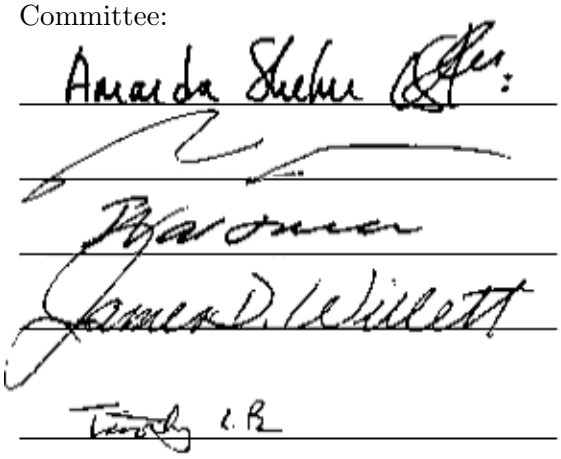
PHYSICOCHEMICAL FEATURE SELECTION FOR
CATHELICIDIN ANTIMICROBIAL PEPTIDES

by

Daniel Paul Veltri
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Master of Science
Bioinformatics and Computational Biology

Committee:



Dr. Amarda Shehu, Thesis Director

Dr. Barney Bishop, Committee Member

Dr. Iosif Vaisman, Committee Member

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

Dr. Timothy L. Born, Associate Dean for
Student and Academic Affairs,
College of Science

Dr. Vikas Chandhoke, Dean,
College of Science

Date: 4/30/2013

Spring Semester 2013
George Mason University
Fairfax, VA

Physicochemical Feature Selection for Cathelicidin Antimicrobial Peptides

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

Daniel Paul Veltri
Bachelor of Arts
University of Colorado at Boulder, 2006

Director: Dr. Amarda Shehu, Professor
Department of School of Systems Biology

Spring Semester 2013
George Mason University
Fairfax, VA

DEDICATION

I dedicate this thesis to my late grandfather and meteorologist, Dr. Duane S. Cooley. His love of science inspired me from an early age.

ACKNOWLEDGEMENTS

I would like to give a special thanks my adviser Dr. Shehu, the other members of the Shehu Lab, and my committee members Dr. Bishop and Dr. Vaisman for all their help. I also want thank to Dr. Ganesan, Dr. Sterling, the other GMU SUNRISE Fellows and the National Science Foundation for giving me the chance to teach science through the GK-12 Program. This work was supported in part by NSF Grant No. DGE-1007911 (NSF GK-12 Award No. 0638680).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

PHYSICOCHEMICAL FEATURE SELECTION FOR CATHELICIDIN ANTIMICROBIAL PEPTIDES

Daniel Paul Veltri, M.S.

George Mason University, 2013

Thesis Director: Dr. Amarda Shehu

Due to recent attention on antimicrobial peptides (AMPs) as targets for antibacterial drug research, many machine learning methods are now turning their attention to AMP recognition. Approaches that rely on whole-peptide properties for recognition are challenged by the great sequence diversity among AMPs for effective feature construction. This thesis proposes a novel and complementary method for feature construction which relies on an extensive list of position-based amino acid physicochemical properties. These features are shown effective in the context of classification by support vector machine (SVM), both in comparison to related work in recognition of AMPs and in a novel study on the cathelicidin family. A detailed analysis and careful construction of a decoy dataset allows for the highlighting of antimicrobial activity-related features in cathelicidins. Special attention is also given to residue positions involved with enzymatic cleavage. The method presented in this thesis is a first step towards understanding what confers to cathelicidins their activity at the physicochemical level and may prove useful for future AMP design efforts.

Chapter 1: Introduction and Background

1.1 Motivation: Antibiotic Resistance

Drug-resistance in bacteria continues to be a growing problem around the world [1–3]. Penicillin, the first major clinical antibiotic, experienced widespread resistance in Western hospitals only a few years after its introduction in 1945 [4, 5]. Ten of the major classes of antibiotics (penicillin, chloramphenicol, erythromycin, methicillin, 1st-3rd generation cephalothin, vancomycin, carbapenems and linezolid), have seen resistance observed only an average of ~ 5.6 years after first introduction [1]. Some proposed causes for the speed at which bacteria have managed this feat include: lateral gene transfer [1, 5], ineffectual clinical prescription regimes [6], and overuse in agriculture for livestock [7]. With cases of difficult-to-treat infections such as methicillin-resistant *S. aureus* (MRSA) on the rise, there has been a series of urgent calls from the World Health Organization for the development of new antibiotics [3]. Computational studies provide an opportunity to aid this effort through data analysis and hypothesis generation methods to assist in novel drug development.

1.2 Problem Statement

The goal of this thesis is to obtain a better understanding of how physicochemical properties (collectively referred to as "features" throughout this work) relate to a naturally occurring family of antimicrobial peptides (AMPs) with innate activity. This thesis details a novel process for characterizing an extensive list of position-based amino acid features. It then utilizes machine learning and feature selection to rank their relative importance for

AMP activity. While the method can be modified for use with any type of AMP, focus is placed here on cathelicidins- an important family of helical AMPs found in humans with great potential for new drug development applications. At the time of this writing, the findings of this thesis have yet to be verified through experiment and should be considered theoretical in nature until their biological relevance has been confirmed.

1.3 Thesis Overview

This thesis elucidates key physicochemical features in AMPs by investigating activity-related information which does not rely on specific, and currently incomplete, domain-specific understanding of AMP-action. The feature construction process is based on the AAIndex [8], an extensive collection of documented physicochemical properties of amino acids. Focus is placed on position-based physicochemical properties, rather than overall sequence, due to the high sequence diversity found in AMPs.

The general relevance of the features proposed here are established through a comparison with previous work that focuses on AMP recognition. While many methods exist for AMP recognition, direct comparison with these methods is challenging. Many datasets contain peptides of variable length, synthetic AMPs of unusual sequence composition, and some negative datasets do not explicitly have member peptides listed [9–11]. Comparison is performed by classifying the dataset provided in [12], which contains variable-length sequences and AMPs from a variety of families. The features presented here, when applied to two separate test sets, allows a support vector machine (SVM) to obtain average accuracies of 80% and 86% and Matthews correlation coefficients of 0.65 and 0.74. This is achieved without incorporating any additional features used in [12] aside from peptide length.

However, the primary focus of this thesis is on the analysis of the cathelicidin family of AMPs. Similar to other work [9], fixed-length subsequences are used which focus on the N-

and C-termini which have been shown relevant for activity in cathelicidins [13]. Differences in secondary structure characteristics unrelated to activity are removed through the careful construction of a helical decoy dataset. SVM-based classification results achieve accuracies of 93-95% and Matthews correlation coefficients of 0.78-0.83 for recognition of cathelicidins over a decoy dataset. To elucidate which features are most important in characterizing cathelicidin activity, ranking is employed by the SVM through their F-scores. Additional statistical analysis is applied to features under the influence of enzymatic cleavage at the N-terminus of active cathelicidins fragments. This provides insight into features of direct relevance for antimicrobial activity and allows features which may be more involved with cleavage to be weighed down if desired for future AMP design studies.

The feature profiles presented in this thesis are a first step towards obtaining a better understanding of what confers cathelicidins their activity and aiding the modification or design of novel cathelicidin-like AMPs.

1.4 Thesis Structure

The structure of this thesis begins with useful background information in the next chapter. This is followed by chapters detailing the methods, results, and finally a discussion on possible biological implications and future directions to expand this work. The appendix provides a more extensive collection of statistical results, feature scores, and feature profiles which are also available through a web interface at:
`http://binf.gmu.edu/dveltri/cgi-bin/cath_explorer.cgi`.

1.5 Related Publications

Much of the work in this thesis was presented at the 5th International Conference on Bioinformatics and Computational Biology (BICoB2013) on March 6th, 2013 in Honolulu, Hawaii [14]. A poster and extended abstract were presented at the 2012 IEEE International

Conference on Bioinformatics and Biomedicine Workshops (BIBMW) in Philadelphia, Pennsylvania. [15].

Chapter 2: Background Introduction

2.1 Antimicrobial Peptides: Prime Drug Candidates

Antimicrobial peptides (AMPs), also referred to as host defense peptides, constitute an array of protein families which play a critical role in innate immunity across all phylogenies [16]. Some examples of major families include: β-defensins, brevinins, caerins, cathelicidins, magainins and stylins [17–21]. Some example 3D structures of AMPs are shown in Figure 2.1.

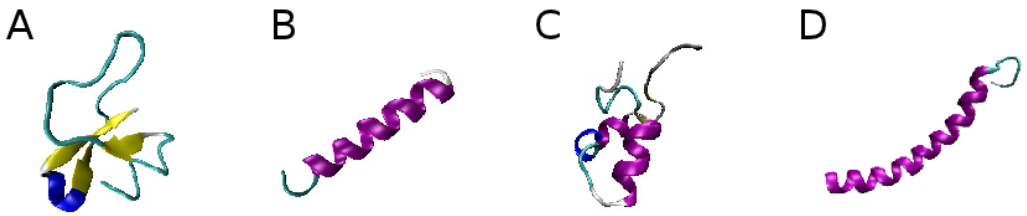


Figure 2.1: Some example AMP crystal structures. A) β-defensin 1 from *Homo sapiens*, PDB ID:1IJV. B) Magainin 2 from *Xenopus laevis*, PDB ID:2MAG. C) Aurelin from *Aurelia aurita*, PDB ID:2LG4. D) Cathelicidin LL-37 from *Homo sapiens*, PDB ID:2K6O.

Researchers have been looking with great interest at AMPs due to their resilience against bacterial resistance; a result of millions of years of coevolution with their targets [13, 17]. AMPs demonstrate a variety of killing mechanisms effective against bacteria (both gram-positive and negative) and fungi. Many also demonstrate an ability to break

down lipopolysaccharide (LPS) endotoxins [22]. As these peptides are often short in length, they are amenable for protein synthesis and manufacturing, making them prime candidates for drug research [21, 23].

2.1.1 AMP Structure

Active AMP fragments are generally less than 100 amino acids in length [17]. While a number of structural classification systems are found in the literature, The Antimicrobial Database (APD) Version 2 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 AA, disulfide bridge (no 3D structure) and unknown 3D structure [22, 24]. To date, amphipathic α-helical, β-sheet and peptides with highly biased amino acid composition have been the most extensively researched [17]. This thesis will focus solely on cathelicidins, a family of mostly amphipathic α-helical structure as they are well-studied and found in humans [17].

2.1.2 AMP Attack Mechanisms

AMPs have been found to play a number of different roles in killing bacteria. Interference with DNA replication, disabling membrane receptors, and signaling for adaptive immune responses have all been observed [22]. One of the most common modes of attack by helical AMPs is believed to be membrane permeabilization [17, 22, 25]. Circular dichroism studies support the notion that most α-helical AMPs lack structure while in solution and prior to contact with a membrane [19]. Amphipathic AMPs are interspersed with cationic and anionic regions which allow them to approach their target in a parallel fashion as depicted in Figure 2.2. This is accomplished via side chain packing of cationic residues, where rotation allows them to face the anionic bacterial phospholipid bilayer and facilitate an electrostatic attraction to the membrane [17, 25]. Accordingly, peptide charge is an important characteristic for AMP activity [19]. Eukaryotic membranes generally escape

AMP attack thanks to the presence of cholesterols; which, disrupt negative charges on their surface [17]. However, hemolytic action on erythrocytes can occur from AMP members with high positive charge given sufficient concentrations [17].

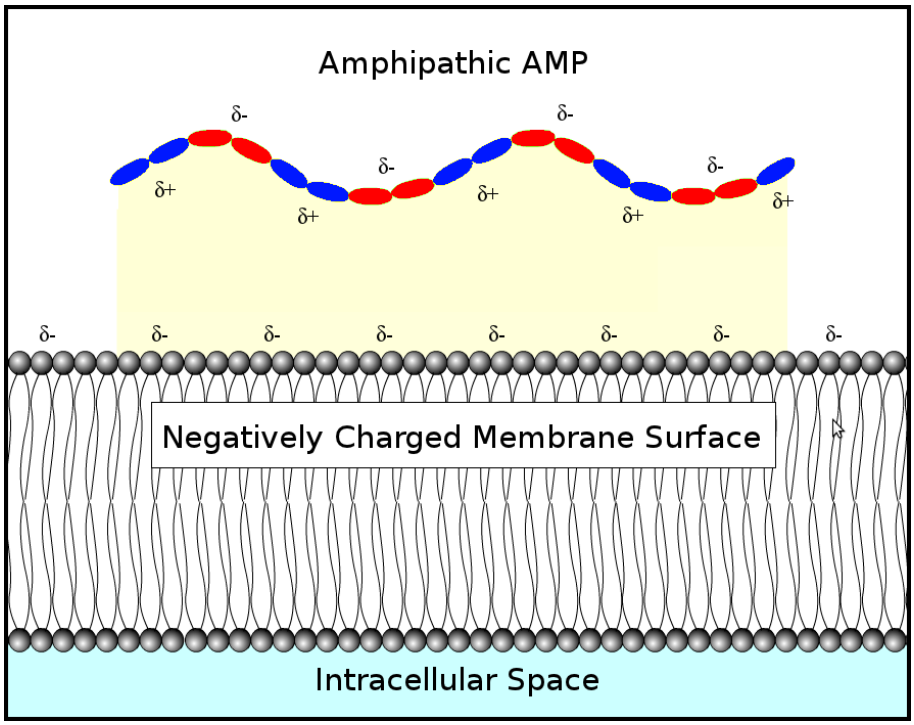


Figure 2.2: Amphipathic α-Helical Peptides Approach Bacterial Membranes in a Parallel Fashion.

Upon contact with the membrane, α-helical AMPs convert from a coil to a helical structure [25]. While this process appears to damage the membrane, the exact method and number of peptides required for lysis has yet to be confirmed. Some proposed models from the literature are now reviewed below.

Carpet Model

The carpet model, illustrated in Figure 2.3A, proposes that individual peptides are only capable of disordering small patches of phospholipids. In order to sufficiently damage a membrane, a large number of peptides are required. As peptides aggregate over the membrane surface, they cause pieces to dissociate and effectively shred the bacterial membrane apart [25, 26].

Barrel-Stave Pore Model

The barrel-stave pore model, illustrated in Figure 2.3B, calls for peptide cooperation in far smaller quantities. Peptides begin by aggregating on the membrane surface and forming dimers and/or trimers. These begin to permeate the membrane causing an indentation, which is gradually enlarged by the addition of more peptides. Eventually, a circular channel is formed perpendicular to the membrane surface [25, 27].

Toroidal Pore Model

The toroidal pore model, illustrated in Figure 2.3C, is a special case of the barrel-stave pore model. Peptides as long, or longer, than the length of the lipid bilayer (e.g. megainins) permeate the membrane and begin to form circular channels as other members aggregate [25].

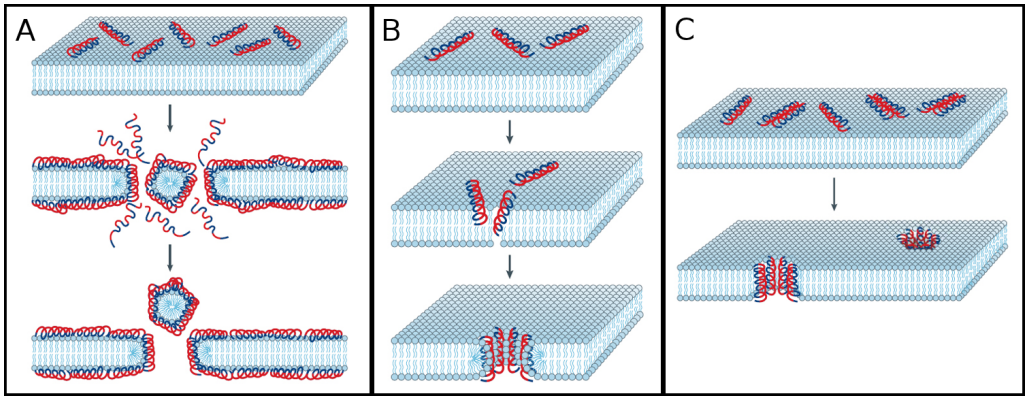


Figure 2.3: Some illustrations of predicted cationic AMP attack mechanisms. A) The Carpet Model. B) The Barrel-Stave Pore Model. C) The Toroidal Pore Model. Illustrations adapted from [28].

Leaky Slit Model

While not yet verified by experiment as of this writing, a study by Mahalka and Kinnunen (2009) notes a number of similarities between how amyloid fibrils and the AMPs temporin B and L permeate a membrane through aggregation [29]. This could also have implications for similar amphipathic AMP types such as cathelicidins.

Wang's Magnet Model

Work by Wang (2008) using KR-12, a truncated version of the human LL-37 (Figure 2.1D), has suggested that clusters of positively charged residues within an AMP may be attacked to negatively charged phosphatidylglycerols (PGs) in a magnet-like fashion [30]. In sufficient numbers, this could result in PG-clusters to form on the membrane surface which could then disrupt voltage-dependent inter-membrane proteins and other related processes (e.g. potassium channels) [30, 31].

2.1.3 The Cathelicidin Family of AMPs

Cathelicidins are one of only a few major AMP families found in humans; making them an important family for study. Expression of cathelicidins in humans has been observed by macrophages, lymphocytes and keratinocytes and their presence has been recorded in semen, epithelial, upper respiratory, and vaginal cells [17, 30, 32, 33]. Abnormally low concentrations of the human cathelicidin LL-37 have been associated with Morbus Kostmann, a disorder causing high infant mortality from bacterial infection [17]. The skin disorder rosacea has also been linked to cathelicidin over-expression [33–35]. Mature cathelicidin peptides are usually around 15-55 residues in length, making them viable for both computer simulations and synthesis in the wet lab [23]. The directed design of cathelicidins has been a challenge for the scientific community as their fragments involved with antibacterial activity are quite heterogeneous in sequence [21, 23, 36].

Cathelicidin Structure

The presence of a conserved "cathelin domain" (ExPASy PROSITE ID: PDOC00729) is seen in all cathelicidin precursors and gives the family their name [13, 21, 23, 36, 37]. The layout of a cathelicidin precursor protein and how it is processed (discussed below) can be seen in Figure 2.4.

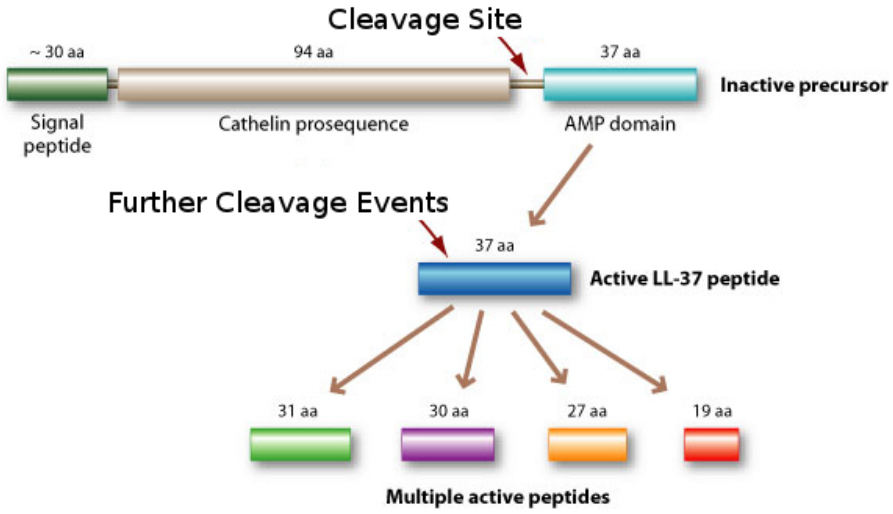


Figure 2.4: General cathelicidin precursor structure and cleavage processing using human LL-37 as an example. Illustration adapted from [38].

Precursor Enzymatic Cleavage

Cathelicidins remain inactive in their precursor form until they are needed. Similar to other proteins with a precursor form (e.g. insulin), this allows for a stockpiled reserve which can quickly be called into action to assist with a bacterial infection. The mature (active) peptide fragment, located at the C-terminus of the precursor, is activated after the peptide is cleaved with an elastase-like serine protease. This mature peptide may undergo further location-dependent cleavage events to generate even shorter peptides with specialized functions [30, 33]. Consequently, amino acids which lay within the protease cleavage window have to compromise between facilitating cleavage and preserving antibacterial action. As the synthetic production of AMP-derived drugs focuses on the later, isolating which amino acid physicochemical features are involved with each process is desirable.

2.2 Machine Learning Classification and Validation

Machine learning 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 [39]. Various machine learning techniques have successfully been used in a number of bioinformatics applications [40]. For example, artificial neural networks have predicted proprotein convertase cleavage sites [41], genetic algorithms and clustering have detected gene expression levels [42], and decision trees have been applied to protein secondary structure prediction [43]. Machine learning is especially helpful when dealing with complex, high-dimensional, datasets.

2.2.1 Dataset Classification

Classification, a common machine learning task, involves the assignment of 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 assignment to three or more labels (e.g. high/medium/low). A classifier is first trained on positive and negative examples from a dataset. Learning is considered "supervised" when the training examples have labels of known class [39]. After training, predictions can then be made on an unlabeled testing set where each observation is assigned to a class.

2.2.2 Dataset Validation

If the classes of the testing set are known, performance can be evaluated by comparing the actual classes against those predicted by the 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). 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. Some common performance metrics used for machine classification are briefly outlined below.

Accuracy

Accuracy (ACC) represents how repeatable measurements are when conditions remain constant [44]. In the context of classification, this represents how accurately a method can predict sample classes and is defined as:

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \times 100. \tag{2.1}$$

Values can range from 0 to 100 percent. Larger values equate to better classification performance.

Mathew's Correlation Coefficient

Matthew's correlation coefficient (MCC) [45] is used to evaluate binary classification performance and is defined as:

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}. \tag{2.2}$$

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

Receiver Operating Characteristic

A receiver operating characteristic (ROC) [46] curve provides a convenient graphical representation of classification accuracy and can be seen in Figure 4.1. The curve is generated

using predictions provided by a classifier 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 [46]. The ROC score refers to the area under the curve. Random ranking is expected to yield a score ~ 0.5 which would be represented as 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. The ROC score reaches 1 if the SVM correctly places all of cathelicidins above the threshold. Conversely, scores less than 0.5 bend the line towards the lower right corner as classification performance decreases.

Sensitivity and Specificity

Sensitivity assesses type II error and is defined as:

$$\text{Sens.} = \frac{TP}{TP + FN}. \tag{2.3}$$

Specificity assesses type I error and is defined as:

$$\text{Spec.} = \frac{TN}{TNs + FP}. \tag{2.4}$$

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

2.3 Support Vector Machines

Currently, one of the most popular machine learning approaches for classification has been support vector machine (SVMs) [47,48]. A few example implementations for bioinformatics include cancer classification [49, 50], protein structure assignment [51] and predicting DNA translation initiation [52]. While a number of excellent resources are available thoroughly covering the theory and mathematics behind SVMs [47, 48, 53–55], a brief description is provided here of how they work.

SVMs are usually used to discriminate between two classes. It accomplishes this by finding an optimal hyperplane which separates the classes as illustrated in Figure 2.5. "Optimal" in this cases refers to the hyperplane with the widest margin of separation between itself and the nearest neighboring samples (known as support vectors) on either side of it. Given sample data to find such a hyperplane, the class of a testing point (or set of points) can be predicted based on which side of the hyperplane it falls. As described in [53], if given a training set with labels $(x_i, y_i), i = \{1, \ldots, m\}$ where $x_k \in \mathbf{R^n}$ and $y \in \{1, -1\}^m$, an SVM optimizes the following problem [48]:

$$\begin{aligned} \underset{\omega, b, \xi}{min} \quad & \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i, \\ \text{subject to} \quad & y_i(\omega^T\phi(x_k) + b) \geq 1 - \xi_i, \\ \text{and} \quad & \xi_i \geq 0. \end{aligned} \tag{2.5}$$

C is a penalty parameter for training errors, ξ is a "soft error" penalty for observations which fall inside the margin [55], and ϕ is a function used to map training data into a higher dimension. This mapping is known as the "kernel trick" and defined as $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$. Two common kernels are the liner kernel $K(x_i, x_j) = x_i^T x_j$ and the radial basis function (RBF) kernel $K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right), \gamma > 0$ (where γ is a tuning parameter). When given an observation x_i for testing, a prediction is made

according to:

$$\text{f(x)} = sgn(\omega^T \phi(x_i) + b). \tag{2.6}$$

Figure 2.5 illustrates a simple two-class svm problem. A linear SVM has drawn a decision boundary to separate the blue class onto the -1 side of the decision boundary and the red class onto the 1 side. The shaded support vectors define the margin which an SVM attempts to minimize. Note how the rest of the data points do not play a role in defining the decision boundary and could be removed without hurting classification accuracy [55].

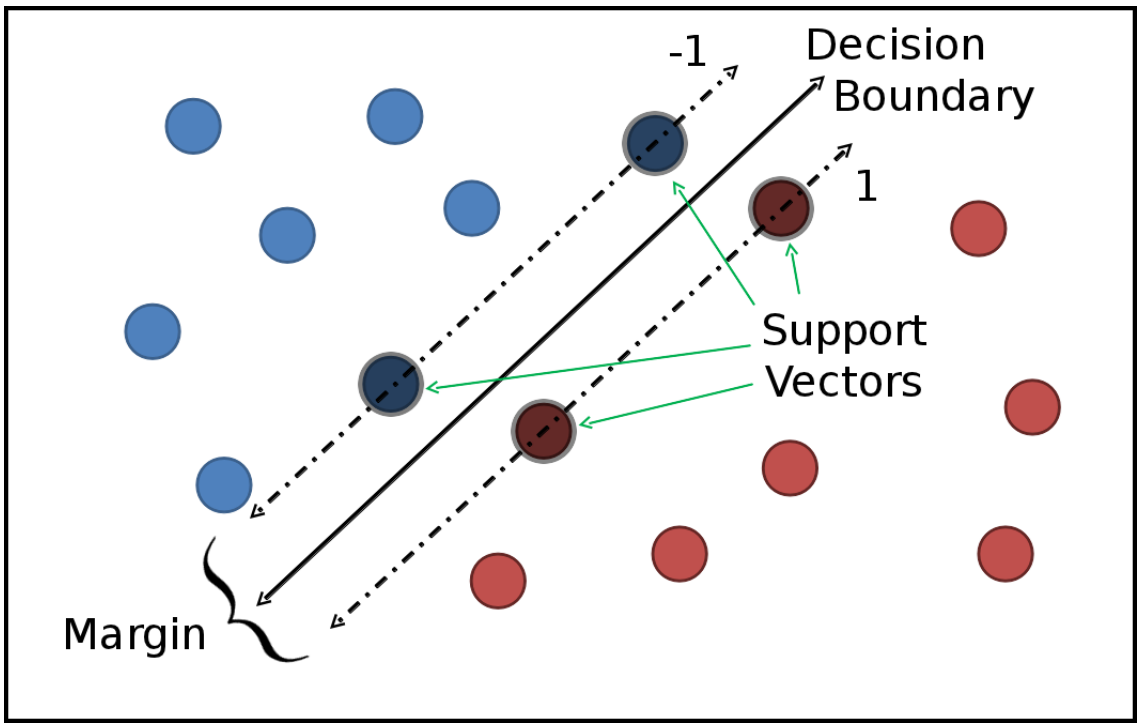


Figure 2.5: A simple two-class SVM example.

