florez, and D. R. Hospenthal, "Methicillin-resistant Staphylococcus aureus (MRSA) nares colonization at hospital admission and its effect on subsequent MRSA infection," *Clinical Infectious Diseases*, vol. 39, no. 6, pp. 776–782, 2004.

[183] M. E. Mulligan, K. A. Murray-Leisure, B. S. Ribner, H. C. Standiford, J. F. John, J. A. Korvick, C. A. Kauffman, and L. Y. Victor, "Methicillin-resistant Staphylococcus aureus: a consensus review of the microbiology, pathogenesis, and epidemiology with implications for prevention and management," *The American journal of medicine*, vol. 94, no. 3, pp. 313–328, 1993.

[184] E. Alpaydin, *Introduction to machine learning*. MIT press, 2004, pp. 419–445.

[185] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[186] H. Drucker, C. Cortes, L. D. Jackel, Y. LeCun, and V. Vapnik, "Boosting and other ensemble methods," *Neural Computation*, vol. 6, no. 6, pp. 1289–1301, 1994.

[187] R. E. Schapire, "A brief introduction to boosting," in *Ijcai*, vol. 99, 1999, pp. 1401–1406.

[188] E. Alpaydin, *Introduction to machine learning*. MIT press, 2004, pp. 431–434.

[189] C. Kaynak and E. Alpaydin, "Multistage cascading of multiple classifiers: One man's noise is another man's data," in *ICML*. Citeseer, 2000, pp. 455–462.

[190] A. Kessel, D. Shental-Bechor, T. Haliloglu, and N. Ben-Tal, "Interactions of hydrophobic peptides with lipid bilayers: Monte carlo simulations with m2δ," *Biophysical journal*, vol. 85, no. 6, pp. 3431–3444, 2003.

[191] A. Kessel, D. Shental-Bechor, T. Haliloglu, N. Ben-Tal, Y. Wu, and G. A. Voth, "Interactions of the m2 segment of the acetylcholine receptor with lipid bilayers: A continuum-solvent model study," *Biophysical journal*, vol. 85, pp. 3687–3695, 2003.

[192] Y. Gofman, T. Haliloglu, and N. Ben-Tal, "Monte carlo simulations of peptide–membrane interactions with the mcpep web server," *Nucleic acids research*, p. gks577, 2012.

[193] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[194] L. Van der Maaten, "Barnes-Hut-SNE," in *Proceedings of the International Conference on Learning Representations*, 2013.

Curriculum Vitae

Daniel Veltri was born on November 20th, 1983 in Fairfax, Virginia. He received his B.A. in Environmental, Populismic and Organismic Biology, with a minor in Computer Science, from the University of Colorado at Boulder in 2006. Through the Japan Exchange Teacher (JET) Program, he spent the following two years living and teaching English in Aomori City, Japan, before returning to the US to study Bioinformatics at George Mason University. From 2010-2012 he helped teach and promote science, technology, engineering and math (STEM) subjects in Fairfax County Public Schools as a GMU SUNRISE (Schools, University 'N' Resources in the Sciences and Engineering) Fellow in the National Science Foundation's GK-12 Program. He also continued his interest in teaching as a GMU BRIDGE Scholar, assisting international graduate students with their studies through the Center for International Student Access (CISA) at Mason. In the Spring of 2013 he completed his M.S. in Bioinformatics and Computational Biology from George Mason University on the topic of Physicochemical Feature Selection for Cathelicidin Antimicrobial Peptides.

Education

- Masters of Science, Bioinformatics and Computational Biology, George Mason University, 2013
- Bachelor of Arts, Environmental Populismic and Organismic Biology, Computer Science Minor, University of Colorado at Boulder, 2006

Awards

- AAAS/Science Program for Excellence in Science
  (2015)
- Best Student Paper (1st Author) and Travel Award, IEEE International Conference on Bioinformatics and Biomedicine
  (2014)
- IEEE International Conference on Comp. Advances in Biology and Medical Sciences
  (2013)
- BRIDGE Scholar, George Mason University
  (2012-2014)
- GK-12 Fellowship, National Science Foundation and George Mason University
  (2010-2012)

Journal Articles (6)

1. Daniel Veltri, Uday Kamath and Amarda Shehu. Improving Recognition of Antimicrobial Peptides and their Target Selectivity through Machine Learning and Genetic Programming. *IEEE Transactions on Computational Biology and Bioinformatics Journal*, 2015.