2.4 Feature Ranking and Reduction with F-select

The F-score that SVM models associate with support vectors provide a measure of discriminating power for features. The "F-select method" is a univariate approach which combines this scoring measure with SVM implementations and was a strong performer in the Neural Information Processing Systems 2003 Feature Selection Challenge in ranking features and creating a minimum feature set [56]. In the same manner, this thesis employs the F-select method to elucidate top ranking features related to cathelicidin AMPs. Briefly, as described in [56], the F-score measures the discrimination of two sets of real numbers. Given training vectors x_k, where $k \in \{1, \ldots, m\}$, with n_+ and n_- denoting the respective number of positive and negative instances, the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x}_i^+ - \bar{x}_i)^2 + (\bar{x}_i^- - \bar{x}_i)^2)}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^+ - \bar{x}_i^+)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^- - \bar{x}_i^-)^2}$$

In the above equation, $\bar{x}_i$, $\bar{x}_i^+$, and $\bar{x}_i^-$ are the average of the i^{th} feature of the whole, positive, and negative datasets, respectively. Similarly, $x_{k,i}^+$, is the i^{th} feature of the k^{th} positive instance, and $x_{k,i}^-$ is the i^{th} feature of the k^{th} negative instance. The numerator measures the discrimination between the positive and negative sets, whereas the denominator measures the discrimination within each of the two sets. A higher score equates with a feature having better discriminatory power.

F-scores can also be used to obtain a minimum feature set as in [56]. Features with the highest F-scores are added iteratively (forward selection), and classification performance is evaluated each round. The process continues until a decrease in performance is detected. After repeated trials, an average F-score threshold is determined for the lowest validation error, and features below this cutoff are removed to create a minimum feature set. Further details into this protocol can be found in [56].

2.5 Previous Machine Learning Work on *in-Silico* AMP Classification

Many machine learning methods have been devised to address AMP recognition. This is generally performed with all AMP families being considered as one class [9, 11, 12]. Table 2.1 summarizes the state of the art. However, direct comparisons are hard to draw due to the great diversity amongst algorithms employed, features considered, and the positive and negative datasets used to demonstrate AMP recognition. Furthermore, as these methods are used as "black boxes" for automatic recognition, their value for drawing rules for designing novel AMP-based drugs in a wet lab setting has yet to be demonstrated. It is still unclear how one can best modify the sequence of a peptide to improve antimicrobial activity [22].

Table 2.1: Summary of AMP Prediction Algorithms and Datasets from [12]

Algorithm	**MCC**			**Database**
	Training Dataset	Validation Dataset	Testing Dataset	
HMM [57]		0.98		AMPer
HMM [58]		0.88		RANDOM
ANN [59]		0.60		CAMEL QSAR
DA [60]	0.75		0.74	CAMP
RF [60]	0.86		0.86	CAMP
SVM [60]	0.88		0.82	CAMP
SVM [9]			0.84	AntiBP2
ANFIS [12]		0.94		APD2
ANN [12]		0.85		APD2

Table 2.1 shows that great recognition accuracy can currently be obtained. For instance, recent work in [12] achieves a high MCC value of 0.94. Interestingly, the high recognition accuracy of AMPs is obtained with only two features calculated over the peptide sequence, length (number of amino acids) and propensity for aggregation. Other interesting physicochemical properties proven useful in the wet lab for modifying antimicrobial activity of various AMPs (such as, hydrophobicity, propensity for certain secondary structures, and more) were shown not to be important for automatic recognition [12].

Other attempts to overcome high AMP sequence diversity have focused, for instance, on a fixed number of terminal amino acids and employing simple features based on amino-acid composition [9, 10, 61]. This approach discriminates between AMPs and decoys 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 [16, 17].

Differences between what is shown important in computational analysis versus the wet lab highlight the question as to what the employed features are really capturing and what role, if any, the employed negative datasets play. A negative dataset of poor quality can potentially bias the model towards a spurious set of features unrelated to activity.

This thesis takes a step back from generalizing AMPs as a single class and argues that better progress can be made by focusing on a specific and well-studied family of AMPs. In doing so, a decoy dataset of higher quality can be constructed to better control for bias.

Chapter 3: Materials and Methods

The overall approach used in this thesis is outlined in Figure 3.1. A more detailed description begins with dataset preparation, followed by the machine learning and feature ranking approach employed. Finally, a statistical approach is detailed which identifies the top N-termini features reported by the SVM that are not statistically different from non-AMP proteins also cleaved by neutrophil elastase. This information is used to mark and lower confidence in such features to aid AMP wet laboratory studies prioritizing antimicrobial activity for design.

Figure 3.1: Overview of Thesis Methodology

3.1 Dataset Generation

The following describes the preparation of the N- and C-termini datasets for recognition of cathelicidins. Two positive datasets are constructed, each consisting of 18-residue long subsequences of the N- and C-termini extracted from mature (active) fragments of natural cathelicidin sequences. This creates a separate SVM model for each termini due to the fact the contribution to activity of each termini, while not fully understood currently, is believed to differ [13, 31]. A third dataset is employed which consists of neutrophil

elastase-cleaved substrates. Since the first four N-termini residues in a cathelicidin are also important for enzymatic cleavage, it is important to differentiate features that may be important for antimicrobial activity as opposed to cleavage.

3.1.1 Positive Datasets of Mature Cathelicidins

A total of 45 mature (active) cathelicidin sequences with no more than 90% sequence identity have been collected from the Antimicrobial Peptide Database (APD2) [24, 62], UniProt [63] and the literature. Protegrin-1 and related sequences (Swiss-Prot: UniRef90_P32194) are excluded as evidence suggests these form a β-sheet upon membrane contact [64]. It should be noted, cathelicidin EA-CATH2 from [65] is included in the positive datasets but was later found to have not been explicitly tested for antimicrobial activity. However, as the authors in [65] classify the sequence as a cathelicidin due to its sequence similarity, the author here does not feel results have been negatively impacted.

As the SVM used for the later classification step requires fixed-length vectors, focus has been placed on terminal residues to resolve the issue of varying peptide length. Two datasets are constructed from mature peptides. One contains 18 consecutive residues of the N-termini, and the other from the C-termini. The length limit of 18 residues is due to the maximum scan-length allowed by HeliQuest, a server used in forming the matching negative datasets. Appendix Table A.1 identifies the positive cathelicidin sequences used for the N-termini dataset while Appendix Table A.2 provides those for the C-termini dataset. The first 4 amino acids of each sequence in the N-termini set form the positive dataset used for analyzing features related to enzymatic cleavage.

3.1.2 Negative Datasets of Decoy Sequences

Two different negative datasets of 18-residue long sequences have been constructed for the two positive datasets of N- and C-termini sequences. Rather than build these at random, the negative sequences are designed to be helical through the HeliQuest server [66]. This

allows for non-AMP decoys which share structural characteristics with cathelicidins and helps prevent top discriminating features from simply exploiting structural differences. The reason for the two separate negative datasets is that each termini has a different helical profile.

The N-terminus has a consensus pattern of KRR[RL]GLF[RL][KR]KAR[KE]KIKKG (amino acids in brackets represent an equal number of observations at that position). This results in 16 possible sequences based on ties at positions 4, 8, 9, and 13. Each is submitted to the HeliQuest "sequence analysis module" with default settings to identify important properties, such as hydrophobicity, hydrophobic moment, and net charge. These results are then passed to the screening module, with "proline accepted at i, i+3 / n−3, n." Results are limited to unique UniProt entries from the human proteome, and the set is further reduced to a sequence identity of less than 50% [12]. UniProt sequence annotations allow for excising entries mentioning antimicrobial, antifungal, anti-viral or cytotoxic activity. So as not to choose secreted proteins, the "cellular location" search option is further limited to "cytoplasm" [9]. From the resulting sequences, 180 are randomly drawn (resulting in a 1:4 positive-to-negative sample ratio). Decoy sequences for the N-termini dataset are identified in Appendix Table A.1.

The above process is repeated to obtain 180 C-termini decoys using the consensus pattern KIGQKIKDFLGI[LP]VPRTG. The selected decoy sequences for the C-termini dataset are available in Appendix Table A.2.

A third negative dataset has been constructed for cleavage analysis and consists of neutrophil elastase substrates. 45 non-AMP substrates are extracted from the PMAP-CutDB Proteolytic Event Database [67], provided as 8-mers centered about the cleavage site. The 4 residues upstream of cleavage are discarded. The analysis below compares features of this set with those over the first 4 residues for the 45 peptides in the N-termini cathelicidin dataset. The objective being to mark or discard features identified as important by the SVM but found present in the substrate dataset. These overlapping

features essentially cannot be determined to be statistically more relevant for antimicrobial activity compared to cleavage. Positive samples correspond of the first 4 amino acids of the cathelicidin sequences identified in Appendix Table A.1. The PMAP-CutDB IDs for the negative samples used in this dataset are provided in Appendix Table A.3.

3.1.3 Feature Construction for Physicochemical Properties of Amino Acids

Sequences in the above datasets are converted into numeric vectors by expanding each residue position into a list of corresponding physicochemical properties for that given amino acid. The features considered are all known physicochemical properties of amino acids documented in the AAIndex (AAIndex1,Vr.9) [8]. The AAIndex is a collection of 544 quantified amino-acid physicochemical properties obtained from the literature. Removing 13 entries which contain "NA" values leaves 531 properties per amino acid. This set, while comprehensive, presents problems for long sequences. All 531 features are employed for the enzymatic cleavage dataset since it only considers four residue positions. This essentially converts each sequence into a numeric vector of $2124 = 531 \times 4$ elements. This feature list is reduced for the datasets with 18 residue-long fragments. Removing entries found to share ± 0.8 or greater correlation scores, both provided and defined in [8], reduces this set to 299 features. Accordingly, each 18-residue long termini fragment can now be converted into a vector of $5382 = 299 \times 18$ elements. Additional information is included into the vectors by arranging them as: $\{C, (R_1, X_1), \ldots, (R_n, X_1), (R_1, X_2), \ldots, (R_n, X_{299})\}$, where C is a class label, R_i is a residue over n positions, and X_j is an AAIndex entry from the 299 entries considered. This format allows any feature to be traced back to a specific physicochemical property at a particular residue position.

3.2 Preliminary Analysis of Feature Space

The similarity between the negative and positive datasets has been initially checked using both linear and nonlinear dimensionality reduction techniques.

3.2.1 Principal Component Analysis

Principal Component Analysis (PCA) [68], a classic linear dimensionality reduction technique, is applied to the N- and C-termini datasets. Results are reported in Figure 3.2 with the first 5 principal components shown. PCA is not able to adequately separate cathelicidins as a whole from non-AMPs.

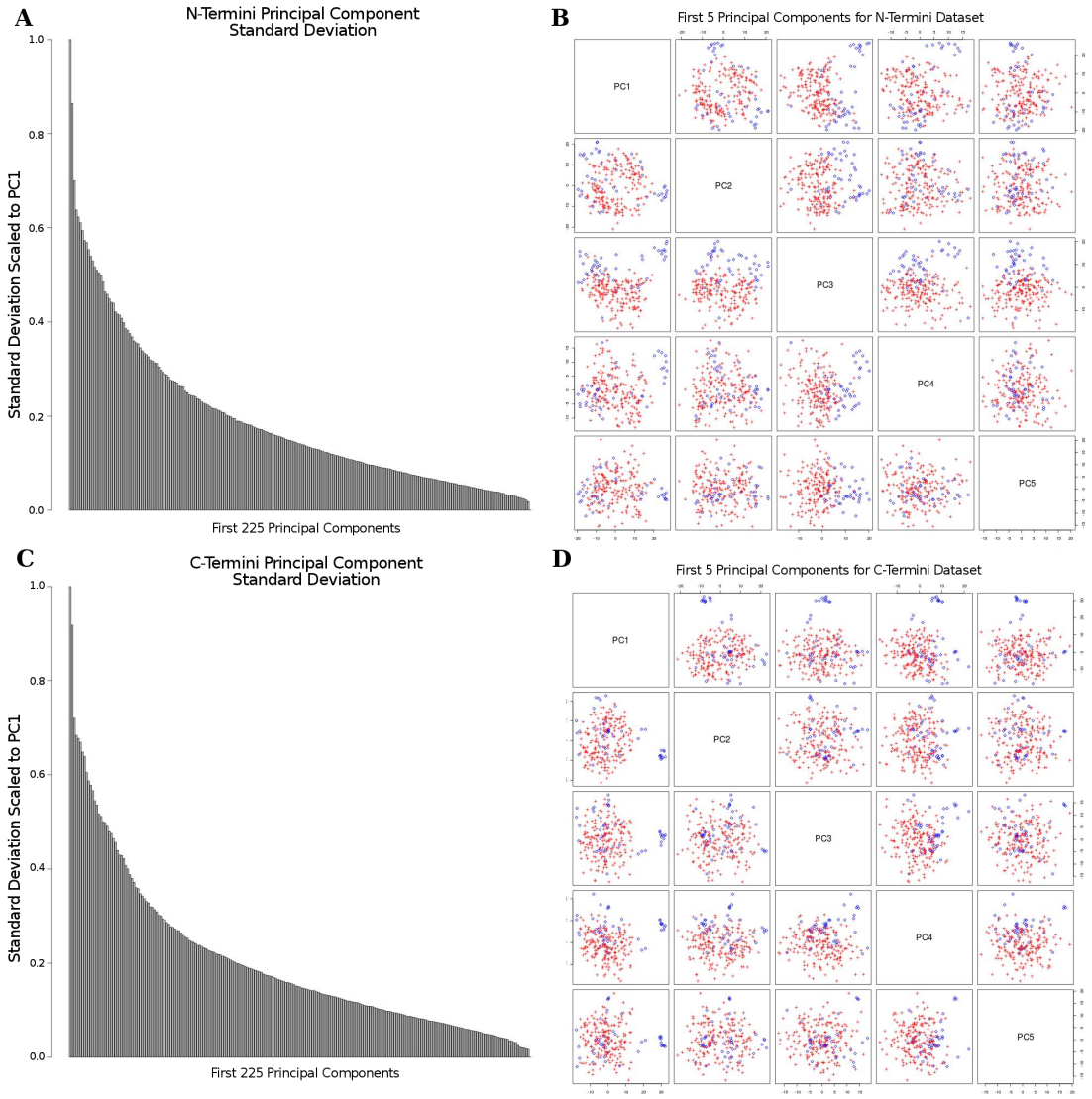


Figure 3.2: Principal Component Analysis [68], a linear dimensionality technique, is applied separately to the N- and C- termini datasets to reduce the feature space to R^2. **(A)** Standard deviation of the first 225 principal components (PCs) scaled to $PC1$ for the N-termini dataset, **(B)** Pairs plot of $PC1 - PC5$ of the N-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) datasets, **(C)** Standard deviation of the first 225 principal components (PCs) scaled to $PC1$ for the C-termini dataset, **(D)** Pairs plot of $PCs1 - 5$ of the C-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) dataset.

3.2.2 Local Linear Embedding

Local Linear Embedding [69], a nonlinear dimensionality technique, has also been applied to the N- and C-termini datasets. Results with $k = 10$ can be seen in Figure 3.3. Note that the density of the C-termini data points at the center of plot **B** make it difficult to see the many cathelicidin points (blue) intermixed amongst the decoys (red), however, the same sample size is used as in plot **A**. While the cathelicidin class as a whole can not be adequately separated at either termini, it is interesting to note how some cathelicidin members (including human LL-37 at both termini) do appear to distance themselves from the rest. LLE might be a useful tool for analyzing differences within the cathelicidin family.

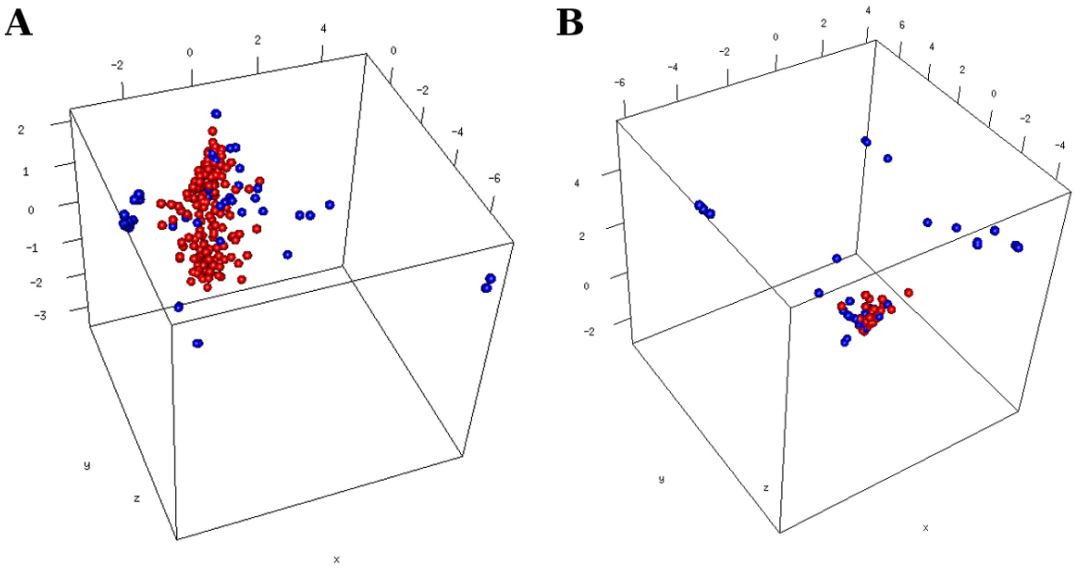


Figure 3.3: Local Linear Embedding, a nonlinear dimensionality reduction technique [69], is applied using a variety of k-neighbors (not all data shown) to reduce the feature space to R^3. $k = 10$ appears to generate better discrimination between classes for both termini datasets. **(A)** A $3D$ representation of the mapped feature space of the N-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) dataset with $k = 10$ neighbors. **(B)** A 3D representation of the mapped feature space of the C-termini decoy (red, $n = 180$) vs. cathelicidin (blue, $n = 45$) dataset with $k = 10$ neighbors.

The inability for PCA and LLE to adequately separate the two classes, at either termini, demonstrates the wide diversity amongst cathelicidins and the high level of classification difficulty for the problem at hand.

3.3 SVM Classification and Feature Selection on Termini Datasets

Two SVM models are trained separately on the N-termini and C-termini datasets using LibSVM [70]. While results reported utilize the RBF kernel, testing with the linear kernel yield only slightly lower performance. Kernel parameters and the SVM cost function

are tuned through the standard grid search mechanism [71] using *grid.py* included with LibSVM. Features are then scaled from -1 to 1, as recommended, using the *svm-scale* program.

3.3.1 Cross-Validation and Performance Measurements

Reported results are obtained after 3-fold cross-validation on each of the termini datasets. Each training set is randomly divided into 3 subsets of equal size. The model is then trained on 2/3 of the data and tested on the remaining subset. This process is repeated so that each fold participates separately as a training set, but no training sequences are used as testing samples within the same fold. Classification performance measurements are reported as averages for the 3-fold validations in terms of ACC and MCC as defined in section 2.2.2.

3.3.2 Feature Selection Based on F-score Ranking

The F-score that SVM models associate with support vectors provides an estimate of the relative importance (discriminating power) of features. This forms the basis of the F-select method from [56] which is described in more detail in section 2.4. F-scores are employed as a feature selection criterion in this thesis to obtain a minimum feature set as in [56]. Essentially, the method iteratively adds features with the highest F-scores and evaluates classification performance each time. The process continues until a set of features with the best performance is detected. After repeated trials, an average F-score threshold is determined for the lowest validation error, and features below this cutoff are removed to create a minimum feature set. Further details into this protocol can be found in [56] and implementation is freely available as *fselect.py* at:
`http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/fselect`.

3.4 Cleavage Site Analysis

A statistical approach is used to evaluate if features of cleavage site amino acids (N-termini residues 1−4) in cathelicidins are different from those in a set of natural, yet non-AMP, neutrophil elastase substrates. The dataset of 45 cathelicidin and non-AMP substrates described above is utilized.

Each feature is treated separately, and visual assessment using Q-Q plots [72] reveals few to be normally distributed. The Brown-Forsythe test is conducted [73] using the *lawstat* package [74] to assess the quality of variance between feature populations of the two datasets. Features with differing variance ($p < 0.05, \alpha = 0.05$), and shown to be statistically independent by this test, are removed. Remaining features are passed on to a second round of assessment with the Mann-Whitney-Wilcoxon Test (using the *exactRankTests* package [75]). Features shown to be statistically independent by this test ($p < 0.05, \alpha = 0.05$) are then removed. The final remaining features represent those that cannot be confidently associated with antimicrobial activity over protease specificity. These features are listed in Appendix Table A.3 and biologists, or other domain experts, conducting future AMP tests may want to treat these features with additional care.

Chapter 4: Results

4.1 Setup and Performance Statistics

Experiments were conducted on an Intel Core2 Duo machine with 4GB RAM and 2.66GHz CPU. Reported statistics include accuracy (ACC), Matthew's correlation coefficient (MCC), sensitivity, specificity, and receiver operating characteristic (ROC) which are detailed in section 2.2.2.

4.2 Comparison with Related Work

In order to first establish that the generated features in this work are relevant to antimicrobial activity, a comparison was performed using the variable-length sequences in the dataset provided by Fernandes et al. (2012). This dataset consists of 231 peptide sequences (115 mixed-family AMPs from the APD2 and 116 non-AMPs) and are detailed in [12]. As these sequences are of variable length, the approach for feature generation described in section 3.1.3 had to be modified. In this case, the 299 AAIndex features were averaged over the number of residue positions in each respective peptide. Peptide length was added as a 300th feature and vectors were arranged as follows: $\{C, \overline{X}_1(n), \ldots, \overline{X}_{299}(n), n\}$, where C is a class label, n is the number of residue positions for a given peptide and $\overline{X}_i(n)$ is the mean value for one of the 299 AAIndex features across residues 1 to n.

Performance was compared to the ANN and ANFIS algorithms in [12]. The datasets were randomly partitioned as in [12] and results from the SVM-based approach used in this thesis is given in the context of two separate testing datasets of 58 peptides, each averaged over 10 runs using the RBF kernel. The comparison with results reported in [12]

is shown in Table 4.1. Columns 2 and 3 show that the approach here obtains average ACCs of 80% and 85.95% and MCCs of 0.64 and 0.74 on each of the respective test sets (the top 100 F-scores are provided in Appendix B.3).

Table 4.1: Classification Performance Comparison with Fernandes et al. (2012) Dataset

Method	**ACC(%)**	**MCC**
SVM Avg. AAIndex	Test Set1: 80.0 Test Set2: 85.9	Test Set1: 0.6462 Test Set2: 0.7356
Fernandes et al. ANN	90.9(overall)	Validation: 0.8320 Testing: 0.8268
Fernandes et al. ANFIS	96.7(overall)	Validation: 0.8868 Testing: 1.0000

These results are slightly lower than those obtained by ANN and ANFIS in [12], where reported ACCs are in the $90-96\%$ range, and MCCs are in the $0.85-0.94$ range. This is unsurprising considering that the ANN and ANFIS implementations in [12] were designed for full-length peptide features in mind. However, the comparison demonstrates that the physicochemical properties employed here allow SVM-based AMP recognition, on a variety of families, with accuracies similar to other SVM-based work [11]. Unfortunately, code for the ANN and ANFIS implementations used in [12] is not available and attempts at reproducing their results were unsuccessful. As such, applying the feature set described here using the exact algorithm implementation in [12] was not possible. The remaining results now focus on the cathelicidin datsets.

4.3 SVM Performance on Cathelicidin Termini Datasets

The N- and C-termini datasets are each classified using SVM paired with the RBF kernel. Evaluation is performed in the context of 3-fold validation. Training sets contain 30 and 120 respective cathelicidins and decoys. Test sets are respectively composed of 15 and 60 cathelicidin and decoy sequences. See section 3.3 for more details on the SVM implementation. Summary statistics for each termini dataset are presented as the average over the 3 folds and can be seen in Table 4.2. Columns 3 and 4 show ACCs of 94.67% and 93.33% and MCCs of 0.83 and 0.78 for the N- and C-termini datasets respectively.

Table 4.2: SVM Average 3-Fold Performance on N- and C-termini Datasets

Dataset	**Sen.(%)**	**Spec.(%)**	**ACC(%)**	**MCC**
N-Term.	95.21	94.80	94.67	0.8277
C-Term.	90.56	93.75	93.33	0.7761

Average ROC curves (defined in section 2.2.2) are shown in Figure 4.1. ACC values correspond to the area under the ROC curves and suggest that the employed features here allow the SVM to achieve high classification accuracy.

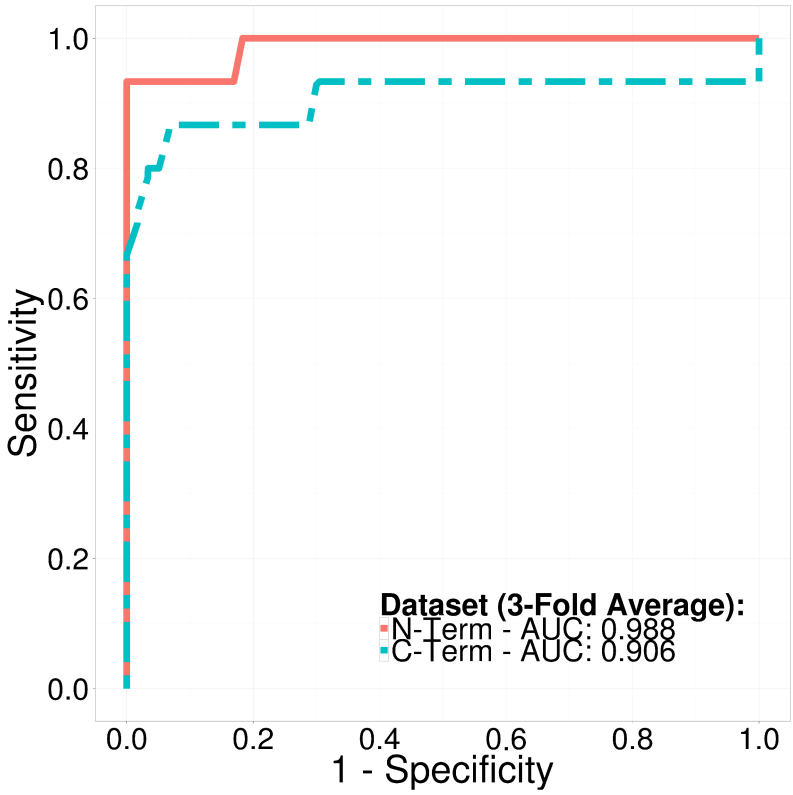


Figure 4.1: ROC Curves for SVM Performance on N- and C-termini Datasets

The rest of the analysis focuses on highlighting the features relevant for antimicrobial activity. To do so, the analysis explicitly discounts features important for cleavage and then employs ranking based on SVM-obtained F-scores to order features based on their relevance for activity.

4.4 Cleavage Site Analysis

This analysis, described in section 3.4, is used to compare the first 4 N-termini residues in the positive cathelicidin dataset to a set of non-AMP substrates also cleaved by neutrophil

elastase. A total of 2124 (4×531) features are each independently tested. The Brown-Forsythe test removes 510 features due to differing group variance ($p < 0.05$). Remaining features are fed to the Mann-Whitney-Wilcoxon test (two-tailed), where 77% are found not to be significantly different ($p > 0.05$). These 1243 collective features potentially encode signals prioritizing protease specificity, rather than antimicrobial activity, and can be found in Appendix C.

4.5 F-score Feature Ranking

The 299 non-redundant physicochemical properties described in section 3.1.3 are used for each residue position to generate feature vectors for SVM training. F-scores obtained by the SVM are analyzed and the selection procedure in section 3.3.2 is employed to elucidate the top features with high discriminatory power.

When trained on the full (aggregate) N-termini dataset, the F-select procedure reports a maximum ACC of 96% using a minimum feature set based on 936 features (out of 299×18). Using the full C-termini dataset, a maximum ACC of 96.4% is obtained using 520 features. When the ranking procedure is implemented only using the 3 training folds used by the SVM in section 4.3, an average high ACC of 96.4% is obtained using an average of 529 features on the N-termini dataset. On the C-termini dataset, an average high ACC of 96.3% is found using an average of 647 features. Very close agreement for feature ranks can be observed at either termini whether F-select is applied to the full-size datasets or the 3-fold averages. Both rankings are reported in Appendix Tables B.1 and B.2.

Table 4.3 lists the top 15 N-termini features for the full-size N-termini dataset. Column 2 shows the residue position of a mature peptide corresponding to a reported top feature. F-scores are shown in column 3. Column 4 shows the AAIndex [8] entry corresponding to the physicochemical property represented by each feature. Column 5 provides a brief

explanation of each AAIndex entry, using source descriptions from [8]. An extended list of top-ranked features can be found in Appendix Table B.1.

Table 4.3: Top-15 N-termini Features Using F-Select

Rank	Residue Position	F-Score	AAIndex Entry	AAIndex Entry Description
1	2	0.245	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
2	3	0.240	RICJ880107	Relative preference value at N4 (Richardson-Richardson, 1988)
3	3	0.198	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)
4	3	0.187	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
5	3	0.186	GRAR740103	Volume (Grantham, 1974)
6	3	0.186	SNEP660103	Principal component III (Sneath, 1966)
7	3	0.165	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
8	3	0.165	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
9	3	0.160	BIGC670101	Residue volume (Bigelow, 1967)
10	3	0.152	RADA880106	Accessible surface area (Radzicka-Wolfenden, 1988)
11	3	0.152	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)
12	3	0.150	MCMT640101	Refractivity (McMeekin et al., 1964), Cited by Jones (1975)
13	15	0.145	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
14	15	0.145	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)
15	12	0.144	FAUJ880111	Positive charge (Fauchere et al., 1988)

Inspection of the full N-termini dataset of ranked features reveals that those potentially important for cleavage, rather than activity, do not start until rank 122. Accordingly, the top 15 features shown in Table 4.3 (or the extended list in Appendix Table B.1), reassuringly, do not include any of the cleavage-related features described above.

To provide some more information about the correlated features which were removed due to redundancy (as described in section 3.1.3), listed alongside each entry are additional ones that share 100% correlation to those reported. For instance, as rank 7 feature ARGP820101 in the C-termini dataset shares 100% correlation with JOND750101, the latter is co-listed despite being absent from the data used by the SVM. Viewing other correlated features is detailed in section 4.6.

Table 4.4 reports the top 15 C-termini features obtained through the feature reduction technique based on F-scores for the full-size C-termini dataset. Column 2 shows the residue position of a mature peptide corresponding to a reported top feature. Negative positions count backwards from the C-terminus (i.e. -1 refers to the final C-terminal residue). F-scores are shown in column 3. Column 4 shows the AAIndex [8] entry corresponding to the physicochemical property represented in each feature. Column 5 provides a brief explanation of each AAIndex entry, using source descriptions from [8]. An extended list of C-termini ranked features is available in Appendix Table B.2.

Table 4.4: Top-15 C-termini Features Using F-select

Rank	**Residue Position**	**F-Score**	**AAIndex Entry**	**AAIndex Entry Description**
1	-3	0.332	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
2	-3	0.314	GEOR030101	Linker propensity from all dataset (George-Heringa, 2003)
3	-3	0.306	FINA910102	Helix initiation parameter at posision i,i+1,i+2 (Finkelstein et al., 1991)
4	-3	0.253	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
5	-3	0.252	AURR980119	Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)
6	-9	0.250	VASM830103	Relative population of conformational state E (Vasquez et al., 1983)
7	-9	0.248	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
8	-3	0.243	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
9	-9	0.242	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
10	-3	0.223	LAWE840101	Transfer free energy, CHP/water (Lawson et al., 1984)
11	-17	0.220	KLEP840101	Net charge (Klein et al., 1984)
12	-17	0.216	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
13	-9	0.215	SNEP660103	Principal component III (Sneath, 1966)
14	-9	0.214	ARGP820101 JOND750101	Hydrophobicity index (Argos et al., 1982) Hydrophobicity (Jones, 1975)
15	-9	0.213	TAKK010101	Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)

4.6 Results for Remaining Ranked Features

An extended list of the top 100 ranked features for the N- and C-termini datasets are available in Appendix Table B.1 and B.2 respectively. Physicochemical profile graphs for these same top-ranked features are available in Appendix C (ordered by AAIndex ID).

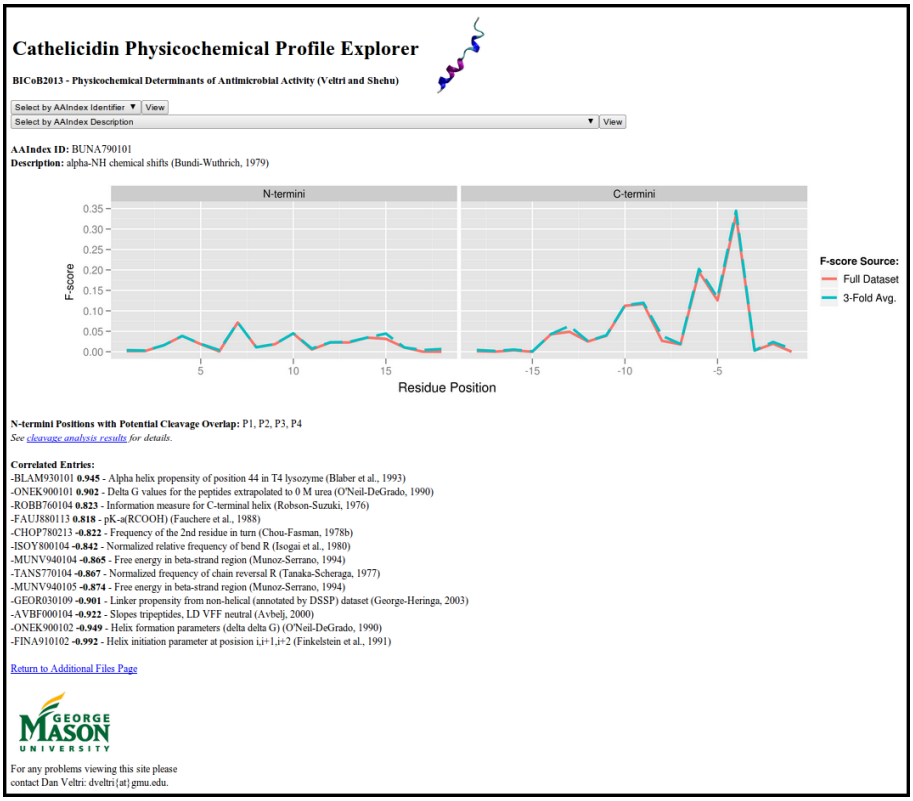


Figure 4.2: Cathelicidin Physicochemical Profile Explorer Example Entry

Furthermore, the "Cathelicidin Physicochemical Profile Explorer" (shown in Figure 4.2) is an online web tool created using Perl and CGI to provide the reader with an easy way to explore all considered features ranked in this study. Profiles also list the correlated AAIndex entries which were removed as described in section 3.1.3. Cathelicidin Physicochemical Profile Explorer is freely available online at:

http://binf.gmu.edu/dveltri/cgi-bin/cath_explorer.cgi.

Chapter 5: Discussion

5.1 Approach Summary

This thesis presents a supervised learning method for AMP recognition and elucidation of activity-related physicochemical features on a per-position basis. It also contributes a carefully selected decoy dataset and statistical approach to identify features confounded by enzymatic cleavage for the cathelicidin family of AMPs. This reduces the chance of features exploiting trivial differences based on structure or cleavage. Through a combination of SVM and F-scores, the list of position-based features is reduced to a few with high discriminatory power. A comparison with another top method in the field confirms that the presented features are also relevant for general AMP recognition.

While it is encouraging to see that many of the top features reported by this analysis overlap with findings from other computational and wet lab studies [11, 12, 16, 19, 21], verification by experiment is still necessary to confirm relevant biological activity. A number of novel features have been introduced and the complete set of ranked and correlated features are also available online with the Cathelicidin Physicochemical Profile Explorer. Freely available at: `http://binf.gmu.edu/dveltri/cgi-bin/cath_explorer.cgi`. Feature profiles should facilitate the aided modification or design of novel AMPs. Biologists may be interested in further analyzing the feature profiles generated by this method. By preserving the list of top features when making mutations to known AMPs, the set of relevant amino acids can be reduced in the context of design. Future wet laboratory studies can provide more data through which computational accuracy can be increased and the feature space reduced further. Ultimately, target bacteria specificity and activity will play

a large role in which features a domain-expert may prioritize for AMP-design. However, a few features which stood out from this analysis are mentioned below.

5.2 Cathelicidin Physicochemical Features of Interest

A number of the reported top features for both termini have already been found biologically important for activity in the literature [19]. Notably, both hydrophobicity and charge are known important for attraction toward bacterial membranes [19]. These same features have also been used successfully for AMP recognition [11, 12].

Through averaging the F-scores for all 299 features at each residue position separately, one can see how some positions appear to play a greater role for cathelicidin recognition than others. This is illustrated for both the N- and C-termini in Figure 5.1. Spikes in average F-score values generally appear after every 4th residue position. This is unsurprising considering most cathelicidins can form helical structures and these are known to follow a typical $[O(i)$ to $N(i+4)]$ hydrogen-bonding pattern [76]. As the decoy dataset was constructed using HeliQuest specifically to find overlapping helical properties (see [66] for details), it was surprising to see quite so many helix-related top features reported in Table 4.3 and 4.4. However, it should be noted that these features should capture differences between AMP and non-AMPs helical patterns. For example, as a number of competing hydrophobicity scales exist [77], some may be more relevant to AMP-activity than others.

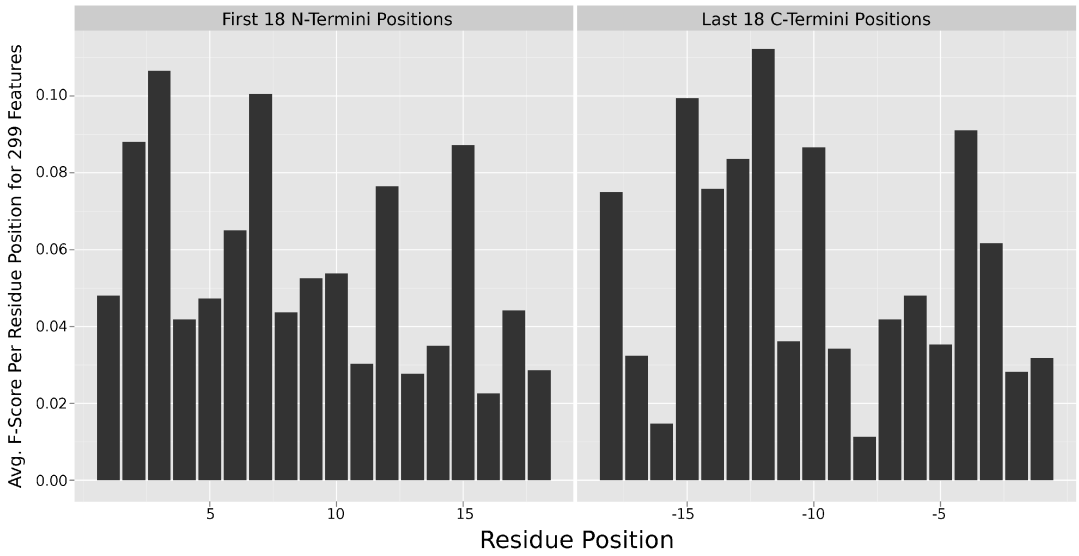


Figure 5.1: Average F-score Over 299 Aggregate Features Per Residue Position.

Figure 5.2 shows a C-termini profile for a novel feature, BUNA790101, ranked $1st$ at position $i = -4$ and $19th$ at position $i = -6$ in Table 4.4. The AAIndex describes this feature as "alpha-NH chemical shifts (Bundi-Wuthrich, 1979)." Chemical shifts describe the electronic environment surrounding a peptide [78] and this may have further relevance to membrane interactions. Table 5.1 outlines some correlated AAIndex entries.

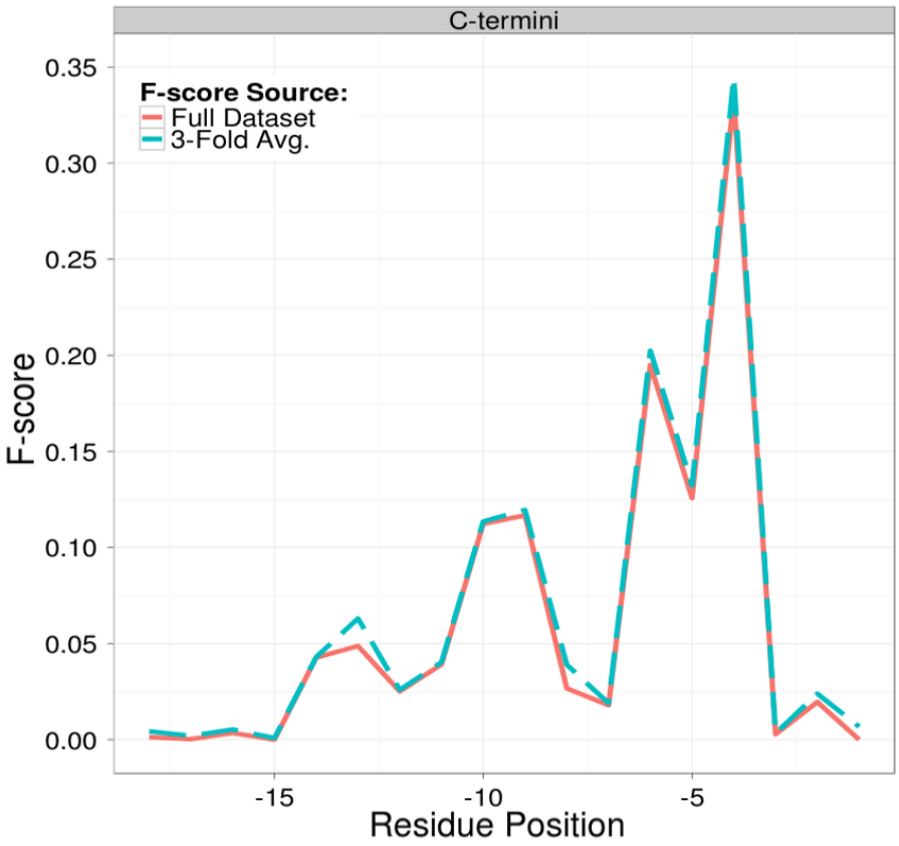


Figure 5.2: Profile for Top-Ranked Feature BUNA790101.

5.2.1 Using Correlated Results to Direct Literature Searches

One way related features can direct further literature searches is through using the correlated entries provided by the AAIndex [8] as shown in Table 5.1. Feature BUNA790101, "alpha-NH chemical shifts (Bundi-Wuthrich, 1979)," is ranked as a top C-termini feature at two residue positions in Table 3. Column 1 contains all of its positively-correlated features, as defined in [8], followed by the correlation coefficient values in column 2. The descriptions (top) and article titles (bottom) in each row of column 3 all happen to contain key words/phrases relevant to AMPs (bold added for emphasis). The articles linked to

these AAIndex entries may be of possible direct or indirect interest to the AMP research community.

Table 5.1: Using Correlated Features for Directed Literature Searches

AAIndex ID	**Corr.**	**AAIndex Description / Article Title**
BLAM930101	0.945	**Alpha helix prop.** of posit. 44 in T4 lysozyme (Blaber et al., 1993) **Structural basis of amino acid alpha helix propensity**
ONEK900101	0.902	Delta G values for the pep. extrapolated to 0 M urea (O'Neil-DeGrado, 1990) A thermodynamic scale for the **helix-forming tendencies** ...
ROBB760104	0.823	Information measure for **C-terminal helix** (Robson-Suzuki, 1976) Conformational properties of amino acid residues in globular proteins
FAUJ880113	0.818	pK-a(RCOOH) (Fauchere et al., 1988) ... **side chain parameters for corr. studies in bio. and pharma.**

5.3 Generalizing Position-Based Features Proves Useful

To show that the AAIndex-based features employed are relevant beyond cathelicidins, a generalized AMP dataset from [12] was considered. AAIndex features were averaging across all amino acids to create fixed-length vectors as described in section 4.2. Accuracies in the 80.0 to 85.9% range have been obtained. However, when the 8 general features used in [11] and [12] are incorporated into the feature space, accuracy increases to 93%. Of these, *in vitro* peptide aggregation ranks as the top feature for activity and is in agreement with previous findings [11, 12]. AAIndex features make up the remainder of the top ten features in this set, including WEBA780101 ("RF value in high salt chromatography") which may have relevance to the salt sensitivity of helical AMPs. Experimental studies have shown that peptides particularly sensitive to salt concentrations may assume helical structures too early and reduce performance in disrupting bacterial membranes [19]. This feature is ranked 54th in the N-termini cathelicidin dataset and demonstrates how some of the features reported here may be relevant for AMPs aside from cathelicidins. There

may be a benefit in mixing both generalized and position-based features.

5.4 Future Directions

This thesis has presented a supervised learning method for elucidating activity-related physicochemical features at the local amino acid level. Work is underway exploring how the generalized features from [11] may be further combined with the position-specific ones described here. While generalized features have the benefit of working on a variable-length dataset, they lack the ability to capture more subtle sequence-based patterns. In cases of AMP families with conserved regions, it may be possible to apply the position-based features presented here at those locations. This could then be combined with the more general features to cover regions of low conservation and improve overall recognition. It is hoped a shift will occur towards applying features which are more useful for identifying motifs related to antimicrobial activity. These will be of more use for directing AMP modification and design of novel AMPs.

Finally, other ongoing work includes applying multi-variate analysis on the features presented in this thesis. The F-select method used for feature ranking only considers features individually, however, some my share synergistic or antagonistic relationships. Statistical approaches such as least absolute shrinkage and selection operator (LASSO) [79] may be better able to capture some of the interplay between important physicochemical features and how they relate to antimicrobial activity.