2. B.D. Wingfield, P.K. Ades, F.A. Al-Naemi, L.A. Beirn, W. Bihon, J.A. Crouch, Z. Wilhelm de Beer, L. De Vos, T.A. Duong, C.J. Fields, G. Fourie, A.M. Kanzi, M. Malapi-Wight, S.J. Pethybridge, O. Radwan, G. Rendon, B. Slippers, Q.C. Santana, E.T. Steenkamp, P.W.J. Taylor, N. Vaghefi, N.A. van der Merwe, D. Veltri, and M.J. Wingfield. Draft genome sequences of *Chrysoporthe austroafricana*, *Diplodia scrobiculata*, *Fusarium nygamai*, *Leptographium lundbergii*, *Limonomyces culmigenus*, *Stagonosporopsis tanaceti* and *Thielaviopsis punctulata*. *IMA Genome-F 4*, 2015.

3. Martha Malapi-Wight, Catalina Salgado-Salazar, Jill Demers, Daniel Veltri, and Jo Anne Crouch. Draft Genome Sequence of *Dactylonectria macrodidyma*, a plant pathogenic fungus in the *Nectriaceae*. *ASM Genome Announcements Journal* 3(2):e00278-15, 2015

4. Nadine Kabbani, Jacob C. Nordman, Brian Corgiat, Daniel Veltri, Amarda Shehu and David J. Adams. Are Nicotinic Receptors Coupled to G Proteins? *BioEssays Journal* 35(12):1025-1034, 2013.

5. Abrar Ashoor, Jacob C. Nordman, Daniel Veltri, Keun-Hang Susan Yang, Lina Al Kury, Yaroslav Shuba, Mohamed Mahgoub, Frank C. Howarth, Carl Lupica, Amarda Shehu, Nadine Kabbani and Murat Oz. Menthol Inhibits 5-HT3 Receptor-Mediated Currents. *Journal of Pharmacology and Experimental Therapeutics* 347(2):398-402, 2013.

6. Abrar Ashoor, Jacob C. Nordman, Daniel Veltri, Keun-Hang Susan Yang, Lina Al Kury, Yaroslav Shuba, Mohamed Mahgoub, Frank C. Howarth, Carl Lupica, Amarda Shehu, Nadine Kabbani and Murat Oz. Menthol Binding and Inhibition of Alpha7-Nicotinic Acetylcholine Receptors. *PLoS ONE Journal* 8(7):e67674, 2013.

Peer-Reviewed Conference Papers (5)

1. Daniel Veltri, Uday Kamath and Amarda Shehu. A Novel Method to Improve Recognition of Antimicrobial Peptides through Distal Sequence-based Features. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM2014)*, Belfast, UK, 2014. *(best student paper award).*

2. Irina Hashmi, Daniel Veltri, Nadine Kabbani and Amarda Shehu. Knowledge-based Search and Multi-objective Filters: Proposed Structural Models of GPCR Dimerization. *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB2014)*, Newport Beach, CA, 2014.

3. Elena G. Randou, Daniel Veltri and Amarda Shehu. Binary Response Models for Recognition of Antimicrobial Peptides. *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB2013)*, Washington DC, 2013.

4. Elena G. Randou, Daniel Veltri and Amarda Shehu. Systematic Analysis of Global Features and Model Building for Recognition of Antimicrobial Peptides. *IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS2013)*. New Orleans, LA, 2013.

5. Daniel Veltri and Amarda Shehu. Physicochemical Determinants of Antimicrobial Activity *5th International Conference on Bioinformatics and Computational Biology (BICoB2013)*, Honolulu, HI, 2013.

Workshop Articles (1)

1. Daniel Veltri and Amarda Shehu. Physicochemical Features for Recognition of Antimicrobial Peptides. *IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Philadelphia, PA, 2012.