Appendix A: Dataset Members

Table A.1: N-Termini Dataset Members

Sample Class	Database	Identifier	Start Pos.	End Pos.	Organism
Positive (Cath.)	UniProt	P49929	1	18	*Ovis aries*
Positive (Cath.)	UniProt	P51437	1	18	*Mus musculus*
Positive (Cath.)	UniProt	P49913	1	18	*Homo sapiens*
Positive (Cath.)	UniProt	P54228	1	18	*Bos taurus*
Positive (Cath.)	UniProt	P54229	1	18	*Bos taurus*
Positive (Cath.)	UniProt	P49930	1	18	*Sus scrofa*
Positive (Cath.)	UniProt	P49931	1	18	*Sus scrofa*
Positive (Cath.)	UniProt	P49932	1	18	*Sus scrofa*
Positive (Cath.)	UniProt	P25230	1	18	*Oryctolagus cuniculus*
Positive (Cath.)	UniProt	Q2IAL7	1	18	*Gallus gallus*
Positive (Cath.)	UniProt	Q6QLQ5	1	18	*Gallus gallus*
Positive (Cath.)	UniProt	Q2IAL6	1	18	*Gallus gallus*
Positive (Cath.)	UniProt	Q9GLV5	1	18	*Macaca mulatta*
Positive (Cath.)	UniProt	Q91X12	1	18	*Cavia porcellus*
Positive (Cath.)	UniProt	Q6TN20	1	18	*Canis familiaris*
Positive (Cath.)	UniProt	P79360	1	18	*Ovis aries*
Positive (Cath.)	UniProt	O62840	1	18	*Equus caballus*
Positive (Cath.)	UniProt	O62841	1	18	*Equus caballus*
Positive (Cath.)	UniProt	O62842	1	18	*Equus caballus*
Positive (Cath.)	UniProt	Q71MD7	1	18	*Myxine glutinosa*
Positive (Cath.)	UniProt	Q1KLX8	1	18	*Hylobates moloch*
Positive (Cath.)	UniProt	Q1KLX4	1	18	*Trachypithecus obscurus*
Positive (Cath.)	UniProt	Q1KLY4	1	18	*Callithrix jacchus*
Positive (Cath.)	UniProt	B6D434	1	18	*Bungarus fasciatus*
Positive (Cath.)	APD2	01533	1	18	*Equus asinus*
Positive (Cath.)	Lit: [65]	EA-CATH2	1	18	*Equus asinus*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Positive (Cath.)	UniProt	Q1KLY8	1	18	*Ateles fusciceps robustus*
Positive (Cath.)	UniProt	Q1KLY5	1	18	*Cebus capucinus*
Positive (Cath.)	UniProt	Q1KLY6	1	18	*Chlorocebus aethiops*
Positive (Cath.)	UniProt	Q1KLY0	1	18	*Nomascus concolor*
Positive (Cath.)	UniProt	Q1KLX0	1	18	*Saguinus oedipus*
Positive (Cath.)	UniProt	Q5F378	1	18	*Gallus gallus*
Positive (Cath.)	UniProt	P56425	1	18	*Bos taurus*
Positive (Cath.)	UniProt	P82017	1	18	*Capra hircus*
Positive (Cath.)	UniProt	Q9XSQ8	1	18	*Capra hircus*
Positive (Cath.)	APD2	01768	1	18	*Felis catus*
Positive (Cath.)	APD2	01799	1	18	*Macropus eugenii*
Positive (Cath.)	APD2	01799	1	18	*Macropus eugenii*
Positive (Cath.)	APD2	01801	1	18	*Ornithorhynchus anatinus*
Positive (Cath.)	APD2	01802	1	18	*Ornithorhynchus anatinus*
Positive (Cath.)	APD2	01898	1	18	*Amolops loloensis*
Positive (Cath.)	APD2	00896	1	18	*Bungarus fasciatus*
Positive (Cath.)	APD2	00897	1	18	*Naja atra*
Positive (Cath.)	APD2	01286	1	18	*Cervus elaphus*
Positive (Cath.)	APD2	00879	1	18	*Ailuropoda melanoleuca*
Negative (Decoy)	UniProt	P46939	1526	1544	*Homo sapiens*
Negative (Decoy)	UniProt	O75170	217	235	*Homo sapiens*
Negative (Decoy)	UniProt	Q02108	420	438	*Homo sapiens*
Negative (Decoy)	UniProt	P04818	93	111	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WWW0	54	72	*Homo sapiens*
Negative (Decoy)	UniProt	Q9HBG6	701	719	*Homo sapiens*
Negative (Decoy)	UniProt	Q9GZY0	6	24	*Homo sapiens*
Negative (Decoy)	UniProt	Q15276	515	533	*Homo sapiens*
Negative (Decoy)	UniProt	Q12955	2515	2533	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q15058	605	623	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TDF6	150	168	*Homo sapiens*
Negative (Decoy)	UniProt	Q13342	381	399	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y2K3	1826	1844	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TDB6	508	526	*Homo sapiens*
Negative (Decoy)	UniProt	Q96JA3	421	439	*Homo sapiens*
Negative (Decoy)	UniProt	Q99707	999	1017	*Homo sapiens*
Negative (Decoy)	UniProt	P30085	43	61	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y4G6	1360	1378	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y613	679	697	*Homo sapiens*
Negative (Decoy)	UniProt	Q9P0U3	83	101	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H8N7	471	489	*Homo sapiens*
Negative (Decoy)	UniProt	Q16543	75	93	*Homo sapiens*
Negative (Decoy)	UniProt	O60941	565	583	*Homo sapiens*
Negative (Decoy)	UniProt	P42680	35	53	*Homo sapiens*
Negative (Decoy)	UniProt	Q96QG7	166	184	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TF76	454	472	*Homo sapiens*
Negative (Decoy)	UniProt	O95714	630	648	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NA72	483	501	*Homo sapiens*
Negative (Decoy)	UniProt	O14964	1	19	*Homo sapiens*
Negative (Decoy)	UniProt	P46821	913	931	*Homo sapiens*
Negative (Decoy)	UniProt	O95235	735	753	*Homo sapiens*
Negative (Decoy)	UniProt	P53370	75	93	*Homo sapiens*
Negative (Decoy)	UniProt	Q86UR1	153	171	*Homo sapiens*
Negative (Decoy)	UniProt	Q86YR5	129	147	*Homo sapiens*
Negative (Decoy)	UniProt	P26045	471	489	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N841	506	524	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q96RG2	880	898	*Homo sapiens*
Negative (Decoy)	UniProt	O43150	127	145	*Homo sapiens*
Negative (Decoy)	UniProt	Q16566	54	72	*Homo sapiens*
Negative (Decoy)	UniProt	Q9BYX2	513	531	*Homo sapiens*
Negative (Decoy)	UniProt	P49619	320	338	*Homo sapiens*
Negative (Decoy)	UniProt	Q5H9J7	52	70	*Homo sapiens*
Negative (Decoy)	UniProt	Q96PM5	1	19	*Homo sapiens*
Negative (Decoy)	UniProt	O14979	338	356	*Homo sapiens*
Negative (Decoy)	UniProt	P07384	587	605	*Homo sapiens*
Negative (Decoy)	UniProt	O15061	961	979	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TBZ2	669	687	*Homo sapiens*
Negative (Decoy)	UniProt	Q14008	245	263	*Homo sapiens*
Negative (Decoy)	UniProt	Q9HC62	445	463	*Homo sapiens*
Negative (Decoy)	UniProt	O95487	350	368	*Homo sapiens*
Negative (Decoy)	UniProt	Q8IVI9	408	426	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H3S7	1170	1188	*Homo sapiens*
Negative (Decoy)	UniProt	Q96AX9	84	102	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NP74	200	218	*Homo sapiens*
Negative (Decoy)	UniProt	Q13459	1068	1086	*Homo sapiens*
Negative (Decoy)	UniProt	Q969Z4	114	132	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NSI6	1793	1811	*Homo sapiens*
Negative (Decoy)	UniProt	Q96N21	180	198	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y5S9	48	66	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H9T3	20	38	*Homo sapiens*
Negative (Decoy)	UniProt	Q9ULL8	1240	1258	*Homo sapiens*
Negative (Decoy)	UniProt	P50991	52	70	*Homo sapiens*
Negative (Decoy)	UniProt	P22392	111	129	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	Database	Identifier	Start Pos.	End Pos.	Organism
Negative (Decoy)	UniProt	P49815	1114	1132	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UPY3	1885	1903	*Homo sapiens*
Negative (Decoy)	UniProt	O95835	1054	1072	*Homo sapiens*
Negative (Decoy)	UniProt	P67936	20	38	*Homo sapiens*
Negative (Decoy)	UniProt	P54687	209	227	*Homo sapiens*
Negative (Decoy)	UniProt	P42229	61	79	*Homo sapiens*
Negative (Decoy)	UniProt	Q96BW1	186	204	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y3S1	1144	1162	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y216	376	394	*Homo sapiens*
Negative (Decoy)	UniProt	A7KAX9	1919	1937	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WUY3	1342	1360	*Homo sapiens*
Negative (Decoy)	UniProt	P23381	364	382	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WUF5	290	308	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TE68	601	619	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H814	168	186	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NEM0	52	70	*Homo sapiens*
Negative (Decoy)	UniProt	Q96CV9	180	198	*Homo sapiens*
Negative (Decoy)	UniProt	Q13310	481	499	*Homo sapiens*
Negative (Decoy)	UniProt	Q9BXL7	69	87	*Homo sapiens*
Negative (Decoy)	UniProt	P08235	135	153	*Homo sapiens*
Negative (Decoy)	UniProt	Q6ZN04	3	21	*Homo sapiens*
Negative (Decoy)	UniProt	P27708	1397	1415	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TD86	33	51	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NHV4	494	512	*Homo sapiens*
Negative (Decoy)	UniProt	Q01484	3107	3125	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NYF8	68	86	*Homo sapiens*
Negative (Decoy)	UniProt	O14818	161	179	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q15075	8	26	*Homo sapiens*
Negative (Decoy)	UniProt	Q96A56	139	157	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NQT8	836	854	*Homo sapiens*
Negative (Decoy)	UniProt	Q96SN8	138	156	*Homo sapiens*
Negative (Decoy)	UniProt	P40855	103	121	*Homo sapiens*
Negative (Decoy)	UniProt	Q06124	182	200	*Homo sapiens*
Negative (Decoy)	UniProt	O94830	64	82	*Homo sapiens*
Negative (Decoy)	UniProt	Q6P1N0	910	928	*Homo sapiens*
Negative (Decoy)	UniProt	P22626	228	246	*Homo sapiens*
Negative (Decoy)	UniProt	P40818	91	109	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WV24	98	116	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N556	279	297	*Homo sapiens*
Negative (Decoy)	UniProt	Q86WV1	87	105	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N0Z6	51	69	*Homo sapiens*
Negative (Decoy)	UniProt	P54136	463	481	*Homo sapiens*
Negative (Decoy)	UniProt	Q7Z5J4	762	780	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NV58	652	670	*Homo sapiens*
Negative (Decoy)	UniProt	O95429	424	442	*Homo sapiens*
Negative (Decoy)	UniProt	Q16526	551	569	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NEX9	147	165	*Homo sapiens*
Negative (Decoy)	UniProt	Q16644	119	137	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TE73	4154	4172	*Homo sapiens*
Negative (Decoy)	UniProt	O75626	64	82	*Homo sapiens*
Negative (Decoy)	UniProt	P20929	1087	1105	*Homo sapiens*
Negative (Decoy)	UniProt	Q12873	832	850	*Homo sapiens*
Negative (Decoy)	UniProt	P23297	51	69	*Homo sapiens*
Negative (Decoy)	UniProt	P16452	421	439	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	P61764	10	28	*Homo sapiens*
Negative (Decoy)	UniProt	Q38SD2	25	43	*Homo sapiens*
Negative (Decoy)	UniProt	Q9P2R3	757	775	*Homo sapiens*
Negative (Decoy)	UniProt	Q63HN8	525	543	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UPQ9	774	792	*Homo sapiens*
Negative (Decoy)	UniProt	P09917	426	444	*Homo sapiens*
Negative (Decoy)	UniProt	P28908	570	588	*Homo sapiens*
Negative (Decoy)	UniProt	Q04609	397	415	*Homo sapiens*
Negative (Decoy)	UniProt	P55265	1190	1208	*Homo sapiens*
Negative (Decoy)	UniProt	O00750	242	260	*Homo sapiens*
Negative (Decoy)	UniProt	Q5TZA2	1250	1268	*Homo sapiens*
Negative (Decoy)	UniProt	O00519	257	275	*Homo sapiens*
Negative (Decoy)	UniProt	P61158	42	60	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NC51	177	195	*Homo sapiens*
Negative (Decoy)	UniProt	Q08378	783	801	*Homo sapiens*
Negative (Decoy)	UniProt	Q5TAX3	1396	1414	*Homo sapiens*
Negative (Decoy)	UniProt	Q9C0C2	1004	1022	*Homo sapiens*
Negative (Decoy)	UniProt	P21127	754	772	*Homo sapiens*
Negative (Decoy)	UniProt	O60237	824	842	*Homo sapiens*
Negative (Decoy)	UniProt	Q86TI2	684	702	*Homo sapiens*
Negative (Decoy)	UniProt	P60323	146	164	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UJ68	141	159	*Homo sapiens*
Negative (Decoy)	UniProt	P35609	52	70	*Homo sapiens*
Negative (Decoy)	UniProt	O75936	227	245	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y2W7	90	108	*Homo sapiens*
Negative (Decoy)	UniProt	Q6VN20	514	532	*Homo sapiens*
Negative (Decoy)	UniProt	P50552	252	270	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	O00571	95	113	*Homo sapiens*
Negative (Decoy)	UniProt	Q8NEV4	104	122	*Homo sapiens*
Negative (Decoy)	UniProt	O15020	1855	1873	*Homo sapiens*
Negative (Decoy)	UniProt	O00425	498	516	*Homo sapiens*
Negative (Decoy)	UniProt	Q4KWH8	144	162	*Homo sapiens*
Negative (Decoy)	UniProt	Q5JSH3	432	450	*Homo sapiens*
Negative (Decoy)	UniProt	P05089	74	92	*Homo sapiens*
Negative (Decoy)	UniProt	P00558	260	278	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WXH0	287	305	*Homo sapiens*
Negative (Decoy)	UniProt	P61201	181	199	*Homo sapiens*
Negative (Decoy)	UniProt	Q6IMN6	861	879	*Homo sapiens*
Negative (Decoy)	UniProt	P22307	166	184	*Homo sapiens*
Negative (Decoy)	UniProt	Q76N32	611	629	*Homo sapiens*
Negative (Decoy)	UniProt	P29536	95	113	*Homo sapiens*
Negative (Decoy)	UniProt	Q53EL6	71	89	*Homo sapiens*
Negative (Decoy)	UniProt	Q14315	2310	2328	*Homo sapiens*
Negative (Decoy)	UniProt	P38935	772	790	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UQL6	473	491	*Homo sapiens*
Negative (Decoy)	UniProt	A6NIX2	281	299	*Homo sapiens*
Negative (Decoy)	UniProt	P07954	464	482	*Homo sapiens*
Negative (Decoy)	UniProt	Q96DT5	3843	3861	*Homo sapiens*
Negative (Decoy)	UniProt	Q9BZF9	552	570	*Homo sapiens*
Negative (Decoy)	UniProt	Q96ED9	305	323	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y6N9	423	441	*Homo sapiens*
Negative (Decoy)	UniProt	Q13813	698	716	*Homo sapiens*
Negative (Decoy)	UniProt	Q2V2M9	578	596	*Homo sapiens*
Negative (Decoy)	UniProt	Q96RS6	533	551	*Homo sapiens*

Continued on next page...

Table A.1: N-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	P27816	1062	1080	*Homo sapiens*
Negative (Decoy)	UniProt	Q92974	816	834	*Homo sapiens*
Negative (Decoy)	UniProt	P14635	40	58	*Homo sapiens*
Negative (Decoy)	UniProt	Q92837	258	276	*Homo sapiens*
Negative (Decoy)	UniProt	O60271	261	279	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y6K8	462	480	*Homo sapiens*
Negative (Decoy)	UniProt	Q504Q3	1158	1176	*Homo sapiens*
Negative (Decoy)	UniProt	Q9HCE6	1245	1263	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TCY9	843	861	*Homo sapiens*

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Positive (Cath.)	UniProt	P49929	12	29	*Ovis aries*
Positive (Cath.)	UniProt	P51437	17	34	*Mus musculus*
Positive (Cath.)	UniProt	P49913	21	38	*Homo sapiens*
Positive (Cath.)	UniProt	P54228	10	27	*Bos taurus*
Positive (Cath.)	UniProt	P54229	11	28	*Bos taurus*
Positive (Cath.)	UniProt	P49930	6	23	*Sus scrofa*
Positive (Cath.)	UniProt	P49931	20	37	*Sus scrofa*
Positive (Cath.)	UniProt	P49932	20	37	*Sus scrofa*
Positive (Cath.)	UniProt	P25230	20	37	*Oryctolagus cuniculus*
Positive (Cath.)	UniProt	Q2IAL7	15	32	*Gallus gallus*
Positive (Cath.)	UniProt	Q6QLQ5	9	26	*Gallus gallus*
Positive (Cath.)	UniProt	Q2IAL6	12	29	*Gallus gallus*
Positive (Cath.)	UniProt	Q9GLV5	21	38	*Macaca mulatta*
Positive (Cath.)	UniProt	Q91X12	26	43	*Cavia porcellus*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Positive (Cath.)	UniProt	Q6TN20	21	38	*Canis familiaris*
Positive (Cath.)	UniProt	P79360	16	33	*Ovis aries*
Positive (Cath.)	UniProt	O62840	9	26	*Equus caballus*
Positive (Cath.)	UniProt	O62841	10	27	*Equus caballus*
Positive (Cath.)	UniProt	O62842	23	40	*Equus caballus*
Positive (Cath.)	UniProt	Q71MD7	21	38	*Myxine glutinosa*
Positive (Cath.)	UniProt	Q1KLX8	21	38	*Hylobates moloch*
Positive (Cath.)	UniProt	Q1KLX4	21	38	*Trachypithecus obscurus*
Positive (Cath.)	UniProt	Q1KLY4	21	38	*Callithrix jacchus*
Positive (Cath.)	UniProt	B6D434	13	30	*Bungarus fasciatus*
Positive (Cath.)	APD2	01533	8	25	*Equus asinus*
Positive (Cath.)	Lit: [65]	EA-CATH2	9	26	*Equus asinus*
Positive (Cath.)	UniProt	Q1KLY8	21	38	*Ateles fusciceps robustus*
Positive (Cath.)	UniProt	Q1KLY5	21	38	*Cebus capucinus*
Positive (Cath.)	UniProt	Q1KLY6	21	38	*Chlorocebus aethiops*
Positive (Cath.)	UniProt	Q1KLY0	21	38	*Nomascus concolor*
Positive (Cath.)	UniProt	Q1KLX0	21	38	*Saguinus oedipus*
Positive (Cath.)	UniProt	Q5F378	23	40	*Gallus gallus*
Positive (Cath.)	UniProt	P56425	18	35	*Bos taurus*
Positive (Cath.)	UniProt	P82017	17	34	*Capra hircus*
Positive (Cath.)	UniProt	Q9XSQ8	11	28	*Capra hircus*
Positive (Cath.)	APD2	01768	20	37	*Felis catus*
Positive (Cath.)	APD2	01799	19	36	*Macropus eugenii*
Positive (Cath.)	APD2	01799	19	36	*Macropus eugenii*
Positive (Cath.)	APD2	01801	17	34	*Ornithorhynchus anatinus*
Positive (Cath.)	APD2	01802	5	22	*Ornithorhynchus anatinus*
Positive (Cath.)	APD2	01898	31	48	*Amolops loloensis*
Positive (Cath.)	APD2	00896	17	34	*Bungarus fasciatus*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Positive (Cath.)	APD2	00897	17	34	*Naja atra*
Positive (Cath.)	APD2	01286	17	34	*Cervus elaphus*
Positive (Cath.)	APD2	00879	21	38	*Ailuropoda melanoleuca*
Negative (Decoy)	UniProt	Q6V1X1	179	197	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UH77	444	462	*Homo sapiens*
Negative (Decoy)	UniProt	Q03426	325	343	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y6K8	144	162	*Homo sapiens*
Negative (Decoy)	UniProt	P21695	98	116	*Homo sapiens*
Negative (Decoy)	UniProt	Q92837	259	277	*Homo sapiens*
Negative (Decoy)	UniProt	P16050	416	434	*Homo sapiens*
Negative (Decoy)	UniProt	Q07812	20	38	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NP74	202	220	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H7D7	562	580	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H0N0	56	74	*Homo sapiens*
Negative (Decoy)	UniProt	Q6ZSZ5	690	708	*Homo sapiens*
Negative (Decoy)	UniProt	P04406	99	117	*Homo sapiens*
Negative (Decoy)	UniProt	P37840	60	78	*Homo sapiens*
Negative (Decoy)	UniProt	Q92945	647	665	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y573	563	581	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UJC3	683	701	*Homo sapiens*
Negative (Decoy)	UniProt	Q8IWV7	93	111	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y240	290	308	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y239	318	336	*Homo sapiens*
Negative (Decoy)	UniProt	P06746	63	81	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y4J8	543	561	*Homo sapiens*
Negative (Decoy)	UniProt	P49796	434	452	*Homo sapiens*
Negative (Decoy)	UniProt	P17544	270	288	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q8WXD5	57	75	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N4N8	289	307	*Homo sapiens*
Negative (Decoy)	UniProt	Q6EMB2	1109	1127	*Homo sapiens*
Negative (Decoy)	UniProt	P54136	234	252	*Homo sapiens*
Negative (Decoy)	UniProt	P36871	491	509	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TD20	96	114	*Homo sapiens*
Negative (Decoy)	UniProt	O75874	278	296	*Homo sapiens*
Negative (Decoy)	UniProt	P13798	31	49	*Homo sapiens*
Negative (Decoy)	UniProt	Q13485	73	91	*Homo sapiens*
Negative (Decoy)	UniProt	Q9C0D2	2155	2173	*Homo sapiens*
Negative (Decoy)	UniProt	P53396	745	763	*Homo sapiens*
Negative (Decoy)	UniProt	Q6DKK2	153	171	*Homo sapiens*
Negative (Decoy)	UniProt	Q3MJ16	364	382	*Homo sapiens*
Negative (Decoy)	UniProt	O95487	123	141	*Homo sapiens*
Negative (Decoy)	UniProt	O00429	362	380	*Homo sapiens*
Negative (Decoy)	UniProt	A8MVW5	169	187	*Homo sapiens*
Negative (Decoy)	UniProt	P49703	172	190	*Homo sapiens*
Negative (Decoy)	UniProt	Q92625	692	710	*Homo sapiens*
Negative (Decoy)	UniProt	Q68DD2	363	381	*Homo sapiens*
Negative (Decoy)	UniProt	Q12815	434	452	*Homo sapiens*
Negative (Decoy)	UniProt	Q13085	1870	1888	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N1E6	189	207	*Homo sapiens*
Negative (Decoy)	UniProt	Q8IWT3	2168	2186	*Homo sapiens*
Negative (Decoy)	UniProt	Q99966	46	64	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N4C8	673	691	*Homo sapiens*
Negative (Decoy)	UniProt	Q8IZL8	447	465	*Homo sapiens*
Negative (Decoy)	UniProt	O95376	33	51	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q9NX74	115	133	*Homo sapiens*
Negative (Decoy)	UniProt	O60331	351	369	*Homo sapiens*
Negative (Decoy)	UniProt	Q15102	107	125	*Homo sapiens*
Negative (Decoy)	UniProt	O95848	133	151	*Homo sapiens*
Negative (Decoy)	UniProt	Q9C000	78	96	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H4B7	8	26	*Homo sapiens*
Negative (Decoy)	UniProt	Q86XP0	324	342	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NXR7	278	296	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H4H8	233	251	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NRR3	42	60	*Homo sapiens*
Negative (Decoy)	UniProt	Q00534	5	23	*Homo sapiens*
Negative (Decoy)	UniProt	P00326	66	84	*Homo sapiens*
Negative (Decoy)	UniProt	O75936	212	230	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y6F1	92	110	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NR20	103	121	*Homo sapiens*
Negative (Decoy)	UniProt	P35125	525	543	*Homo sapiens*
Negative (Decoy)	UniProt	Q14517	4097	4115	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UBF8	706	724	*Homo sapiens*
Negative (Decoy)	UniProt	O43312	8	26	*Homo sapiens*
Negative (Decoy)	UniProt	O60259	217	235	*Homo sapiens*
Negative (Decoy)	UniProt	P53355	1269	1287	*Homo sapiens*
Negative (Decoy)	UniProt	Q00653	764	782	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H9F9	212	230	*Homo sapiens*
Negative (Decoy)	UniProt	Q5VST9	4561	4579	*Homo sapiens*
Negative (Decoy)	UniProt	O75106	501	519	*Homo sapiens*
Negative (Decoy)	UniProt	Q7LDG7	271	289	*Homo sapiens*
Negative (Decoy)	UniProt	O60308	589	607	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q6P3W7	879	897	*Homo sapiens*
Negative (Decoy)	UniProt	Q2PPJ7	1691	1709	*Homo sapiens*
Negative (Decoy)	UniProt	Q69YQ0	55	73	*Homo sapiens*
Negative (Decoy)	UniProt	P30085	12	30	*Homo sapiens*
Negative (Decoy)	UniProt	Q9Y2Z0	109	127	*Homo sapiens*
Negative (Decoy)	UniProt	P07384	210	228	*Homo sapiens*
Negative (Decoy)	UniProt	O15523	301	319	*Homo sapiens*
Negative (Decoy)	UniProt	P62826	61	79	*Homo sapiens*
Negative (Decoy)	UniProt	Q02108	151	169	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NZJ4	110	128	*Homo sapiens*
Negative (Decoy)	UniProt	O15360	967	985	*Homo sapiens*
Negative (Decoy)	UniProt	P11277	1135	1153	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UKA4	1676	1694	*Homo sapiens*
Negative (Decoy)	UniProt	Q9HC62	446	464	*Homo sapiens*
Negative (Decoy)	UniProt	P07954	271	289	*Homo sapiens*
Negative (Decoy)	UniProt	P50995	51	69	*Homo sapiens*
Negative (Decoy)	UniProt	P05787	416	434	*Homo sapiens*
Negative (Decoy)	UniProt	Q86XP3	363	381	*Homo sapiens*
Negative (Decoy)	UniProt	Q14145	323	341	*Homo sapiens*
Negative (Decoy)	UniProt	Q9H467	70	88	*Homo sapiens*
Negative (Decoy)	UniProt	O60733	205	223	*Homo sapiens*
Negative (Decoy)	UniProt	O75366	101	119	*Homo sapiens*
Negative (Decoy)	UniProt	O14727	101	119	*Homo sapiens*
Negative (Decoy)	UniProt	Q14244	626	644	*Homo sapiens*
Negative (Decoy)	UniProt	P28066	159	177	*Homo sapiens*
Negative (Decoy)	UniProt	Q16775	125	143	*Homo sapiens*
Negative (Decoy)	UniProt	O94874	150	168	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	P12277	49	67	*Homo sapiens*
Negative (Decoy)	UniProt	Q07020	116	134	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TC71	49	67	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UBP0	469	487	*Homo sapiens*
Negative (Decoy)	UniProt	Q99459	720	738	*Homo sapiens*
Negative (Decoy)	UniProt	P35080	57	75	*Homo sapiens*
Negative (Decoy)	UniProt	P08319	316	334	*Homo sapiens*
Negative (Decoy)	UniProt	Q96P47	287	305	*Homo sapiens*
Negative (Decoy)	UniProt	P51610	672	690	*Homo sapiens*
Negative (Decoy)	UniProt	Q86SG6	511	529	*Homo sapiens*
Negative (Decoy)	UniProt	Q86TI2	683	701	*Homo sapiens*
Negative (Decoy)	UniProt	Q8IUG5	1065	1083	*Homo sapiens*
Negative (Decoy)	UniProt	Q96PN6	392	410	*Homo sapiens*
Negative (Decoy)	UniProt	Q14678	892	910	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UJT0	146	164	*Homo sapiens*
Negative (Decoy)	UniProt	Q4J6C6	529	547	*Homo sapiens*
Negative (Decoy)	UniProt	Q99871	342	360	*Homo sapiens*
Negative (Decoy)	UniProt	Q7L5N1	195	213	*Homo sapiens*
Negative (Decoy)	UniProt	Q96P48	599	617	*Homo sapiens*
Negative (Decoy)	UniProt	P00519	509	527	*Homo sapiens*
Negative (Decoy)	UniProt	Q969H0	416	434	*Homo sapiens*
Negative (Decoy)	UniProt	Q96IF1	257	275	*Homo sapiens*
Negative (Decoy)	UniProt	P06493	4	22	*Homo sapiens*
Negative (Decoy)	UniProt	Q92793	2258	2276	*Homo sapiens*
Negative (Decoy)	UniProt	Q96RG2	878	896	*Homo sapiens*
Negative (Decoy)	UniProt	Q5T5U3	1363	1381	*Homo sapiens*
Negative (Decoy)	UniProt	Q9P2A4	143	161	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	Q8IX03	598	616	*Homo sapiens*
Negative (Decoy)	UniProt	Q14008	82	100	*Homo sapiens*
Negative (Decoy)	UniProt	Q6IWH7	406	424	*Homo sapiens*
Negative (Decoy)	UniProt	P78537	103	121	*Homo sapiens*
Negative (Decoy)	UniProt	P19838	112	130	*Homo sapiens*
Negative (Decoy)	UniProt	Q96F86	13	31	*Homo sapiens*
Negative (Decoy)	UniProt	Q7Z6M1	103	121	*Homo sapiens*
Negative (Decoy)	UniProt	Q16531	363	381	*Homo sapiens*
Negative (Decoy)	UniProt	Q13939	366	384	*Homo sapiens*
Negative (Decoy)	UniProt	O95714	2753	2771	*Homo sapiens*
Negative (Decoy)	UniProt	Q05469	325	343	*Homo sapiens*
Negative (Decoy)	UniProt	Q9NX02	591	609	*Homo sapiens*
Negative (Decoy)	UniProt	P56524	679	697	*Homo sapiens*
Negative (Decoy)	UniProt	Q76N32	381	399	*Homo sapiens*
Negative (Decoy)	UniProt	Q16222	266	284	*Homo sapiens*
Negative (Decoy)	UniProt	O43182	733	751	*Homo sapiens*
Negative (Decoy)	UniProt	Q5TZA2	331	349	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N142	289	307	*Homo sapiens*
Negative (Decoy)	UniProt	Q66K89	299	317	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UKD1	427	445	*Homo sapiens*
Negative (Decoy)	UniProt	Q63HN8	1489	1507	*Homo sapiens*
Negative (Decoy)	UniProt	O15143	332	350	*Homo sapiens*
Negative (Decoy)	UniProt	Q13813	1392	1410	*Homo sapiens*
Negative (Decoy)	UniProt	P50452	64	82	*Homo sapiens*
Negative (Decoy)	UniProt	P48634	2056	2074	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TEQ6	926	944	*Homo sapiens*
Negative (Decoy)	UniProt	Q8TDY4	644	662	*Homo sapiens*

Continued on next page...

Table A.2: C-Termini Dataset Members

Sample Class	**Database**	**Identifier**	**Start Pos.**	**End Pos.**	**Organism**
Negative (Decoy)	UniProt	O15287	42	60	*Homo sapiens*
Negative (Decoy)	UniProt	Q8N1V2	551	569	*Homo sapiens*
Negative (Decoy)	UniProt	P46060	331	349	*Homo sapiens*
Negative (Decoy)	UniProt	P61962	153	171	*Homo sapiens*
Negative (Decoy)	UniProt	Q5TA45	326	344	*Homo sapiens*
Negative (Decoy)	UniProt	P09960	342	360	*Homo sapiens*
Negative (Decoy)	UniProt	Q14032	312	330	*Homo sapiens*
Negative (Decoy)	UniProt	Q9UGK8	284	302	*Homo sapiens*
Negative (Decoy)	UniProt	Q9GZT8	134	152	*Homo sapiens*
Negative (Decoy)	UniProt	Q92817	222	240	*Homo sapiens*
Negative (Decoy)	UniProt	P17655	185	203	*Homo sapiens*
Negative (Decoy)	UniProt	Q8WXH0	6641	6659	*Homo sapiens*
Negative (Decoy)	UniProt	Q14315	2432	2450	*Homo sapiens*
Negative (Decoy)	UniProt	Q6IMN6	229	247	*Homo sapiens*
Negative (Decoy)	UniProt	O15117	548	566	*Homo sapiens*
Negative (Decoy)	UniProt	O00154	72	90	*Homo sapiens*
Negative (Decoy)	UniProt	O60674	550	568	*Homo sapiens*
Negative (Decoy)	UniProt	O60711	126	144	*Homo sapiens*
Negative (Decoy)	UniProt	P49454	1455	1473	*Homo sapiens*
Negative (Decoy)	UniProt	Q6VN20	513	531	*Homo sapiens*
Negative (Decoy)	UniProt	P48730	50	68	*Homo sapiens*

Appendix B: F-score Results

Table A.3: CutDB [80] IDs for Non-AMP Samples in the Cleavage Dataset

3322	3323	3318	3321	3313	3312	3311	3310	3309
17237	17236	17235	17234	18246	19732	18849	18919	18989
18990	19009	19010	19736	19733	19734	19735	19346	19372
19373	19476	19507	19508	19542	19557	19669	19856	20101
20260	20493	20904	20905	20906	20907	20908	20909	20910

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	N-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
1	1	5083	7	7	0.275373	0.159828	0.227867	0.221023	0.216531	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
2	3	4053	3	3	0.143420	0.184060	0.230671	0.186050	0.183979	SNEP660103	Principal component III (Sneath, 1966)
3	2	1254	12	12	0.155153	0.107515	0.304615	0.189094	0.177169	FAUJ880111	Positive charge (Fauchere et al., 1988)
4	5	2185	7	7	0.199261	0.167069	0.164224	0.176851	0.176278	MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
5	4	5127	15	15	0.272192	0.144367	0.129216	0.181925	0.175574	WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce et al. 1995)
6	6	4057	7	7	0.208434	0.120637	0.196029	0.175033	0.171999	SNEP660103	Principal component III (Sneath, 1966)
7	8	33	15	15	0.212815	0.136862	0.165492	0.171723	0.170206	ARGP820101	Hydrophobicity index (Argos et al., 1982)
8	10	1069	7	7	0.174736	0.143431	0.187502	0.168556	0.167213	FASG760103	Optical rotation (Fasman, 1976)
9	12	3232	10	10	0.187745	0.143329	0.162431	0.164502	0.163582	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
10	7	3747	3	3	0.109946	0.154042	0.253304	0.172431	0.162818	RICJ880107	Relative preference value at N4 (Richardson-Richardson, 1988)
11	14	1587	3	3	0.159649	0.130147	0.195971	0.161922	0.160352	GRAR740103	Volume (Grantham, 1974)
12	17	5263	7	7	0.139980	0.161736	0.181244	0.160987	0.160209	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
13	11	5078	2	2	0.225918	0.113947	0.156948	0.165604	0.159869	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
14	15	187	7	7	0.209694	0.121306	0.152455	0.161152	0.158318	BROC820101	Retention coefficient in TFA (Browne et al., 1982)
15	18	327	3	3	0.171547	0.112658	0.195833	0.160013	0.157954	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
16	13	4813	7	7	0.245657	0.118722	0.125109	0.163163	0.156260	TAKK010101	Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
17	9	5079	3	3	0.064576	0.169714	0.280318	0.171536	0.155603	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
18	20	2203	7	7	0.218798	0.115917	0.141536	0.158750	0.154286	MEEJ810101	Retention coefficient in NaClO4 (Meek-Rossetti, 1981)
19	21	1987	7	7	0.170061	0.141421	0.145010	0.152164	0.151844	LAWE840101	Transfer free energy, CHP/water (Lawson et al., 1984)

Continued on next page...

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
20	22	5065	7	7	0.183629	0.144983	0.126189	0.151600	0.149876	WILM950101	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
21	23	129	3	3	0.145695	0.117283	0.180218	0.147732	0.146167	BIGC670101	Residue volume (Bigelow, 1967)
22	24	3746	2	2	0.183882	0.124476	0.132188	0.146849	0.144471	RICJ880107	Relative preference value at N4 (Richardson-Richardson, 1988)
23	16	1020	12	12	0.153682	0.048358	0.281250	0.161097	0.143045	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
24	26	1497	3	3	0.106112	0.144580	0.185239	0.145310	0.142884	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)
25	30	28	10	10	0.126920	0.152978	0.143049	0.140982	0.140723	ARGP820101	Hydrophobicity index (Argos et al., 1982)
26	29	2829	3	3	0.113701	0.112630	0.204431	0.143587	0.140526	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
27	31	1173	3	3	0.144665	0.111215	0.163455	0.139778	0.138851	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al., 1988)
28	19	5304	12	12	0.132303	0.050492	0.294663	0.159153	0.138762	GUYH850105	Apparent partition energies calculated from Chothia index (Guy, 1985)
29	25	1393	7	7	0.231256	0.100457	0.108079	0.146597	0.138738	GARJ730101	Partition coefficient (Garel et al., 1973)
30	32	4443	15	15	0.170513	0.116455	0.131757	0.139575	0.138493	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
31	33	4830	6	6	0.129431	0.103439	0.184868	0.139246	0.135902	FODM020101	Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)
32	34	3657	3	3	0.156683	0.095175	0.157332	0.136397	0.134745	RADA880106	Accessible surface area (Radzicka-Wolfenden, 1988)
33	28	2496	12	12	0.154664	0.051478	0.227590	0.144577	0.132546	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
34	35	2163	3	3	0.129234	0.084858	0.190990	0.135027	0.130612	MCMT640101	Refractivity (McMeekin et al., 1964), Cited by Jones (1975)
35	39	2589	15	15	0.122492	0.151773	0.119092	0.131119	0.130057	OOBM850104	Optimized average non-bonded energy per atom (Oobatake et al., 1985)
36	36	25	7	7	0.199345	0.106674	0.097840	0.134620	0.129675	ARGP820101	Hydrophobicity index (Argos et al., 1982)
37	40	1677	3	3	0.175193	0.092176	0.125332	0.130900	0.128646	HUTJ700102	Absolute entropy (Hutchens, 1970)
38	43	231	15	15	0.139853	0.122888	0.121042	0.127928	0.127621	BULH740102	Apparent partial specific volume (Bull-Breese, 1974)
39	44	4705	7	7	0.144865	0.111593	0.126544	0.127667	0.127312	WIMW960101	Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)
40	45	3201	15	15	0.113013	0.141062	0.125678	0.126584	0.126127	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)

Continued on next page...

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
41	27	516	12	12	0.109096	0.052868	0.271778	0.144581	0.125515	CHOC760102	Residue accessible surface area in folded protein (Chothia, 1976)
42	37	1686	12	12	0.143096	0.062860	0.191593	0.132516	0.125443	HUTJ700102	Absolute entropy (Hutchens, 1970)
43	46	480	12	12	0.127551	0.097877	0.151402	0.125610	0.124285	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton, 1983)
44	38	1251	9	9	0.071012	0.228467	0.097265	0.132248	0.120910	FAUJ880111	Positive charge (Fauchere et al., 1988)
45	48	417	3	3	0.084215	0.142857	0.144880	0.123984	0.120569	CHAM830104	The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)
46	51	2553	15	15	0.142949	0.127599	0.093008	0.121185	0.120093	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)
47	53	4111	7	7	0.145189	0.107278	0.104422	0.118963	0.118005	SWER830101	Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)
48	50	4060	10	10	0.164543	0.074074	0.125950	0.121522	0.117926	SNEP660103	Principal component III (Sneath, 1966)
49	52	3231	9	9	0.142849	0.147032	0.072080	0.120654	0.116167	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
50	49	1938	12	12	0.104863	0.074581	0.185453	0.121632	0.114671	KLEP840101	Net charge (Klein et al., 1984)
51	58	4438	10	10	0.089334	0.106458	0.146259	0.114017	0.112457	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
52	56	1533	3	3	0.078677	0.107972	0.160198	0.115616	0.112379	GEIM800110	Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)
53	63	597	3	3	0.102851	0.122947	0.110989	0.112262	0.112144	CHOP780204	Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)
54	54	4291	7	7	0.173160	0.080497	0.096434	0.116697	0.112111	WEBA780101	RF value in high salt chromatography (Weber-Lacey, 1978)
55	62	941	5	5	0.109472	0.094220	0.133722	0.112471	0.111329	DAYM780201	Relative mutability (Dayhoff et al., 1978b)
56	61	1935	9	9	0.088251	0.152115	0.097727	0.112698	0.110388	KLEP840101	Net charge (Klein et al., 1984)
57	42	1830	12	12	0.091750	0.043750	0.252137	0.129212	0.109865	JANJ790102	Transfer free energy (Janin, 1979)
58	60	5137	7	7	0.168170	0.085402	0.086620	0.113397	0.109332	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)
59	47	3648	12	12	0.116398	0.033251	0.225264	0.124971	0.109057	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
60	41	984	12	12	0.095658	0.034275	0.260999	0.130311	0.108873	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
61	71	704	2	2	0.128003	0.105918	0.093267	0.109063	0.108547	CHOP780210	Normalized frequency of N-terminal non beta region (Chou-Fasman, 1978b)

Continued on next page...

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
62	59	1182	12	12	0.135733	0.052648	0.152211	0.113531	0.108419	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al., 1988)
63	73	3967	7	7	0.132022	0.094658	0.099931	0.108870	0.108245	ROBB790101	Hydration free energy (Robson-Osguthorpe, 1979)
64	55	3975	15	15	0.198212	0.068533	0.080865	0.115870	0.107637	ROBB790101	Hydration free energy (Robson-Osguthorpe, 1979)
65	75	3027	3	3	0.099863	0.110980	0.114510	0.108451	0.107616	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)
66	74	753	15	15	0.118602	0.125414	0.081744	0.108587	0.107393	CHOP780212	Frequency of the 1st residue in turn (Chou-Fasman, 1978b)
67	67	219	3	3	0.079341	0.119463	0.130328	0.109711	0.107330	BULH740102	Apparent partial specific volume (Bull-Breese, 1974)
68	65	5052	12	12	0.102373	0.071466	0.157820	0.110553	0.106779	MITS020101	Amphiphilicity index (Mitaku et al., 2002)
69	81	5204	2	2	0.109882	0.098501	0.111433	0.106605	0.106406	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)
70	77	4821	15	15	0.140460	0.078938	0.103175	0.107524	0.106099	TAKK010101	Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
71	69	5348	2	2	0.153651	0.078407	0.096875	0.109644	0.105744	CORJ870103	PRIFT index (Cornette et al., 1987)
72	70	3823	7	7	0.123442	0.066426	0.138421	0.109430	0.105525	RICJ880111	Relative preference value at C4 (Richardson-Richardson, 1988)
73	72	843	15	15	0.159578	0.078292	0.088965	0.108945	0.105335	CIDH920103	Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)
74	76	3229	7	7	0.152847	0.089839	0.081015	0.107900	0.104685	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
75	66	1131	15	15	0.181328	0.078636	0.071561	0.110508	0.104330	FAUJ880101	Graph shape index (Fauchere et al., 1988)
76	87	3057	15	15	0.133289	0.087848	0.092745	0.104627	0.103535	QIAN880114	Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)
77	84	489	3	3	0.124460	0.066355	0.123901	0.104905	0.103273	CHOC760101	Residue accessible surface area in tripeptide (Chothia, 1976)
78	86	3368	2	2	0.135127	0.095552	0.083269	0.104649	0.103218	RACS770103	Side chain orientational preference (Rackovsky-Scheraga, 1977)
79	93	3237	15	15	0.096909	0.110918	0.102436	0.103421	0.103195	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
80	89	3079	1	1	0.081913	0.122979	0.108493	0.104462	0.102916	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)

Continued on next page...

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
81	85	5095	1	1	0.088775	0.089873	0.135561	0.104736	0.102914	WILM950103	Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)
82	92	4054	4	4	0.113758	0.083275	0.113303	0.103445	0.102845	SNEP660103	Principal component III (Sneath, 1966)
83	82	336	12	12	0.136606	0.053648	0.128068	0.106107	0.102250	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
84	88	2743	7	7	0.146199	0.084207	0.083120	0.104509	0.102085	PLIV810101	Partition coefficient (Pliska et al., 1981)
85	94	3574	10	10	0.127781	0.100315	0.081922	0.103339	0.101994	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
86	97	1315	1	1	0.086429	0.106335	0.116068	0.102944	0.101794	FINA910101	Helix initiation parameter at posision i-1 (Finkelstein et al., 1991)
87	79	681	15	15	0.158840	0.104000	0.058085	0.106975	0.101356	CHOP780208	Normalized frequency of N-terminal beta-sheet (Chou-Fasman, 1978b)
88	99	4055	5	5	0.087694	0.097204	0.122479	0.102459	0.101074	SNEP660103	Principal component III (Sneath, 1966)
89	57	5133	3	3	0.042178	0.094986	0.209505	0.115556	0.100785	BASU050102	Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)
90	98	205	7	7	0.139693	0.093475	0.075175	0.102781	0.100407	BULH740101	Transfer free energy to surface (Bull-Breese, 1974)
91	96	2031	15	15	0.145127	0.064238	0.099887	0.103084	0.100038	LEVM760106	van der Waals parameter R0 (Levitt, 1976)
92	105	2697	15	15	0.114542	0.095765	0.089873	0.100060	0.099602	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al., 1981)
93	95	2181	3	3	0.061807	0.099582	0.148179	0.103189	0.099309	MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
94	90	1359	9	9	0.059515	0.145704	0.106643	0.103954	0.099294	FINA910103	Helix termination parameter at posision j-2 & j-1 & j (Finkelstein et al., 1991)
95	107	2828	2	2	0.110139	0.095526	0.092889	0.099518	0.099086	PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
96	106	3579	15	15	0.086517	0.122242	0.091371	0.100043	0.098723	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
97	103	1373	5	5	0.085242	0.080481	0.135662	0.100462	0.098529	FINA910104	Helix termination parameter at posision j+1 (Finkelstein et al., 1991)
98	83	1362	12	12	0.090503	0.054566	0.172728	0.105932	0.098390	FINA910103	Helix termination parameter at posision j-2 & j-1 & j (Finkelstein et al., 1991)
99	110	4327	7	7	0.062837	0.125547	0.108821	0.099068	0.096533	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)

Continued on next page...

Table B.1: Top 100 N-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
100	78	1146	12	12	0.114049	0.033166	0.174576	0.107264	0.096367	FAUJ880104	STERIMOL length of the side chain (Fauchere et al., 1988)

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
1	1	249	15	-4	0.351667	0.208460	0.473561	0.344563	0.334142	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
2	2	1347	15	-4	0.324448	0.192223	0.439879	0.318850	0.309309	FINA910102	Helix initiation parameter at posision i,i+1,i+2 (Finkelstein et al., 1991)
3	3	5271	15	-4	0.293044	0.169765	0.447078	0.303296	0.291451	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
4	4	5181	15	-4	0.421673	0.183610	0.293864	0.299716	0.286997	GEOR030101	Linker propensity from all dataset (George-Heringa, 2003)
5	5	1825	7	-12	0.205486	0.181574	0.410925	0.265995	0.250426	JANJ790102	Transfer free energy (Janin, 1979)
6	7	1995	15	-4	0.359106	0.138411	0.272591	0.256703	0.245414	LAWE840101	Transfer free energy, CHP/water (Lawson et al., 1984)
7	6	511	7	-12	0.210862	0.143810	0.421328	0.258667	0.237414	CHOC760102	Residue accessible surface area in folded protein (Chothia, 1976)
8	9	3327	15	-4	0.260715	0.138083	0.315589	0.238129	0.231214	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
9	10	4659	15	-4	0.275256	0.160979	0.265595	0.233943	0.230056	AURR980119	Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)
10	11	3579	15	-4	0.229924	0.155617	0.313971	0.233171	0.228200	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
11	8	3227	5	-14	0.192423	0.358675	0.163793	0.238297	0.226416	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
12	13	2473	7	-12	0.204372	0.157088	0.307524	0.222995	0.216360	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi, 1977)
13	12	1969	7	-12	0.147658	0.179169	0.344913	0.223913	0.212392	KRIW790102	Fraction of site occupied by water (Krigbaum-Komoriya, 1979)
14	16	2869	7	-12	0.187548	0.166721	0.277774	0.210681	0.205944	PRAM900101	Hydrophobicity (Prabhakaran, 1990)

Continued on next page...

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
15	15	3231	9	-10	0.122057	0.225346	0.302128	0.216510	0.204307	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
16	14	1249	7	-12	0.161402	0.132719	0.368182	0.220768	0.201480	FAUJ880111	Positive charge (Fauchere et al., 1988)
17	19	5265	9	-10	0.119122	0.271738	0.212978	0.201279	0.196466	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
18	17	1015	7	-12	0.160326	0.164318	0.286648	0.203764	0.195723	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
19	18	247	13	-6	0.287341	0.086491	0.233541	0.202458	0.195035	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
20	21	3570	6	-13	0.203351	0.176672	0.202105	0.194043	0.191921	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
21	22	1761	15	-4	0.207691	0.132011	0.241042	0.193581	0.190929	ISOY800106	Normalized relative frequency of helix end (Isogai et al., 1980)
22	20	3578	14	-5	0.275174	0.236657	0.074120	0.195317	0.185379	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
23	26	1345	13	-6	0.254668	0.079015	0.232995	0.188893	0.182171	FINA910102	Helix initiation parameter at posision i,i+1,i+2 (Finkelstein et al., 1991)
24	24	23	5	-14	0.179910	0.288673	0.100016	0.189533	0.177608	ARGP820101	Hydrophobicity index (Argos et al., 1982)
25	28	3577	13	-6	0.302762	0.095531	0.163335	0.187209	0.177530	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
26	31	508	4	-15	0.239209	0.154530	0.141646	0.178462	0.174142	CHOC760102	Residue accessible surface area in folded protein (Chothia, 1976)
27	29	979	7	-12	0.159036	0.112878	0.285364	0.185759	0.173843	EISD840101	Consensus normalized hydrophobicity scale (Eisenberg, 1984)
28	23	1927	1	-18	0.093727	0.154044	0.329815	0.192529	0.172954	KLEP840101	Net charge (Klein et al., 1984)
29	32	4443	15	-4	0.261322	0.125961	0.147852	0.178378	0.172714	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
30	27	771	15	-4	0.124238	0.095440	0.342145	0.187274	0.171300	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)
31	33	1186	16	-3	0.156406	0.109875	0.267017	0.177766	0.169216	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al., 1988)
32	37	340	16	-3	0.154459	0.123068	0.235719	0.171082	0.166070	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
33	30	5299	7	-12	0.161307	0.093872	0.284682	0.179954	0.165475	GUYH850105	Apparent partition energies calculated from Chothia index (Guy, 1985)
34	35	1690	16	-3	0.145493	0.109438	0.262668	0.172533	0.164302	HUTJ700102	Absolute entropy (Hutchens, 1970)
35	43	3228	6	-13	0.175450	0.143277	0.173498	0.164075	0.163315	QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)

Continued on next page...

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
36	41	4437	9	-10	0.128741	0.168400	0.197860	0.165000	0.162633	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
37	42	27	9	-10	0.136778	0.158192	0.197440	0.164137	0.162571	ARGP820101	Hydrophobicity index (Argos et al., 1982)
38	47	1989	9	-10	0.148779	0.162368	0.175861	0.162336	0.161998	LAWE840101	Transfer free energy, CHP/water (Lawson et al., 1984)
39	40	4055	5	-14	0.097701	0.229470	0.168942	0.165371	0.161109	SNEP660103	Principal component III (Sneath, 1966)
40	25	1357	7	-12	0.127962	0.076036	0.363164	0.189054	0.159663	FINA910103	Helix termination parameter at posision j-2,j-1,j (Finkelstein et al., 1991)
41	38	1077	15	-4	0.166414	0.069268	0.275156	0.170279	0.158992	FASG760103	Optical rotation (Fasman, 1976)
42	45	1138	4	-15	0.226077	0.142133	0.120250	0.162820	0.158949	FAUJ880104	STERIMOL length of the side chain (Fauchere et al., 1988)
43	34	1243	1	-18	0.081013	0.139495	0.311954	0.177487	0.158400	FAUJ880111	Positive charge (Fauchere et al., 1988)
44	46	1071	9	-10	0.087972	0.236205	0.163132	0.162436	0.155708	FASG760103	Optical rotation (Fasman, 1976)
45	48	5199	15	-4	0.201460	0.096578	0.181245	0.159761	0.154598	GEOR030104	Linker propensity from 3-linker dataset (George-Heringa, 2003)
46	53	961	7	-12	0.145998	0.141975	0.175046	0.154340	0.154008	DESM900101	Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)
47	52	2488	4	-15	0.186172	0.153875	0.124298	0.154782	0.153647	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
48	50	33	15	-4	0.240227	0.111670	0.126279	0.159392	0.153279	ARGP820101	Hydrophobicity index (Argos et al., 1982)
49	39	1993	13	-6	0.253351	0.055979	0.193114	0.167481	0.153101	LAWE840101	Transfer free energy, CHP/water (Lawson et al., 1984)
50	36	4375	1	-18	0.079212	0.124014	0.311370	0.171532	0.152983	WOLS870103	Principal property value z3 (Wold et al., 1987)
51	51	4867	7	-12	0.128801	0.106998	0.239489	0.158429	0.151666	NADH010103	Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)
52	55	4273	7	-12	0.147026	0.113620	0.199112	0.153253	0.150605	WARP780101	Average interactions per side chain atom (Warme-Morgan, 1978)
53	44	4059	9	-10	0.077121	0.188361	0.223340	0.162941	0.150437	SNEP660103	Principal component III (Sneath, 1966)
54	57	2187	9	-10	0.173389	0.143118	0.130687	0.149065	0.147852	MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
55	59	4815	9	-10	0.124872	0.145550	0.174561	0.148328	0.147327	TAKK010101	Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
56	49	1293	15	-4	0.145680	0.057305	0.276129	0.159705	0.146917	FAUJ880113	pK-a(RCOOH) (Fauchere et al., 1988)
57	58	5296	4	-15	0.190680	0.130995	0.124921	0.148865	0.146897	GUYH850105	Apparent partition energies calculated from Chothia index (Guy, 1985)

Continued on next page...

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
58	56	1645	7	-12	0.104437	0.169961	0.177039	0.150479	0.146661	HOPT810101	Hydrophilicity value (Hopp-Woods, 1981)
59	60	3552	6	-13	0.103540	0.166763	0.173266	0.147856	0.145493	RACS820111	Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)
60	54	4239	9	-10	0.090178	0.156235	0.213889	0.153434	0.144316	VASM830103	Relative population of conformational state E (Vasquez et al., 1983)
61	63	544	4	-15	0.175066	0.120857	0.139346	0.145090	0.143598	CHOC760104	Proportion of residues 100% buried (Chothia, 1976)
62	61	1822	4	-15	0.190083	0.132268	0.117427	0.146593	0.143481	JANJ790102	Transfer free energy (Janin, 1979)
63	69	2470	4	-15	0.153865	0.145951	0.128181	0.142666	0.141876	OOBM770101	Average non-bonded energy per atom (Oobatake-Ooi, 1977)
64	67	4723	7	-12	0.144772	0.108922	0.174366	0.142687	0.140971	MONM990101	Turn propensity scale for transmembrane helices (Monne et al., 1999)
65	66	1897	7	-12	0.099930	0.146639	0.185845	0.144138	0.140764	KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)
66	62	3569	5	-14	0.172342	0.190831	0.074854	0.146009	0.138882	RACS820112	Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)
67	72	24	6	-13	0.100955	0.132631	0.190307	0.141298	0.138431	ARGP820101	Hydrophobicity index (Argos et al., 1982)
68	80	4184	8	-11	0.133086	0.134182	0.146037	0.137768	0.137642	TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga, 1977)
69	78	3320	8	-11	0.167397	0.127798	0.119483	0.138226	0.137043	QIAN880137	Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)
70	68	3957	15	-4	0.122669	0.086129	0.219256	0.142685	0.136914	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki, 1976)
71	81	4651	7	-12	0.110819	0.175557	0.126657	0.137678	0.136604	AURR980119	Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)
72	70	4687	7	-12	0.079474	0.162219	0.185402	0.142365	0.136233	VINM940104	Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)
73	79	2491	7	-12	0.153014	0.103627	0.156901	0.137847	0.136147	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
74	76	4236	6	-13	0.085822	0.153510	0.179655	0.139662	0.136082	VASM830103	Relative population of conformational state E (Vasquez et al., 1983)
75	75	2500	16	-3	0.128131	0.088287	0.202849	0.139756	0.134643	OOBM770102	Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)
76	87	436	4	-15	0.146679	0.138136	0.120265	0.135027	0.134394	CHAM830105	The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)
77	64	4426	16	-3	0.110641	0.070664	0.252117	0.144474	0.133591	YUTK870103	Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)

Continued on next page...

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
78	82	3903	15	-4	0.155802	0.076729	0.177979	0.136837	0.133103	RICJ880116	Relative preference value at C' (Richardson-Richardson, 1988)
79	84	1679	5	-14	0.120645	0.193976	0.094433	0.136351	0.131866	HUTJ700102	Absolute entropy (Hutchens, 1970)
80	86	547	7	-12	0.124828	0.086449	0.197378	0.136218	0.130635	CHOC760104	Proportion of residues 100% buried (Chothia, 1976)
81	85	329	5	-14	0.100908	0.216000	0.092129	0.136346	0.129966	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
82	96	904	4	-15	0.165832	0.118690	0.105825	0.130116	0.128629	DAWD720101	Size (Dawson, 1972)
83	98	1931	5	-14	0.111392	0.143106	0.131507	0.128668	0.127937	KLEP840101	Net charge (Klein et al., 1984)
84	99	3192	6	-13	0.119839	0.108958	0.156612	0.128470	0.127702	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)
85	89	4137	15	-4	0.152863	0.072168	0.173292	0.132774	0.127695	TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)
86	71	1675	1	-18	0.065887	0.105882	0.253267	0.141679	0.127189	HUTJ700102	Absolute entropy (Hutchens, 1970)
87	94	502	16	-3	0.126641	0.091895	0.173774	0.130770	0.127156	CHOC760101	Residue accessible surface area in tripeptide (Chothia, 1976)
88	74	3643	7	-12	0.131777	0.061197	0.227717	0.140230	0.127131	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
89	73	529	7	-12	0.119550	0.066681	0.237396	0.141209	0.127069	CHOC760103	Proportion of residues 95% buried (Chothia, 1976)
90	103	3640	4	-15	0.148368	0.130464	0.102490	0.127107	0.126467	RADA880104	Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)
91	88	4677	15	-4	0.205476	0.082235	0.113164	0.133625	0.126263	AURR980120	Normalized positional residue frequency at helix termini C4' (Aurora-Rose, 1998)
92	104	3723	15	-4	0.123617	0.102969	0.154671	0.127086	0.125902	RICJ880104	Relative preference value at N1 (Richardson-Richardson, 1988)
93	93	3950	8	-11	0.155858	0.155964	0.081116	0.130979	0.125762	ROBB760109	Information measure for N-terminal turn (Robson-Suzuki, 1976)
94	90	248	14	-5	0.204192	0.166714	0.026012	0.132306	0.125651	BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
95	92	1024	16	-3	0.120141	0.070837	0.204552	0.131843	0.125530	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
96	91	2239	7	-12	0.102187	0.083460	0.210192	0.131946	0.124746	MEIH800103	Average side chain orientation angle (Meirovitch et al., 1980)
97	65	1009	1	-18	0.038614	0.128491	0.265324	0.144143	0.124252	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
98	83	3985	7	-12	0.124856	0.060462	0.224393	0.136570	0.124072	ROSM880102	Side chain hydropathy, corrected for solvation (Roseman, 1988)

Continued on next page...

Table B.2: Top 100 C-Termini Dataset F-score Results

Full Dataset Rank	3-Fold Avg. Rank	LibSVM Fea. No.	Win. Pos.	C-Term. Resid.	Fold1 F-score	Fold2 F-score	Fold3 F-score	3-Fold Avg. F-score	Full Dataset F-score	AAIndex Entry	AAIndex Description
99	106	142	16	-3	0.115255	0.103074	0.160976	0.126435	0.123965	BIGC670101	Residue volume (Bigelow, 1967)
100	100	3139	7	-12	0.094967	0.100320	0.188478	0.127922	0.123015	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
1	61	0.821117	0.638160	0.678736	0.659310	0.934040	0.636917	0.909086	0.654967	0.742489	0.786572	0.746139	FASG760104	pK-N (Fasman, 1976)
2	225	0.653391	0.533899	0.591305	0.780792	0.730538	0.646936	0.581665	0.532083	0.660186	0.502686	0.621348	SNEP660102	Principal component II (Sneath, 1966)
3	239	0.649209	0.603464	0.561393	0.578957	0.747805	0.392561	0.619526	0.394651	0.492890	0.344407	0.538486	WEBA780101	RF value in high salt chromatography (Weber-Lacey, 1978)
4	113	0.471013	0.516818	0.573641	0.533295	0.517343	0.590914	0.517337	0.435128	0.612626	0.367932	0.513605	LEVM760106	van der Waals parameter R0 (Levitt, 1976)
5	43	0.492162	0.643259	0.362643	0.418367	0.463326	0.551919	0.618716	0.449975	0.576852	0.400375	0.497759	CHOP780213	Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)
6	80	0.494569	0.600402	0.509506	0.424301	0.453984	0.574960	0.448431	0.479243	0.520227	0.389059	0.489468	GEIM800102	Alpha-helix indices for alpha-proteins (Geisow-Roberts, 1980)
7	20	0.520011	0.436490	0.402587	0.559230	0.683409	0.378144	0.602170	0.379568	0.450398	0.466626	0.487863	CHAM820102	Free energy of solution in water, kcal/mole (Charton-Charton, 1982)
8	78	0.541133	0.464366	0.327981	0.515240	0.456541	0.518785	0.400784	0.340436	0.503365	0.508862	0.457749	GARJ730101	Partition coefficient (Garel et al., 1973)
9	68	0.518372	0.367255	0.393649	0.543078	0.531796	0.389920	0.494176	0.297869	0.466445	0.516661	0.451922	FAUJ880108	Localized electrical effect (Fauchere et al., 1988)
10	278	0.356551	0.369999	0.505872	0.484052	0.499053	0.585136	0.398707	0.356421	0.494508	0.346144	0.439644	FUKS010105	Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)
11	173	0.503075	0.494741	0.338476	0.432509	0.529619	0.354325	0.552496	0.346962	0.430828	0.385215	0.436825	QIAN880118	Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)
12	171	0.512281	0.505223	0.354595	0.331697	0.491031	0.379236	0.565287	0.372654	0.435587	0.367914	0.431551	QIAN880116	Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)
13	18	0.558170	0.403161	0.329720	0.553508	0.514552	0.381908	0.446821	0.259202	0.408968	0.380179	0.423619	CHAM810101	Steric parameter (Charton, 1981)
14	236	0.507047	0.469834	0.333099	0.506293	0.516438	0.390481	0.394452	0.318288	0.424591	0.351246	0.421177	VASM830103	Relative population of conformational state E (Vasquez et al., 1983)
15	96	0.309727	0.403359	0.479794	0.421960	0.386615	0.603318	0.426922	0.444243	0.395975	0.251621	0.412353	ISOY800102	Normalized relative frequency of extended structure (Isogai et al., 1980)

Continued on next page...

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
16	118	0.443686	0.377860	0.380168	0.419756	0.457175	0.389762	0.390745	0.339003	0.437704	0.345545	0.398140	LIFS790103	Conformational preference for antiparallel beta-strands (Lifson-Sander, 1979)
17	60	0.378326	0.339253	0.480240	0.421439	0.501875	0.352313	0.461339	0.315761	0.412944	0.315724	0.397921	FASG760103	Optical rotation (Fasman, 1976)
18	175	0.456116	0.390377	0.303400	0.388348	0.480495	0.345382	0.477022	0.330815	0.374186	0.374151	0.392029	QIAN880122	Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
19	100	0.461755	0.407835	0.297383	0.422389	0.479812	0.338299	0.450205	0.290378	0.378083	0.370874	0.389701	ISOY800108	Normalized relative frequency of coil (Isogai et al., 1980)
20	26	0.409336	0.362872	0.380939	0.378059	0.433365	0.433333	0.411129	0.378986	0.381890	0.297754	0.386766	CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
21	210	0.449578	0.339193	0.351203	0.473187	0.492630	0.368834	0.415845	0.284270	0.357457	0.280362	0.381256	RICJ880108	Relative preference value at N5 (Richardson-Richardson, 1988)
22	19	0.348698	0.273434	0.324210	0.341149	0.394510	0.428814	0.376280	0.340140	0.417286	0.542572	0.378709	CHAM820101	Polarizability parameter (Charton-Charton, 1982)
23	9	0.313723	0.338668	0.255613	0.432227	0.472083	0.476162	0.305780	0.390433	0.367287	0.393542	0.374552	BIOV880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)
24	53	0.365724	0.313762	0.318631	0.644061	0.390775	0.492013	0.328135	0.274246	0.313026	0.262616	0.370299	DAYM780201	Relative mutability (Dayhoff et al., 1978b)
25	244	0.374350	0.354888	0.369565	0.401363	0.482160	0.332148	0.405242	0.266047	0.398308	0.285201	0.366927	WOLS870103	Principal property value z3 (Wold et al., 1987)
26	37	0.395397	0.283912	0.308191	0.467815	0.448417	0.338015	0.362201	0.240987	0.336912	0.309930	0.349178	CHOP780207	Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)
27	112	0.263744	0.241722	0.474196	0.233203	0.335065	0.323363	0.374863	0.443000	0.441976	0.270320	0.340145	LEVM760103	Side chain angle theta(AAR) (Levitt, 1976)
28	56	0.357489	0.242869	0.309787	0.394734	0.365325	0.377181	0.329031	0.249825	0.342544	0.335740	0.330453	EISD860101	Solvation free energy (Eisenberg-McLachlan, 1986)
29	243	0.313252	0.336491	0.395135	0.242527	0.368862	0.340304	0.279800	0.316403	0.372818	0.265614	0.323121	WOLS870102	Principal property value z2 (Wold et al., 1987)
30	140	0.273914	0.368187	0.243276	0.334589	0.339157	0.449380	0.308934	0.360581	0.310741	0.240199	0.322896	OOBM770103	Long range non-bonded energy per atom (Oobatake-Ooi, 1977)
31	174	0.338605	0.267548	0.285782	0.289971	0.374526	0.347266	0.327321	0.343934	0.307518	0.329083	0.321155	QIAN880121	Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)
32	66	0.354593	0.303184	0.230450	0.403258	0.391385	0.325682	0.350293	0.234377	0.308210	0.309386	0.321082	FAUJ880106	STERIMOL maximum width of the side chain (Fauchere et al., 1988)
33	75	0.309969	0.295338	0.295473	0.318742	0.350443	0.430276	0.329204	0.329305	0.276810	0.260213	0.319577	FINA910102	Helix initiation parameter at posision i & i+1 & i+2 (Finkelstein et al., 1991)
34	224	0.275718	0.446880	0.189305	0.377287	0.288655	0.358802	0.352985	0.260267	0.346234	0.280058	0.317619	SNEP660101	Principal component I (Sneath, 1966)
35	169	0.365192	0.310394	0.248576	0.379565	0.421937	0.288891	0.374375	0.219467	0.292381	0.268936	0.316971	QIAN880112	Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)
36	71	0.287631	0.290641	0.242118	0.341851	0.442971	0.268277	0.250189	0.297634	0.345069	0.283005	0.304939	FAUJ880112	Negative charge (Fauchere et al., 1988)

Continued on next page...

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
37	135	0.363938	0.451282	0.219436	0.268778	0.351170	0.225463	0.419678	0.199799	0.294713	0.242245	0.303650	NAKH920101	AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)
38	273	0.324722	0.356221	0.318674	0.283931	0.363752	0.239538	0.345545	0.244121	0.330727	0.217986	0.302522	NADH010107	Hydropathy scale based on self-information values in the two-state model (50% accessibility) (Naderi-Manesh et al., 2001)
39	31	0.354066	0.303111	0.250171	0.389805	0.383505	0.285807	0.316148	0.195854	0.285507	0.234988	0.299896	CHOC760104	Proportion of residues 100% buried (Chothia, 1976)
40	74	0.319608	0.299053	0.231336	0.282717	0.346236	0.341177	0.343254	0.301441	0.252178	0.263124	0.298012	FINA910101	Helix initiation parameter at posision i-1 (Finkelstein et al., 1991)
41	89	0.260637	0.350314	0.272389	0.275094	0.293168	0.299870	0.363031	0.346898	0.261216	0.156537	0.287915	GRAR740103	Volume (Grantham, 1974)
42	300	0.251914	0.241838	0.253855	0.300047	0.184628	0.312944	0.231454	0.244522	0.318067	0.475862	0.281513	(NA)	Peptide Length
43	172	0.268356	0.261452	0.274611	0.202087	0.272470	0.332614	0.345631	0.306109	0.289088	0.258881	0.281130	QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)
44	116	0.358534	0.228260	0.290537	0.288389	0.317710	0.252448	0.286852	0.178577	0.315603	0.250847	0.276776	LEWP710101	Frequency of occurrence in beta-bends (Lewis et al., 1971)
45	157	0.258467	0.379132	0.199781	0.218141	0.327425	0.195044	0.343743	0.284221	0.327760	0.233660	0.276737	PONP800106	Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980)
46	290	0.264719	0.256321	0.266838	0.221317	0.286374	0.321803	0.287128	0.360428	0.243380	0.212750	0.272106	GEOR030105	Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)
47	122	0.289576	0.283774	0.192912	0.253980	0.323371	0.316636	0.327772	0.261137	0.232728	0.238199	0.272009	MEEJ800101	Retention coefficient in HPLC, pH7.4 (Meek, 1980)
48	146	0.341952	0.324479	0.183792	0.309140	0.411583	0.169822	0.340988	0.146277	0.244250	0.233918	0.270620	PALJ810105	Normalized frequency of turn from LG (Palau et al., 1981)
49	177	0.341000	0.428402	0.180532	0.261430	0.306972	0.204315	0.353080	0.143377	0.229070	0.215985	0.266416	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
50	209	0.332121	0.334280	0.139910	0.334203	0.326489	0.233252	0.321570	0.136927	0.220556	0.232864	0.261217	RICJ880107	Relative preference value at N4 (Richardson-Richardson, 1988)
51	3	0.341222	0.245818	0.256638	0.243113	0.262103	0.243037	0.373162	0.170636	0.234082	0.236321	0.260613	ARGP820102	Signal sequence helical potential (Argos et al., 1982)
52	58	0.300432	0.238464	0.307543	0.271050	0.316934	0.198071	0.277916	0.191164	0.282016	0.215337	0.259893	EISD860103	Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)
53	49	0.291168	0.185093	0.242006	0.316168	0.295701	0.279663	0.282224	0.180419	0.254307	0.268160	0.259491	CRAJ730102	Normalized frequency of beta-sheet (Crawford et al., 1973)
54	297	0.287290	0.286008	0.171644	0.248042	0.305009	0.311988	0.297815	0.217097	0.217591	0.212559	0.255504	CASG920101	Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)
55	178	0.244628	0.471380	0.160297	0.183552	0.216035	0.259428	0.359010	0.218877	0.257726	0.170827	0.254176	QIAN880125	Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)

Continued on next page...

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
56	197	0.248799	0.208528	0.265710	0.243484	0.249370	0.370216	0.256221	0.249367	0.231245	0.211668	0.253461	RACS820110	Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)
57	226	0.309442	0.207870	0.217430	0.353755	0.288080	0.252362	0.246582	0.159950	0.242814	0.252945	0.253123	SNEP660103	Principal component III (Sneath, 1966)
58	247	0.281307	0.301752	0.179860	0.249504	0.313929	0.271012	0.301200	0.225491	0.199778	0.200828	0.252466	ZIMJ680101	Hydrophobicity (Zimmerman et al., 1968)
59	256	0.284348	0.191964	0.227490	0.272516	0.355483	0.216311	0.246824	0.174354	0.256044	0.276737	0.250207	AURR980116	Normalized positional residue frequency at helix termini Cc (Aurora-Rose, 1998)
60	277	0.278143	0.241751	0.225671	0.260068	0.309753	0.271619	0.260092	0.262373	0.193444	0.198701	0.250162	FUKS010103	Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)
61	268	0.311609	0.214183	0.201643	0.261823	0.402251	0.146952	0.354273	0.164062	0.201853	0.219725	0.247837	TAKK010101	Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)
62	102	0.339060	0.386190	0.183046	0.210979	0.296552	0.174377	0.318856	0.139870	0.217634	0.185797	0.245236	JANJ790102	Transfer free energy (Janin, 1979)
63	27	0.225961	0.232447	0.262054	0.219747	0.243871	0.357673	0.219051	0.251102	0.238570	0.197108	0.244758	CHAM830108	A parameter of charge transfer donor capability (Charton-Charton, 1983)
64	81	0.284464	0.275455	0.178922	0.310393	0.319572	0.192563	0.261027	0.192568	0.204466	0.179405	0.239884	GEIM800103	Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)
65	95	0.190388	0.288477	0.182625	0.191692	0.214837	0.301254	0.285083	0.299281	0.226029	0.184557	0.236422	ISOY800101	Normalized relative frequency of alpha-helix (Isogai et al., 1980)
66	145	0.268598	0.339608	0.159671	0.220612	0.273637	0.169798	0.268427	0.169424	0.264780	0.224038	0.235859	OOBM850105	Optimized side chain interaction parameter (Oobatake et al., 1985)
67	202	0.192315	0.107177	0.234311	0.168533	0.201019	0.275774	0.235027	0.250613	0.312963	0.366652	0.234438	RADA880103	Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)
68	11	0.277446	0.200872	0.226774	0.302647	0.272367	0.225122	0.211487	0.126285	0.252504	0.194405	0.228991	BROC820101	Retention coefficient in TFA (Browne et al., 1982)
69	162	0.228877	0.227210	0.211130	0.172711	0.246251	0.248139	0.308090	0.253428	0.202369	0.180579	0.227878	PTIO830101	Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)
70	150	0.182548	0.165920	0.174735	0.171822	0.225753	0.316687	0.194036	0.222882	0.281887	0.340177	0.227645	PALJ810114	Normalized frequency of turn in all-beta class (Palau et al., 1981)
71	164	0.223539	0.212000	0.178429	0.226887	0.263818	0.260204	0.284329	0.175899	0.208087	0.218500	0.225169	QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)
72	8	0.176095	0.219294	0.276340	0.216934	0.246790	0.306511	0.171430	0.252780	0.234408	0.142761	0.224334	BIGC670101	Residue volume (Bigelow, 1967)
73	141	0.243705	0.172131	0.263053	0.235625	0.294745	0.207805	0.225050	0.144241	0.248158	0.185903	0.222042	OOBM770104	Average non-bonded energy per residue (Oobatake-Ooi, 1977)
74	114	0.261804	0.111727	0.283877	0.167266	0.248929	0.228753	0.215188	0.227748	0.210121	0.259436	0.221485	LEVM780102	Normalized frequency of beta-sheet, with weights (Levitt, 1978)
75	131	0.196121	0.227074	0.163860	0.173825	0.323399	0.174293	0.280305	0.236637	0.181373	0.228794	0.218568	NAKH900109	AA composition of membrane proteins (Nakashima et al., 1990)

Continued on next page...

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
76	266	0.234434	0.203639	0.183797	0.188304	0.218667	0.305655	0.259583	0.212810	0.165152	0.158194	0.213024	KUMS000101	Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)
77	153	0.233754	0.167098	0.174943	0.226298	0.295985	0.242624	0.180377	0.157484	0.190074	0.254789	0.212343	PLIV810101	Partition coefficient (Pliska et al., 1981)
78	257	0.230000	0.124070	0.190251	0.216493	0.267411	0.266515	0.184442	0.168909	0.186039	0.221655	0.205579	AURR980117	Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)
79	204	0.209074	0.194948	0.167336	0.169182	0.222565	0.214048	0.290223	0.213644	0.174702	0.172745	0.202847	RADA880106	Accessible surface area (Radzicka-Wolfenden, 1988)
80	208	0.242176	0.185743	0.126704	0.285940	0.377055	0.180271	0.167846	0.137919	0.139246	0.150216	0.199312	RICJ880105	Relative preference value at N2 (Richardson-Richardson, 1988)
81	36	0.211800	0.161400	0.195826	0.127013	0.259668	0.139290	0.220070	0.242131	0.226099	0.182309	0.196561	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
82	84	0.153150	0.158273	0.194213	0.197774	0.251216	0.258972	0.191706	0.152750	0.218499	0.184917	0.196147	GEIM800106	Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)
83	45	0.218540	0.181667	0.141344	0.179763	0.192908	0.247120	0.244746	0.167207	0.162319	0.179516	0.191513	CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman, 1978b)
84	127	0.245761	0.168238	0.165562	0.236137	0.272719	0.145378	0.203915	0.098902	0.180942	0.183045	0.190060	NAGK730103	Normalized frequency of coil (Nagano, 1973)
85	254	0.105906	0.262239	0.198243	0.079641	0.113760	0.238667	0.160756	0.269730	0.286743	0.176302	0.189199	AURR980110	Normalized positional residue frequency at helix termini N5 (Aurora-Rose, 1998)
86	101	0.188291	0.291153	0.110354	0.164461	0.249747	0.121071	0.241602	0.165699	0.195937	0.149840	0.187816	JANJ790101	Ratio of buried and accessible molar fractions (Janin, 1979)
87	255	0.206238	0.144097	0.156398	0.258796	0.295272	0.149023	0.177429	0.120989	0.180204	0.179571	0.186802	AURR980112	Normalized positional residue frequency at helix termini C4 (Aurora-Rose, 1998)
88	168	0.198086	0.139706	0.216910	0.150261	0.243437	0.102715	0.181201	0.209842	0.214871	0.196470	0.185350	QIAN880110	Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)
89	46	0.212422	0.162877	0.142784	0.176692	0.227472	0.215354	0.221843	0.155671	0.143768	0.188135	0.184702	CIDH920101	Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)
90	192	0.191069	0.166257	0.174713	0.215040	0.293000	0.169965	0.190481	0.112946	0.159127	0.168106	0.184070	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)
91	275	0.144171	0.145444	0.212532	0.157481	0.204493	0.277461	0.147622	0.198811	0.173413	0.125002	0.178643	KOEP990102	Beta-sheet propensity derived from designed sequences (Koehl-Levitt, 1999)
92	293	0.163880	0.273087	0.113592	0.174622	0.138006	0.171169	0.160454	0.131068	0.228027	0.224828	0.177873	GEOR030109	Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)
93	288	0.144945	0.197524	0.149350	0.105359	0.267818	0.139237	0.172475	0.280049	0.166127	0.125116	0.174800	GEOR030101	Linker propensity from all dataset (George-Heringa, 2003)
94	126	0.135226	0.262149	0.123172	0.131589	0.194202	0.118684	0.235077	0.220916	0.179683	0.108720	0.170942	MIYS850101	Effective partition energy (Miyazawa-Jernigan, 1985)

Continued on next page...

Table B.3: Top 100 Fernandes et al. (2012) Dataset F-score Results

Avg. Rank	Feat. Num.	Run1 F-score	Run2 F-score	Run3 F-score	Run4 F-score	Run5 F-score	Run6 F-score	Run7 F-score	Run8 F-score	Run9 F-score	Run10 F-score	Avg. F-score	AAIndex Entry	AAIndex Description
95	291	0.171677	0.121157	0.163912	0.136296	0.176047	0.207240	0.188437	0.185151	0.161974	0.152909	0.166480	GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)
96	129	0.174287	0.109081	0.163768	0.225094	0.208850	0.175538	0.183751	0.083035	0.146936	0.177851	0.164819	NAKH900103	AA composition of mt-proteins (Nakashima et al., 1990)
97	104	0.229087	0.151373	0.132790	0.186195	0.229725	0.147490	0.153418	0.085654	0.149465	0.147500	0.161270	KANM800104	Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)
98	76	0.144611	0.199998	0.103336	0.154375	0.180726	0.167700	0.236539	0.182888	0.132347	0.108913	0.161143	FINA910103	Helix termination parameter at posision j-2 & j-1 & j (Finkelstein et al., 1991)
99	285	0.209755	0.124062	0.144096	0.157616	0.166067	0.159241	0.205947	0.106158	0.132114	0.188166	0.159322	WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce et al. 1995)
100	160	0.169000	0.087536	0.138823	0.164531	0.185503	0.242538	0.142420	0.126060	0.128465	0.187763	0.157264	PRAM900101	Hydrophobicity (Prabhakaran, 1990)

Appendix C: Cleavage Site Results

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	ARGP820103	0.5101419906	0.4373295147	0.0815408746	797.5	0.62	0.81
1	BEGF750101	0.4035970134	0.704359058	0.1058440309	1210	0.6	0.44
1	BEGF750102	0.7367497254	0.1137212107	0.4706307213	924	0.72	0.72
1	BEGF750103	0.521748489	0.4137428909	0.6669206605	959.5	0.75	0.75
1	BHAR880101	0.7823653392	0.0767756141	0.1126602136	1208.5	0.466	0.463
1	BIGC670101	0.4737700092	0.517608361	0.1071958539	1211.5	105.1	84.7
1	BIOV880101	0.3026807015	1.0749049544	0.5209096771	933	-13	-38
1	BIOV880102	0.0700100248	3.364141577	0.5248156445	933.5	-8	-54
1	BROC820101	0.8328599029	0.0448009508	0.8227956031	984.5	-1.2	-0.3
1	BROC820102	0.8725610377	0.0258805853	0.5899304008	1079.5	2.1	1.8
1	BULH740101	0.711332355	0.1378378545	0.7118088585	1058.5	-0.35	-0.39
1	BUNA790101	0.8721589306	0.0260456082	0.7119240967	1058.5	8.391	8.38
1	BUNA790103	0.2665940092	1.2500072107	0.6152044321	950.5	6.5	6.9
1	BURA740101	0.6260889206	0.2390784651	0.5487571434	1087	0.318	0.272
1	BURA740102	0.9844522259	0.0003819265	0.0985857093	808.5	0.318	0.362
1	CHAM810101	0.7641067305	0.0906151071	0.2962930688	885	0.68	0.68
1	CHAM820101	0.1586096221	2.0216060885	0.2758534407	1147.5	0.219	0.14
1	CHAM830101	0.8928033227	0.0182650464	0.3927918211	906.5	0.99	1.07
1	CHAM830102	0.3727581406	0.8026105299	0.3751801249	902.5	0.104	0.124
1	CHAM830103	0.6679409687	0.1852631579	0.1129871669	830.5	1	1
1	CHAM830104	0.8346639585	0.0438247012	0.9225404126	1034.5	1	1
1	CHAM830105	0.5257944724	0.4057377049	0.1387191035	1192.5	1	0
1	CHAM830106	0.0607707992	3.6082004556	0.3306742185	1132	4	2
1	CHAM830107	0.7989306982	0.0652818991	1	990	0	0
1	CHAM830108	0.8346639585	0.0438247012	0.5272488417	1102.5	1	0
1	CHOC750101	0.3046834681	1.0659842054	0.130317803	1199.5	171.3	141.7
1	CHOC760101	0.1807795113	1.8199567354	0.208995372	1168	200	160
1	CHOP780101	0.565542259	0.3327066374	0.6722657853	960	0.95	0.96
1	CHOP780201	0.2565447233	1.3042034467	0.1943104	1173	1.06	0.83
1	CHOP780202	0.727977257	0.121750379	0.1667602158	842	0.93	1.05
1	CHOP780203	0.9136401972	0.0118288111	0.4388228166	1108.5	1.01	0.98
1	CHOP780206	0.9727259164	0.0011756007	0.0981875867	808.5	0.93	0.99
1	CHOP780207	0.5105315385	0.4365226391	0.0572635095	1246.5	1.24	0.93

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	CHOP780208	0.0776900982	3.1865952188	0.7877389524	1046	1	1.07
1	CHOP780209	0.7774417877	0.0803853804	0.3568104467	899	0.79	0.9
1	CHOP780210	0.7248937294	0.1246457386	0.175071999	845.5	0.89	1.09
1	CHOP780211	0.4422500704	0.595807529	0.2132523943	858.5	0.82	1.05
1	CHOP780213	0.919759438	0.0102063053	0.6702479723	1065.5	0.085	0.085
1	CHOP780214	0.2281247484	1.4729718685	0.9307886135	1023.5	0.072	0.077
1	CHOP780215	0.5427516099	0.3733636358	0.5955046877	1078.5	0.085	0.085
1	CHOP780216	0.5330463256	0.391663093	0.4537326159	1105.5	1.05	0.98
1	CIDH920101	0.2288175392	1.468555869	0.7853130346	978.5	-0.24	-0.24
1	CIDH920102	0.1011538818	2.7444541443	0.3591183215	1125.5	-0.09	-0.4
1	CIDH920103	0.1376160491	2.2451106947	0.156597529	1187.5	-0.52	-0.6
1	CIDH920104	0.5186512426	0.4199460301	0.1841936493	1176.5	-0.57	-0.62
1	CIDH920105	0.2588408244	1.291588449	0.1106443454	1209.5	-0.41	-0.77
1	CRAJ730101	0.2280700869	1.4733209886	0.8797197282	993.5	0.96	0.94
1	CRAJ730102	0.9442085996	0.0049254148	0.9020462858	1028	1.18	0.87
1	CRAJ730103	0.8953442571	0.0174043805	0.6698503038	959.5	1.05	1.05
1	DAWD720101	0.0939271243	2.8674157303	0.0899025297	1220.5	6.5	5
1	DESM900101	0.1569307463	2.0381946453	0.1905204389	850.5	1.03	1.03
1	DESM900102	0.0694884565	3.3769614299	0.213513071	858.5	1.08	1.01
1	EISD860101	0.5516549301	0.357101773	0.1002532581	809.5	0	0.38
1	EISD860103	0.746561227	0.1051011064	0.1475908807	833.5	0	0
1	FASG760101	0.0998433725	2.766014185	0.8353485903	1038.5	146.19	131.17
1	FASG760102	0.907481722	0.0135840545	0.6180206755	1074.5	284	253
1	FASG760104	0.7237763483	0.1257043916	0.1668392657	1183	9.18	9.18
1	FASG760105	0.6542778735	0.2019230416	0.608537359	1076	2.16	2.17
1	FAUJ830101	0.1523925972	2.084079254	0.0985847788	808.5	0	0.13
1	FAUJ880101	0.5818257335	0.3055492036	0.4560085747	920	2.34	2.34
1	FAUJ880102	0.2127291257	1.5755274164	0.2278844369	1161.5	0.71	0.66
1	FAUJ880103	0.1140806775	2.5471199074	0.1366383976	1196.5	4.77	3
1	FAUJ880105	0.2836328724	1.1637712329	0.1596654517	866.5	1.52	1.52
1	FAUJ880106	0.0872339171	2.991016184	0.3698228788	1123.5	4.87	3.53
1	FAUJ880107	0.6828352079	0.1680631552	0.4833961531	925.5	11.1	11.4
1	FAUJ880109	0.0715244322	3.3275	0.8117691592	1041.5	0	1

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	FAUJ880110	0.3174132001	1.0110294118	0.0700723643	801.5	0	2
1	FAUJ880112	0.1560967345	2.0465116279	0.4943820225	967.5	0	0
1	FAUJ880113	0.2378380189	1.4125077719	0.1715773739	1181.5	4.3	4.25
1	FINA770101	0.8170882916	0.0538170382	0.1345906328	1196.5	1.05	0.95
1	GEIM800101	0.6580526347	0.1972344484	0.3025228835	1140	1	0.81
1	GEIM800102	0.2529513264	1.3242316722	0.285634678	880.5	1.08	1.06
1	GEIM800104	0.7958245593	0.0673612015	0.7300379252	969.5	1	0.94
1	GEIM800105	0.5792573667	0.3097317808	0.5354190963	935.5	1.04	1.04
1	GEIM800106	0.5114209681	0.4346843678	0.469966909	1102	1.11	1.14
1	GEIM800107	0.6319513151	0.2310336014	0.7852406229	978.5	0.93	0.96
1	GEIM800108	0.7984318739	0.0656134773	0.258399199	872.5	0.82	1
1	GEIM800109	0.181281827	1.8157277525	0.3833532396	1120.5	0.96	0.96
1	GEIM800110	0.1505533701	2.103125254	0.9052477041	1027.5	0.93	0.88
1	GEIM800111	0.6565935835	0.1990388792	0.5678150075	941.5	0.91	1.01
1	GOLD730101	0.9815156991	0.0005398493	0.1007059532	1214	1.5	0.69
1	GOLD730102	0.4897100602	0.4811916507	0.1071958539	1211.5	175.6	140.5
1	GRAR740102	1	7.55391054e-31	0.3986303702	1117	9	9
1	GRAR740103	0.520423069	0.416389413	0.1222112591	1203.5	119	83
1	GUYH850101	0.0726601327	3.3005721488	0.0969272554	1217.5	0.33	0.33
1	HUTJ700101	0.1382307677	2.2380248505	0.5623284842	940.5	40.35	38.3
1	ISOY800101	0.4609994109	0.5482628831	0.2653032175	1150.5	1.17	0.89
1	ISOY800102	0.5295646095	0.3983770662	0.1134811191	817	0.98	1.06
1	ISOY800103	0.5945276812	0.2854045126	0.7149056434	967	1.01	1.09
1	ISOY800104	0.8643793935	0.0293444633	0.4386948146	1108	0.97	0.97
1	ISOY800105	0.4674175679	0.5326874987	0.995168184	1011.5	0.83	0.85
1	ISOY800106	0.2230404983	1.5058877931	0.8603151724	1034.5	1.07	0.99
1	ISOY800107	0.3987987568	0.7189198783	0.9596569856	1006	0.79	0.87
1	ISOY800108	0.5387122489	0.3809063067	0.4537144696	1105.5	0.76	0.76
1	JANJ790101	0.1120933001	2.5757841397	0.2481839335	869.5	1.7	0.8
1	JOND750102	0.1467569709	2.1432974571	0.8226958154	1040.5	2.18	2.17
1	KANM800101	0.5750371652	0.3166856842	0.3210869417	1135.5	1	0.89
1	KANM800102	0.8368544995	0.042654396	0.1655854235	841	0.88	0.98
1	KANM800104	0.7214816045	0.127894366	0.5621990864	940.5	0.82	0.83

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	KARP850101	0.3107877097	1.0392619178	0.8447976427	988	1.041	1.073
1	KARP850102	0.5252211418	0.4068653628	0.2105438043	1167.5	1.028	1.025
1	KARP850103	0.2132180795	1.5721338797	0.292693116	1142.5	0.914	0.923
1	KHAG800101	0.1001371331	2.7611541182	0.0677272817	1237.5	54.7	31
1	KRIW790103	0.5149785557	0.4273870808	0.1262081346	1201.5	100	62
1	LEVM760102	0.0555237268	3.7654370567	0.1850412321	1176.5	2.94	1.98
1	LEVM760104	0.3183253382	1.0072037152	0.6410527351	1070.5	210.9	215
1	LEVM760106	0.1724785315	1.8919043356	0.7788636077	1047	6	6
1	LEVM780101	0.9800479782	0.0006290039	0.5403793301	1088.5	0.96	0.91
1	LEVM780102	0.7620738623	0.0922332865	0.9052486149	997.5	0.95	0.97
1	LEVM780103	0.941271353	0.0054586631	0.4488685012	918.5	0.88	1.04
1	LEVM780104	0.9557280143	0.0030995533	0.3948640453	1118	0.98	0.93
1	LEVM780105	0.7682855445	0.0873377383	0.1975002589	853	0.97	1
1	LEVM780106	0.8627378644	0.0300664093	0.9212460508	1000	0.9	0.93
1	LEWP710101	0.5010491857	0.4564716581	0.0743627735	793	0.27	0.42
1	LIFS790101	0.8404465545	0.040771046	0.4580516457	920.5	0.93	0.95
1	LIFS790102	0.9059120323	0.014051114	0.3525044816	1127.5	0.79	0.7
1	LIFS790103	0.9441917212	0.0049284003	0.1145120565	817.5	1.02	1.09
1	MANP780101	0.9488763425	0.0041346277	0.7361299093	1054.5	12.43	11.76
1	MAXF760101	0.3841911285	0.7648643881	0.1557973916	1188	1.18	0.92
1	MAXF760102	0.764083364	0.0906336182	0.1454755598	832.5	0.97	1.04
1	MAXF760106	0.3793484522	0.780654682	0.1823276844	847.5	1.01	1.01
1	MCMT640101	0.3861507609	0.7585554759	0.4440221146	1107.5	21.29	13.92
1	MEEJ800101	0.1489239301	2.120223451	0.191566382	851.5	0.8	1.2
1	MEEJ800102	0.7539266079	0.0988762983	0.2758917286	877.5	-0.5	-0.5
1	MEEJ810101	0.5606521733	0.341164439	0.3093861108	1138.5	-0.2	-0.4
1	MEEJ810102	0.6214828983	0.2455194679	0.7170551653	967.5	0.2	-0.6
1	MEIH800101	0.3756669861	0.7928522455	0.4166206513	1113	0.98	0.98
1	MEIH800103	0.2241952542	1.4983322517	0.4877385898	926.5	87	83
1	MIYS850101	0.7986000097	0.0655016141	0.6526776948	956.5	2.06	2.04
1	NAGK730101	0.2407195783	1.3951508976	0.2757983039	1146.5	0.97	0.87
1	NAGK730102	0.2742409354	1.2104633754	0.511550402	931	0.9	1.07
1	NAGK730103	0.9517366741	0.0036843526	0.6818024742	961.5	0.84	1.03

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	NAKH900103	0.5884969334	0.2948578846	0.0520142847	773	5.12	5.3
1	NAKH900107	0.1895616348	1.7479444468	0.5623551882	940.5	6.18	5.44
1	NAKH900109	0.2226214945	1.5086410769	0.0694305044	1236.5	6.36	5.66
1	NAKH900113	0.0665507985	3.4511809777	0.7667242813	975.5	1.24	0.92
1	NAKH920104	0.1102962869	2.6021967018	0.0920877307	1220.5	6.19	5.2
1	NAKH920105	0.0543507597	3.8027840642	0.9116370469	998.5	6.47	4.17
1	NAKH920107	0.6270462414	0.2377530954	0.4344563275	1109.5	4.93	5.75
1	NISK800101	0.9557123841	0.0031017446	0.9822946174	1009.5	-0.07	-0.26
1	NISK860101	0.7245563212	0.1249648817	0.6468630564	955.5	-0.93	-1.2
1	NOZY710101	0.781553197	0.0773648873	0.3854189681	917.5	0	0
1	OOBM770103	0.2910143591	1.1284424506	0.9244557644	1000.5	-0.534	-0.534
1	OOBM770104	0.5716077257	0.3224118048	0.711898371	1058.5	-12.366	-12.48
1	OOBM770105	0.6902787406	0.1598333443	0.9887318149	1010.5	-9.666	-9.424
1	OOBM850101	0.5023172632	0.4537663011	0.2416713792	1157.5	2.55	1.47
1	OOBM850102	0.3703946978	0.8106187701	0.3250660821	890.5	0.95	1.09
1	OOBM850103	0.8254721539	0.0489163375	0.3250430031	890.5	0.43	-0.33
1	OOBM850104	0.6107826844	0.2608983408	0.1547098148	836.5	-3.97	-1.6
1	PALJ810101	0.6520738249	0.2046913606	0.2101455873	1167.5	0.95	0.9
1	PALJ810102	0.0764485794	3.2139614445	0.2195607223	1164.5	1.05	0.86
1	PALJ810103	0.7154723383	0.1337307975	0.2369204359	866.5	1.03	1.03
1	PALJ810104	0.9969982229	1.42346359e-05	0.4577135229	921	0.83	0.94
1	PALJ810105	0.4919198439	0.4762986167	0.4635327317	921.5	0.91	1.05
1	PALJ810106	0.4127813263	0.6771874273	0.3330408517	892.5	0.93	0.96
1	PALJ810108	0.7268905348	0.1227664211	0.1850559457	1176.5	1.02	0.9
1	PALJ810109	0.7231485584	0.1263013998	0.7544406918	973.5	0.92	0.97
1	PALJ810110	0.623836909	0.2422143397	0.8445606097	1037	1.02	1.06
1	PALJ810111	0.5322211829	0.393247055	0.1531619972	836	0.94	0.99
1	PALJ810112	0.7232112235	0.1262417356	0.5168029341	1093	0.98	0.98
1	PALJ810113	0.2419639334	1.3877345443	0.8827349849	994	1.18	0.88
1	PALJ810114	0.5741791547	0.3181119442	0.9372810413	1022.5	0.69	1.08
1	PALJ810115	0.0674629173	3.4277609488	0.8447077442	1037	0.91	0.87
1	PARJ860101	0.7317525763	0.1182573739	0.385291949	905.5	4.2	5.2
1	PLIV810101	0.580203208	0.3081871471	0.2385962322	867	-3.25	-2.91

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	PONP800101	0.9723293765	0.0012100476	0.661428048	958	12.01	11.65
1	PONP800102	0.6829695166	0.1679125208	0.40189489	908.5	7.31	7.08
1	PONP800103	0.3896849386	0.7472930349	0.1538490764	837	2.55	2.55
1	PONP800104	0.18529885	1.7823964666	0.7179214637	1057.5	12.88	12.88
1	PONP800105	0.212451626	1.577457555	0.5608740359	940.5	14.18	14.18
1	PONP800106	0.9045962324	0.0144487987	0.1606874295	839	11.05	11.18
1	PONP800107	0.2911363114	1.1278684886	0.658438633	957.5	3.13	2.7
1	PONP800108	0.3352433561	0.9387952919	0.7667598153	975.5	6.05	5.81
1	PRAM820102	0.6582928534	0.1969383109	0.4367094317	916	0.096	0.1
1	PRAM820103	0.2028059954	1.6464927551	0.3453952447	895.5	0.577	0.59
1	PRAM900102	0.9800479782	0.0006290039	0.5403793301	1088.5	0.96	0.91
1	PRAM900103	0.7620738623	0.0922332865	0.9052486149	997.5	0.95	0.97
1	PRAM900104	0.9168904049	0.0109519631	0.4344584858	915.5	0.88	1.03
1	PTIO830101	0.2673458902	1.246055855	0.6200335775	1073.5	0.95	0.95
1	PTIO830102	0.9315631108	0.007417428	0.2039374581	856.5	0.7	1
1	QIAN880102	0.4011285218	0.7118180665	0.1313924865	1198.5	0.02	-0.03
1	QIAN880103	0.1098904228	2.6082287304	0.2305718527	864.5	-0.09	-0.07
1	QIAN880104	0.5055009533	0.4470255406	0.4586281636	920.5	-0.03	-0.04
1	QIAN880105	0.2102228529	1.5930705188	0.9148377735	1026	0.03	-0.1
1	QIAN880106	0.9042214188	0.014563112	0.4226307064	1112	0.13	-0.18
1	QIAN880107	0.2391438153	1.4046105116	0.1366574881	1196.5	0.16	-0.35
1	QIAN880108	0.3560862596	0.8606792922	0.1522359559	1189.5	0.23	-0.19
1	QIAN880109	0.5754745256	0.3159602916	0.0651338962	1240	0.15	-0.07
1	QIAN880111	0.1037057268	2.7033490963	0.152366033	1189.5	0	-0.04
1	QIAN880113	0.2314960118	1.4516349439	0.0514546679	1252.5	0	-0.08
1	QIAN880116	0.3481723324	0.889583936	0.1769682887	845.5	-0.09	0.02
1	QIAN880118	0.7226802863	0.1267477502	0.230381232	1161	0.25	0.11
1	QIAN880119	0.8751578945	0.0248277992	0.1643621244	840.5	-0.04	0.19
1	QIAN880120	0.6983128065	0.1512186495	0.4635519062	921.5	-0.12	-0.02
1	QIAN880121	0.3811345583	0.774797445	0.1642788321	840.5	-0.26	-0.09
1	QIAN880122	0.2204917141	1.522733751	0.6908525886	1062	0.05	-0.1
1	QIAN880124	0.3528956403	0.8722250281	0.3129333542	887.5	-0.1	0.06
1	QIAN880125	0.7207284897	0.1286177515	0.0944814243	806	-0.03	0.03

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	QIAN880128	0.1359270647	2.2647665479	0.422508576	1112	0.07	0.04
1	QIAN880131	0.0898950781	2.9406600454	0.2424710159	868	0.09	0.12
1	QIAN880132	0.103017532	2.7143228016	0.7853057008	1046.5	-0.27	-0.1
1	QIAN880133	0.9192789047	0.0103293433	0.6468443757	955.5	-0.23	-0.01
1	QIAN880134	0.6610525786	0.1935552609	0.93728163	1022.5	-0.04	0.01
1	QIAN880135	0.9723632077	0.0012070893	0.5720438865	942.5	-0.14	0.21
1	QIAN880136	0.8381837188	0.0419523071	0.1693609955	843	-0.13	-0.11
1	QIAN880138	0.0762000177	3.2194987868	0.4885981848	1098.5	-0.02	-0.02
1	QIAN880139	0.6383218526	0.2224830476	0.81023667	1042.5	-0.01	-0.01
1	RACS770101	0.8651643844	0.029002416	0.7177737677	1057.5	0.962	0.986
1	RACS770103	0.0675128298	3.4264893206	0.1034113168	1213.5	1.63	1.61
1	RACS820101	0.4775694538	0.5087460768	0.2526303506	871	0.89	1.5
1	RACS820102	0.7224882839	0.1269310209	0.3622032201	1125.5	1.14	1
1	RACS820103	0.6658058727	0.1878102418	0.0600441822	1243	1.63	1.63
1	RACS820104	0.2768340985	1.1973699361	0.0858726012	800.5	1.11	1.31
1	RACS820107	0.1744443092	1.8745055295	0.4064562476	909.5	1.2	1.24
1	RACS820108	0.2142578214	1.5649487057	0.4488744948	1106.5	1.02	0.99
1	RACS820109	0.6907946515	0.1592718329	0.6998514639	970.5	0	0
1	RACS820110	0.4874541752	0.4862251387	0.4659956018	922	1	1.05
1	RACS820111	0.391540173	0.7414397342	0.6067462528	1076.5	1.03	1.31
1	RACS820112	0.885202965	0.0209659145	0.3210539593	889.5	1.19	1.18
1	RACS820113	0.6160314998	0.2532812074	0.2618266272	873.5	17.06	19.95
1	RACS820114	0.1382119734	2.23824096	0.294161663	882.5	19.61	20.61
1	RADA880101	0.1349302263	2.2764982226	0.8797135437	993.5	0.94	-2.57
1	RADA880102	0.7152504949	0.1339490875	0.6667045243	959	0.08	0.04
1	RADA880103	0.1372247787	2.249639651	0.6879548698	962.5	-3.74	-2.77
1	RADA880104	0.114455501	2.5417764866	0.7729082688	976.5	0.89	-2.84
1	RADA880106	0.1272354884	2.3705025834	0.1202604392	1204.5	215.2	157.2
1	RICJ880104	0.7527849043	0.0998274957	0.0586634002	783	0.6	0.7
1	RICJ880107	0.4816475049	0.4993620308	0.1115774055	1207	1.3	0.6
1	RICJ880108	0.23333093	1.4401805869	0.1905934727	1173.5	1.1	1
1	RICJ880109	0.6591904927	0.1958340751	0.3839057695	1119.5	1.2	1
1	RICJ880110	0.4896957016	0.4812235656	0.569095562	1082.5	1	0.6

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	RICJ880111	0.6895117426	0.1606702579	0.1312302374	1198	1	0.6
1	RICJ880114	0.2523093751	1.3278470133	0.2628295338	875.5	0.7	1
1	RICJ880115	0.1160240575	2.5196263032	0.0832909701	1225.5	0.9	0.8
1	RICJ880116	0.5373065041	0.3835557469	0.1911029468	1172.5	1.2	0.9
1	ROBB760101	0.8394395539	0.0412945573	0.1202655795	1204.5	1.4	-1.3
1	ROBB760103	0.202349762	1.6498549867	0.383950161	1120.5	0.5	-3
1	ROBB760104	0.5281569912	0.4011141677	0.0725724375	1234	1.6	-0.6
1	ROBB760105	0.3466584026	0.895216492	0.2978674945	883.5	0.4	1.2
1	ROBB760106	0.9402252032	0.005655243	0.3860769257	905	0.4	0.8
1	ROBB760107	0.2967869137	1.1016103158	0.1523205088	835.5	0.7	0.7
1	ROBB760108	0.8828429247	0.021843266	0.9308499664	1001.5	1	0.4
1	ROBB760110	0.8880271619	0.0199401348	0.3705569451	1123.5	2	0.6
1	ROBB760111	0.5131571539	0.4311120561	0.3168790636	1136.5	1.3	1.2
1	ROBB760112	0.2089442408	1.6021174859	0.4534318164	919.5	-1.2	-0.5
1	ROBB760113	0.7342165652	0.116008302	0.9951694524	1013.5	1	0.2
1	ROBB790101	0.1261208503	2.384652205	0.9788824388	1016	0.3	0.3
1	ROSG850101	0.5749934828	0.3167581948	0.8165029121	1041.5	119.2	119.2
1	ROSM880101	0.1558549555	2.0489322956	0.8543136608	989.5	0	3.86
1	SIMZ760101	0.7684625915	0.0872003346	0.0772048679	1230.5	1.5	0.54
1	SNEP660101	0.720592156	0.1287489498	0.1932823633	851.5	0.228	0.234
1	SNEP660102	0.631600528	0.2315102	0.1395829321	1195	-0.011	0.022
1	SUEM840101	0.2231584503	1.5051138599	0.9052403211	1027.5	0.95	0.922
1	SWER830101	0.4914594695	0.4773149524	0.4268246949	914.5	-0.59	-0.55
1	TANS770101	0.7526219853	0.0999636384	0.0857778903	1224.5	1.06	0.78
1	TANS770104	0.6408951703	0.2190850433	0.3928787355	906.5	1.015	0.831
1	TANS770105	0.4242557166	0.6444809772	0.9212395328	1025	0.932	0.749
1	TANS770106	0.5381294657	0.3820031488	0.9565557645	1019.5	0.901	1.08
1	TANS770107	0.9564352767	0.0030012118	0.0693474938	1236.5	1.288	0.285
1	TANS770108	0.1359293675	2.2647395597	0.3329349813	892.5	0.835	1.089
1	TANS770109	0.1165811033	2.5118407624	0.2164955771	859.5	0.704	0.863
1	TANS770110	0.3343337393	0.9423544441	0.2653041126	874.5	0.936	1.055
1	VASM830101	0.9915214926	0.0001135646	0.5300646606	1090.5	0.17	0.173
1	VASM830103	0.5873343063	0.2967033132	0.3209748005	889.5	0.194	0.198

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	VELV850101	0.6849641982	0.1656846628	0.8892610371	995	0.0371	0.08226
1	VENT840101	0.4740414859	0.5169712794	0.6294742667	1080	0	0
1	WARP780101	0.0617099167	3.5816129921	0.5407436546	936.5	7.98	7.08
1	WEBA780101	0.9030903926	0.0149108315	0.0769801131	1230	0.88	0.88
1	WERD780101	0.7764811229	0.0811001573	0.3542933123	898	0.49	0.49
1	WERD780103	0.242699245	1.3833741658	0.0800869692	796.5	-0.08	0.13
1	WERD780104	0.4597743227	0.5512754906	0.8161559498	1041.5	-0.11	-0.11
1	WOEC730101	0.8319244268	0.0453116024	0.0858621336	1224.5	7.9	7.5
1	WOLS870101	0.9631712861	0.002144245	0.439239161	1108.5	2.23	1.96
1	WOLS870102	0.1860272122	1.7764439801	0.7118786017	1058.5	1.3	-0.27
1	YUTK870101	0.5685291456	0.3276101432	0.0662158499	786	7.9	8.8
1	YUTK870102	0.2491338623	1.3459007098	0.4272454165	914	7.5	7
1	YUTK870103	0.1511292337	2.097133307	0.8385107147	1038	17.96	17.93
1	YUTK870104	0.1479523844	2.1305206452	0.4586095904	1104.5	18.36	18.24
1	ZIMJ680101	0.7194866749	0.1298155826	0.0919642414	1220.5	1.6	0.64
1	ZIMJ680102	0.2355617879	1.4264022856	0.8479832682	1036.5	15.71	14.45
1	ZIMJ680105	0.8389999048	0.0415242096	0.0577307075	778.5	5.6	9
1	AURR980103	0.2612230674	1.2786474464	0.5485868086	938	0.96	0.96
1	AURR980106	0.9527833607	0.0035260923	0.410739073	1114.5	0.9	0.81
1	AURR980108	0.1520642932	2.0874595475	0.6152500337	1075	1.1	0.75
1	AURR980110	0.1805561385	1.8218417392	0.1487195953	1191	0.96	0.81
1	AURR980111	0.5274318868	0.4025292635	0.2975880882	1141.5	1.1	1.03
1	AURR980112	0.8012814556	0.0637313451	0.655500456	1068	0.94	0.86
1	AURR980113	0.2786514353	1.1882865145	0.1324081123	1198.5	1.22	0.76
1	AURR980114	0.6281521958	0.2362276304	0.1454636162	1192.5	1.22	1.02
1	AURR980116	0.3265619287	0.9733057964	0.2196604345	1163.5	1.2	1.02
1	AURR980118	0.5560366021	0.3492788328	0.0859319785	1224.5	1.05	0.96
1	ONEK900101	0.1349757623	2.2759601707	0.1188028234	1205	12.2	12
1	VINM940101	0.2416694461	1.3894854338	0.3708265596	1123.5	1.008	1.008
1	VINM940102	0.9462960492	0.0045631817	0.7300111104	1055.5	1.315	1.324
1	VINM940103	0.5973433016	0.2810587423	0.1770798207	1179.5	1.018	1.018
1	MUNV940101	0.7849971565	0.0748826837	0.4168403901	912	0.706	0.802
1	MUNV940102	0.2387524436	1.4069718918	0.3132300477	887.5	0.939	0.969

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	MUNV940103	0.3987992519	0.7189183627	0.6702342491	1065.5	0.987	0.976
1	MUNV940104	0.5256996425	0.4059240694	0.1850536641	1176	0.784	0.784
1	MUNV940105	0.6684392569	0.1846716906	0.1163783973	1206.5	1.33	1.2
1	WIMW960101	0.8527711408	0.0346444088	0.5434321783	937	4.12	4.12
1	KIMC930101	0.1348581939	2.2773497774	0.5730171037	1082.5	-0.41	-0.44
1	MONM990101	0.6064003611	0.2673675357	0.1973078435	1170.5	1.3	0.7
1	BLAM930101	0.7681178624	0.0874679826	0.5221352968	1092	0.63	0.54
1	PARS000101	0.7777651731	0.080145537	0.5063476373	1095	0.353	0.362
1	PARS000102	0.3174547912	1.0108546528	0.4885925042	1098.5	0.34	0.339
1	TAKK010101	0.6277613113	0.2367660902	0.2904143528	1143.5	10.5	6.9
1	FODM020101	0.147709439	2.1331076913	0.6063559341	948.5	0.91	0.91
1	NADH010101	0.8248016687	0.0492992231	0.9243917357	1000.5	-11	-11
1	NADH010104	0.1400535529	2.2172228545	0.1205515644	821	-22	-22
1	NADH010105	0.4237174237	0.6459854572	0.7361253602	1054.5	-9	-28
1	NADH010106	0.0917508382	2.9065077314	0.9210379329	1025	-38	-8
1	NADH010107	0.5269276301	0.4035154237	0.7239860395	1056.5	36	-7
1	MONM990201	0.3799433139	0.7786995705	0.7202677628	1057	1.1	0.9
1	KOEP990101	0.4615863056	0.5468241866	0.0669513564	786.5	-0.06	0.05
1	CEDJ970103	0.1958472634	1.6987808313	0.0989062898	1216	5.6	5.6
1	FUKS010103	0.7051747385	0.1440778363	0.141015541	1194.5	5.42	5.42
1	FUKS010105	0.3160447078	1.0167965827	0.1222296154	1203.5	5.27	4.1
1	FUKS010106	0.1101938509	2.6037167816	0.5300963425	1090.5	4.97	4.93
1	FUKS010108	0.1528834503	2.079040907	0.8637674091	1034	5.62	5
1	FUKS010111	0.4066718074	0.695161333	0.9180422863	1025.5	5.45	6.54
1	TSAJ990101	0.3975759803	0.7226715854	0.2288407442	1161.5	165.1	138.2
1	TSAJ990102	0.3574131026	0.8559199826	0.1303148956	1199.5	167.3	139
1	COSI940101	0.6847971029	0.1658706246	0.8892610371	995	0.0371	0.0823
1	PONP930101	0.5852068937	0.3000995364	0.9951708733	1013.5	0.2	0
1	WILM950101	0.0524362394	3.8656499025	0.0572654645	778.5	0.21	0.25
1	WILM950102	0.2432384344	1.3801871365	0.6067796301	1076.5	1.26	-0.45
1	KUHL950101	0.1815797242	1.8132262598	0.8605891896	990.5	0.69	1
1	BASU050101	0.7380407083	0.1125653938	0.6999044872	964.5	0.0363	0.0363
1	BASU050102	0.8430891597	0.0394137521	0.9052440422	997.5	0.0394	0.0239

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	BASU050103	0.8962876868	0.017090202	0.0525236677	773.5	0.0248	0.0844
1	SUYM030101	0.4956129754	0.4682032722	0.7667551957	1049.5	0	0
1	PUNT030102	0.5646051954	0.3343163687	0.1891335284	1175	0.08	0.03
1	GEOR030101	0.2055540936	1.6264298833	0.7975068537	980.5	0.955	1.014
1	GEOR030102	0.3411912015	0.9158407275	0.8668399753	1033.5	0.946	0.972
1	GEOR030103	0.5088043184	0.4401084324	0.5677987212	941.5	0.959	0.986
1	GEOR030104	0.8319137525	0.0453174467	0.5300455076	934.5	0.981	0.992
1	GEOR030105	0.3541609202	0.8676292207	0.867008838	1033.5	1.131	1.022
1	GEOR030106	0.757357062	0.0960483348	0.213494859	1166.5	1.003	0.988
1	GEOR030107	0.0516053925	3.8937057815	0.6468531131	955.5	0.944	0.994
1	GEOR030108	0.667280536	0.1860488209	0.5353785202	1089.5	1.008	0.919
1	GEOR030109	0.7196176268	0.1296889728	0.8416840678	987.5	0.999	0.978
1	ZHOH040101	0.7320428973	0.11799112	0.9244461243	1000.5	2.18	2.18
1	ZHOH040102	0.406883536	0.6945317585	0.9565465514	1019.5	2.5	2.83
1	ZHOH040103	0.6715036571	0.1810588646	0.0638835937	784.5	8.5	9.9
1	BAEK050101	0.6014541617	0.274790415	0.6295524378	952.5	-0.0442	-0.0701
1	HARY940101	0.4748952875	0.5149715222	0.1303178443	1199.5	170	139.1
1	PONJ960101	0.3398189481	0.9210882149	0.4110256919	1114.5	162.5	138.4
1	GUYH850102	0.9196108636	0.0102442674	0.3935781991	1118	0.12	0.12
1	BLAS910101	0.424074632	0.6449867696	0.962986648	1006.5	0.501	0.45
1	CASG920101	0.1748441929	1.8709942482	0.3375754069	894	-0.1	-0.4
1	CORJ870101	0.2799582894	1.1818011626	0.7059067492	965.5	50.27	49.26
1	CORJ870102	0.5027165937	0.4529167807	0.5301079645	934.5	-0.584	-0.563
1	CORJ870103	0.6907751175	0.1592930726	0.8416816075	1037.5	-0.96	-1.28
1	CORJ870104	0.7382360298	0.112391095	0.7976307636	1044.5	-0.26	-0.4
1	CORJ870105	0.434554379	0.6162451156	0.6295484871	952.5	-1.03	-1.03
1	CORJ870106	0.7999991301	0.064574703	0.5091287547	930.5	-3.89	-3.89
1	CORJ870107	0.6203560517	0.2471115031	0.8733654121	1032.5	-0.56	-1.35
1	CORJ870108	0.8010456903	0.0638859601	0.9565596988	1005.5	1.37	3.39
1	MIYS990101	0.7782904689	0.0797567694	0.392872823	1118.5	0.38	0.38
1	MIYS990102	0.7348365455	0.1154461965	0.4603388941	1104	0.06	0.06
1	MIYS990103	0.7042770476	0.1450007473	0.2822504249	1144.5	0.09	0.09
1	MIYS990104	0.5672783356	0.3297380094	0.2738151774	1148	0.07	0.07

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
1	MIYS990105	0.2332075408	1.4409473739	0.5428195341	1088	-0.02	0.05
1	FASG890101	0.1514770059	2.0935273777	0.1164164404	1206.5	0	0.78
2	ANDN920101	0.1892528045	1.7504095636	0.0688285908	788.5	4.35	4.36
2	ARGP820102	0.8840257084	0.0214012665	0.3811557514	1121	1.08	0.76
2	ARGP820103	0.6707139985	0.1819858046	0.2561168986	1153	1.14	0.76
2	BEGF750103	0.1849832469	1.7849842756	0.0591752236	784	0.53	0.75
2	BHAR880101	0.1614071736	1.9944099245	0.7449576033	972	0.386	0.466
2	BIGC670101	0.5442899757	0.3705183172	0.7600051934	1050.5	102	85.1
2	BIOV880101	0.3982843429	0.7204961466	0.1121785682	1208	16	-13
2	BIOV880102	0.1088508168	2.6237937928	0.1651217151	1184	44	-8
2	BROC820101	0.8270016324	0.0480487768	0.0601862985	1244	5.1	-1.2
2	BULH740101	0.9154017399	0.0113493395	0.6696399411	1065.5	-0.2	-0.39
2	BUNA790101	0.552877737	0.3549067724	0.1076848938	814	8.274	8.391
2	BUNA790102	0.1149563607	2.5346668501	0.1024152096	811	4.349	4.385
2	BUNA790103	0.3030713484	1.0731588826	0.4039959779	1113	6.5	6.5
2	BURA740101	0.68604787	0.1644815876	0.1675948027	1183	0.318	0.318
2	BURA740102	0.4716162813	0.5226837675	0.3185447462	1136	0.362	0.327
2	CHAM810101	0.8619619664	0.0304107966	0.6744853322	960.5	0.68	0.68
2	CHAM820101	0.1004735651	2.7556074925	0.7051717004	1059.5	0.186	0.14
2	CHAM820102	0.1931076077	1.7199760359	0.0551978348	776.5	-0.368	0
2	CHAM830101	0.3581204366	0.8533928187	0.1159294407	819	0.71	1.06
2	CHAM830102	0.9771024782	0.0008284869	0.5483277501	938	0.104	0.094
2	CHAM830103	0.0564724051	3.7358490566	0.329675581	903	1	1
2	CHAM830104	0.5219401567	0.4133611691	0.8502362473	986	1	1
2	CHAM830105	0.823974479	0.0497737557	1	990	0	0
2	CHAM830107	0.2971380524	1.1	0.4298691739	922.5	0	0
2	CHAM830108	0.6590626688	0.1959910913	0.8233134526	967.5	0	0
2	CHOC750101	0.1712855329	1.9025766225	0.8257100067	1040	167.9	141.7
2	CHOC760101	0.26660408	1.2499541929	0.7198886672	1057	170	160
2	CHOC760102	0.8395126371	0.0412564463	0.3607750929	900.5	25	43
2	CHOC760103	0.8366947319	0.0427391938	0.3447994901	1129.5	0.38	0.23
2	CHOC760104	0.6299811458	0.2337182448	0.0546821042	1248	0.16	0.08
2	CHOP780201	0.3947919782	0.731276528	0.1785207422	1178.5	1.06	1.01

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	CHOP780202	0.4137400749	0.6744026965	0.0961689819	1217	0.93	0.83
2	CHOP780203	0.1252846076	2.3953608386	0.0793030248	796.5	0.74	1.14
2	CHOP780204	0.7071346017	0.1420746267	0.1101454922	815.5	0.61	0.74
2	CHOP780207	0.850212499	0.0358738105	0.2199277575	861	0.87	0.96
2	CHOP780208	0.1115953271	2.5830556625	0.3167693856	1136	0.9	0.9
2	CHOP780209	0.4914702116	0.4772912195	0.3774655418	1121.5	0.9	0.79
2	CHOP780211	0.2353774385	1.4275347964	0.1902622456	851	0.74	0.96
2	CHOP780213	0.2526263608	1.3260603839	0.1831928427	848	0.076	0.085
2	CHOP780214	0.4692342956	0.5283411976	0.0877583225	802	0.036	0.072
2	CHOP780216	0.1739006098	1.8792947376	0.0749409012	793	0.64	1.08
2	CIDH920101	0.1568307092	2.0391895467	0.6551847066	1068	-0.24	-0.36
2	CIDH920102	0.0749110712	3.2485337308	0.1989956604	1171	-0.08	-0.09
2	CIDH920103	0.8167995511	0.053990209	0.0690786762	1236.5	0.36	-0.56
2	CIDH920104	0.9512424741	0.0037602881	0.1112782349	1209	0.17	-0.57
2	CIDH920105	0.2507872323	1.3364653645	0.079042168	1229	0.02	-0.41
2	COHE430101	0.3423178708	0.9115537849	0.2432573642	1156.5	0.75	0.71
2	CRAJ730101	0.7153609129	0.1338404129	0.6727784801	1065	0.96	0.96
2	CRAJ730102	0.2838485115	1.1627224646	0.4931431834	1097.5	1	0.89
2	DAWD720101	0.3857452352	0.7598572624	0.7654905328	1049.5	5.5	5.5
2	DAYM780101	0.847670191	0.0371173777	0.0939640797	1219	7.4	6.6
2	DAYM780201	0.0568871642	3.7230815197	0.6434254452	955	65	56
2	DESM900102	0.8448327546	0.0385312852	0.2144717377	1166	1.26	1.01
2	EISD840101	0.2249989636	1.4931014503	0.3510691522	1128	0.25	0.02
2	EISD860101	0.1564470969	2.0430114939	0.8476122684	988.5	0.67	0.01
2	EISD860103	0.2138546685	1.5677297123	0.3342527476	1131	0	0
2	FASG760101	0.0583890523	3.6776763972	0.8858034443	994.5	131.17	131.17
2	FASG760103	0.4117345736	0.6802387914	0.0736036798	1233	1.8	-5.6
2	FASG760104	0.4500885635	0.5755512593	0.5614910598	1084.5	9.62	9.21
2	FASG760105	0.2377922866	1.4127853197	0.7079378105	1059	2.34	2.19
2	FAUJ830101	0.5810227381	0.3068528786	0.7327748804	1055	0.31	0
2	FAUJ880101	0.9346282661	0.0067663991	0.9275547482	1001	2.34	1.89
2	FAUJ880102	0.8368481974	0.0426577392	0.7352679443	970.5	0.69	0.69
2	FAUJ880103	0.090988138	2.9204508931	0.7172432148	1057.5	4	3

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	FAUJ880105	0.6985499958	0.1509685023	0.7356622459	1051.5	1.52	1.52
2	FAUJ880107	0.888290092	0.0198459713	0.363672749	900	10.1	10.9
2	FAUJ880109	0.3258068617	0.9763655462	0.3708187207	915.5	0	0
2	FAUJ880110	0.3717026431	0.8061780739	0.1976317713	875	0	0
2	FAUJ880111	0.1820591526	1.8092105263	0.2813085616	1125	0	0
2	FAUJ880112	0.1560967345	2.0465116279	0.4943820225	967.5	0	0
2	FAUJ880113	0.2788582692	1.1872575043	0.0634413522	1241	4.31	4.27
2	FINA910102	0.5726310289	0.3206960953	0.1099520185	841	1	1
2	FINA910103	0.400488079	0.7137642956	0.4081425256	1112.5	1	1
2	FINA910104	0.3687527788	0.8162247443	0.0791844082	1205	1	1
2	GARJ730101	0.794579897	0.0682041728	0.850975908	989	0.28	0.25
2	GEIM800105	0.2938898428	1.1149912577	0.6906245257	1062	1.04	1.02
2	GEIM800106	0.1553140735	2.0543632658	0.5256364469	1090.5	1.15	1.01
2	GEIM800107	0.5751349076	0.3165234765	0.0942751618	1217	0.99	0.93
2	GEIM800108	0.0845066182	3.0444104658	0.4035893918	909	0.91	0.94
2	GEIM800109	0.0860235662	3.0144820456	0.0657617624	789.5	0.85	1.2
2	GEIM800110	0.0589522884	3.6609733509	0.3594978616	899	0.93	0.94
2	GEIM800111	0.1284475434	2.3552744751	0.394521985	907	0.93	0.98
2	GOLD730101	0.0822174406	3.0907111038	0.3723291754	1121.5	0.75	0.75
2	GOLD730102	0.6069372711	0.2665695472	0.7600051934	1050.5	168.5	141.4
2	GRAR740102	0.406913344	0.6944431635	0.2406163485	867.5	8.1	9
2	GRAR740103	0.7018227139	0.147541302	0.7477078339	1052.5	111	84
2	GUYH850101	0.250330417	1.339064545	0.8320238372	986	0.1	0.33
2	HOPT810101	0.4554565674	0.5619961207	0.6012973153	948	-0.5	0
2	ISOY800101	0.0744201149	3.2597369538	0.0924888993	1220	1.17	1
2	ISOY800102	0.1335568853	2.2928232074	0.7725913332	976.5	0.98	1.01
2	ISOY800103	0.4395704741	0.6028609732	0.1994827331	854	0.78	1.01
2	ISOY800104	0.2791054908	1.1860288405	0.3294421948	892	0.97	1.09
2	ISOY800106	0.248946155	1.34697682	0.48230687	925.5	1.07	1.09
2	JANJ780101	0.9678712288	0.001631603	0.3811711592	904	27.8	42
2	JANJ780102	0.62837496	0.235921103	0.4011239014	1116.5	51	35
2	JANJ780103	0.3866990341	0.7567985739	0.2716952765	877	15	32
2	JANJ790101	0.4615399057	0.5469378234	0.4173906447	1113	1.7	0.8

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	JANJ790102	0.9283251007	0.0081379004	0.2989519255	1139.5	0.3	-0.1
2	JOND920101	0.4688827801	0.5291800275	0.1422546776	1193.5	0.074	0.066
2	JOND920102	0.1601178657	2.0068755466	0.9983867531	1013	83	72
2	JUKT750101	0.6335253568	0.2289024671	0.1040764337	1213	4.7	4.2
2	JUNJ780101	0.9568978777	0.0029377494	0.4704393629	1102	581	575
2	KANM800101	0.2255338557	1.489632773	0.2427370109	1157	1	1
2	KANM800102	0.8794435255	0.0231393291	0.8131486088	983	0.88	0.92
2	KANM800104	0.9202472961	0.0100821537	0.556003217	1085.5	0.83	0.83
2	KLEP840101	0.2971380524	1.1	0.0768269492	1181.5	0	0
2	KRIW710101	0.7897840874	0.0715047193	0.5375761247	936	4.6	5.25
2	KRIW790101	0.7799865861	0.0785084531	0.6787363695	961	4.32	5.37
2	KRIW790102	0.1020412352	2.7300311872	0.9334023571	1023	0.28	0.27
2	KRIW790103	0.6944432041	0.1553334495	0.8097704009	1042.5	93.5	71.5
2	KYTJ820101	0.3406150922	0.9180402623	0.3385622579	1131	1.8	-0.7
2	LAWE840101	0.4550415554	0.563035072	0.422108872	913	0	0.05
2	LEVM760101	0.5197136533	0.4178108559	0.7904936353	979.5	-0.5	0
2	LEVM760104	0.7342923248	0.1159395327	0.50599377	1095	217.9	215
2	LEVM760106	0.4759638343	0.5124770696	0.7251743736	1056	6	6
2	LEVM760107	0.7020476096	0.1473074512	0.6741688711	1063.5	0.19	0.1
2	LEVM780101	0.2062843203	1.621152671	0.3385057658	1131	0.96	0.96
2	LEVM780102	0.3652166449	0.8284181013	0.214481793	1166	0.99	0.95
2	LEVM780103	0.1120642248	2.5762077101	0.0517435987	773	0.77	0.98
2	LEVM780104	0.2454615632	1.3671383052	0.1966334948	1172	0.98	0.95
2	LEVM780105	0.2814607816	1.174392562	0.6148655829	1075	0.97	1
2	LIFS790101	0.3007020047	1.0837945374	0.142341062	1193.5	0.93	0.82
2	LIFS790102	0.9025390925	0.0150818313	0.0948354148	1218.5	1	0.79
2	LIFS790103	0.0830380121	3.073953272	0.2298694059	1161	1.02	0.87
2	MAXF760102	0.149748001	2.1115495361	0.3638080416	900	0.97	0.98
2	MAXF760104	0.6780190885	0.1735170709	0.1014932392	810.5	0.26	0.57
2	MAXF760105	0.9106705225	0.0126598588	0.998381586	1013	0.65	0.81
2	MAXF760106	0.0837140107	3.0602840036	0.597868676	947	1	1.01
2	MCMT640101	0.2208467018	1.5203733798	0.8827599035	1031	18.78	13.92
2	MEEJ800101	0.3328832684	0.9480569956	0.6230816731	1073.5	0.8	1.2

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	MEEJ800102	0.7463883818	0.1052497076	0.850982167	1036	-0.1	-0.5
2	MEEJ810101	0.360745645	0.8440737731	0.0818938501	1227	1.1	-0.2
2	MEEJ810102	0.5165526369	0.424186495	0.1598670118	1186	1	0.2
2	MEIH800102	0.806223981	0.0605358436	0.169823242	843	0.94	1.01
2	NAGK730101	0.4416282624	0.5974383824	0.2061943967	1167.5	0.97	0.97
2	NAGK730102	0.8435537592	0.0391775943	0.7974381797	1044.5	0.96	0.9
2	NAGK730103	0.8894644969	0.0194281547	0.1881848988	850	0.83	0.84
2	NAKH900101	0.2610428485	1.2796212622	0.0749401528	1232	6.91	6.65
2	NAKH900103	0.7784454948	0.079642231	0.8068821412	982	5.74	5.66
2	NAKH900104	0.9286317338	0.0080682304	0.0812383087	798	-0.6	0.08
2	NAKH900105	0.7582732596	0.0953006644	0.7415474457	971.5	5.88	5.29
2	NAKH900106	0.9746055803	0.0010190915	0.0833456734	799	-0.57	0.1
2	NAKH900107	0.616730198	0.2522779355	0.3304845374	892	5.44	6.34
2	NAKH900110	0.2252006409	1.491792433	0.5190362651	1092.5	0.34	0.07
2	NAKH900111	0.4528593712	0.5685227966	0.0648137647	1240	9.6	5.38
2	NAKH900112	0.7627599136	0.0916854349	0.9468511392	1004	6.61	4.88
2	NAKH900113	0.851863491	0.0350779851	0.3674733196	1124	1.37	1.24
2	NAKH920101	0.1088726431	2.6234653039	0.0990191125	1216	6.8	6.75
2	NAKH920102	0.4721061678	0.5215260007	0.1461211017	1192	7.72	7.21
2	NAKH920103	0.1525661971	2.0822952125	0.2527291017	871	5.15	6.38
2	NAKH920104	0.9817263151	0.0005276147	0.1351169712	828	5.04	6.31
2	NAKH920105	0.7832741898	0.0761190487	0.1858792104	1176	8.14	4.17
2	NAKH920107	0.4713225888	0.5233788011	0.7757468591	977	5.08	5.96
2	NAKH920108	0.50545928	0.447113302	0.3345320254	1132	9.36	5.58
2	NOZY710101	0.934901269	0.0067098742	0.383349291	1113.5	0.5	0
2	OOBM770101	0.3831118693	0.7683587415	0.330493687	892	-1.895	-1.767
2	OOBM770104	0.5788394788	0.310415839	0.9854947272	1015	-9.475	-11.893
2	OOBM770105	0.8012876161	0.0637273077	0.5537063219	1086	-7.02	-8.652
2	OOBM850101	0.5015166566	0.4554729762	0.5537169872	1086	2.01	1.47
2	OOBM850103	0.407562866	0.6925150376	0.4804748457	1100	0.46	0.43
2	OOBM850104	0.1221555028	2.4361587909	0.9919418393	1011	-2.49	-1.6
2	PALJ810101	0.4062082859	0.6965413037	0.34834557	1127.5	0.95	0.95
2	PALJ810102	0.0872134984	2.9914090743	0.1858758207	1176	1.05	1.04

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	PALJ810103	0.5600992347	0.3421297668	0.2421443723	1155.5	1.03	1.02
2	PALJ810104	0.788836369	0.0721668474	0.3077430079	1138.5	0.9	0.83
2	PALJ810105	0.102785493	2.7180412086	0.0613173358	782	0.88	1
2	PALJ810107	0.7423693138	0.1087378641	0.5007465503	1096	0.98	0.96
2	PALJ810108	0.433719011	0.6184972013	0.6319690679	1072	1.02	1.05
2	PALJ810109	0.0515601474	3.8952475528	0.5059668604	1095	1.06	0.92
2	PALJ810110	0.3135444354	1.0274192326	0.0760836854	1231	1.02	1
2	PALJ810111	0.1260664281	2.3853466737	0.3296132375	1132	0.94	0.94
2	PALJ810114	0.7767680595	0.0808863054	0.5164207666	1093	0.87	0.91
2	PLIV810101	0.8579453158	0.0322259677	0.3497271771	1128	-2.89	-3.15
2	PONP800102	0.4177043163	0.6629898917	0.1416483447	1194	7.62	7.31
2	PONP800103	0.384505169	0.7638502504	0.3804057453	1121	2.63	2.6
2	PONP800104	0.2615905817	1.2766641869	0.2807058307	1146	13.65	12.88
2	PONP800105	0.4827582375	0.4968288293	0.4134310907	1113.5	14.18	14.18
2	PONP800106	0.6285503419	0.2356799476	0.87639826	993	11.05	11.18
2	PONP800108	0.1306947798	2.3274659965	0.0722934195	1234	6.05	6.05
2	PRAM820102	0.9179055869	0.0106850813	0.686709806	1062.5	0.104	0.104
2	PRAM900101	0.3101355314	1.0420838939	0.4410931186	917	-6.7	-4.2
2	PRAM900102	0.2062843203	1.621152671	0.3385057658	1131	0.96	0.96
2	PRAM900103	0.3652166449	0.8284181013	0.214481793	1166	0.99	0.95
2	PRAM900104	0.1018170532	2.7336618216	0.0547856406	776	0.78	0.97
2	PTIO830101	0.2339828819	1.4361373391	0.2349635846	1158	0.95	0.95
2	PTIO830102	0.8533677186	0.0343609486	0.5799575633	1081	1	0.9
2	QIAN880102	0.2602872232	1.2837134978	0.8572298226	990	-0.02	-0.02
2	QIAN880104	0.7452033215	0.1062716764	0.0749424061	1232	-0.01	-0.17
2	QIAN880105	0.0726083076	3.3017909122	0.1483877183	1191	0.06	-0.08
2	QIAN880108	0.1917494526	1.7306159315	0.2681625689	1149.5	0.24	0.1
2	QIAN880111	0.1159050368	2.5212952324	0.1288299546	1200	0.07	-0.04
2	QIAN880112	0.1975374938	1.6858792797	0.0963879743	1217.5	0.18	-0.03
2	QIAN880116	0.5244343617	0.4084164384	0.1625935388	840	-0.15	0.02
2	QIAN880117	0.9576384068	0.002837575	0.6610127166	1067	0.14	0.02
2	QIAN880118	0.6365408477	0.2248535905	0.2801620144	1146	0.25	0.11
2	QIAN880119	0.1088217898	2.6242307679	0.059064394	1245	0.19	-0.12

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	QIAN880120	0.1388608638	2.2307988967	0.1247677216	1202	-0.02	-0.31
2	QIAN880121	0.3967469788	0.7252247086	0.1024021235	1214	-0.09	-0.28
2	QIAN880123	0.0905135871	2.9291915862	0.1178513677	820	-0.13	-0.12
2	QIAN880128	0.087511391	2.9856871256	0.1226049481	1203	0.17	0.05
2	QIAN880129	0.8610533786	0.0308166491	0.3213328924	890	-0.1	-0.1
2	QIAN880131	0.3737852112	0.7991527473	0.0505387346	772	-0.25	0.09
2	QIAN880134	0.2314634236	1.4518393764	0.0818945051	798	-0.24	0.13
2	QIAN880135	0.3879541202	0.7527902013	0.4545906003	920	-0.14	0.06
2	QIAN880136	0.1144958941	2.5412018204	0.5318885405	935	-0.11	-0.09
2	QIAN880137	0.1971940996	1.6884897661	0.0912953205	805	-0.12	-0.09
2	QIAN880138	0.1015060216	2.7387137302	0.3068501557	886	-0.2	-0.02
2	RACS770102	0.9347526671	0.0067406124	0.3767838156	903	0.941	1.055
2	RACS820101	0.0728734736	3.2955650118	0.9854594667	1010	1.03	1.34
2	RACS820102	0.8696835583	0.027073353	0.2298568697	1161	1.14	1.05
2	RACS820103	0.9953877418	3.36062112e-05	0.7733497694	1048	1.62	1.23
2	RACS820105	0.183431548	1.7977835968	0.2843597031	880	0.99	1.09
2	RACS820107	0.7805075105	0.0781271938	0.5482987794	938	1.2	0.92
2	RACS820108	0.602762609	0.2728141949	0.277095472	1147	1.02	1.02
2	RACS820109	0.4740384715	0.5169783499	0.2595257504	895	0	0
2	RACS820110	0.3686197615	0.8166804408	0.1076750033	814	1.02	1.06
2	RACS820111	0.0917526506	2.9064747511	0.1530355427	1189	1.03	1
2	RACS820112	0.1492677299	2.1165980406	0.1449879729	832.5	1.19	1.25
2	RACS820113	0.5360301905	0.385972283	0.8383153016	987	17.06	19.95
2	RACS820114	0.5015925853	0.4553109166	0.4315223191	915	17.82	19.61
2	RADA880101	0.1816295109	1.8128086632	0.3441465055	1129.5	1.81	0
2	RADA880102	0.0716414866	3.3247031172	0.7475206077	972.5	0.52	0.08
2	RADA880103	0.1914089084	1.7332977814	0.70855064	1059	-2.64	-2.64
2	RADA880104	0.2158293365	1.5541680841	0.2443708024	1156.5	0.94	0
2	RADA880105	0.3169115323	1.0131397328	0.2788735885	1146.5	0.52	-2.85
2	RADA880106	0.2561701689	1.3062746767	0.5922557342	1079	173.7	157.2
2	RADA880107	0.2991154859	1.0909774776	0.4728019141	1101.5	-0.29	-0.34
2	RADA880108	0.9416158599	0.0053946928	0.5950505708	1078.5	-0.06	-0.41
2	RICJ880104	0.5659486517	0.3320101294	0.251882789	1153	0.9	0.7

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	RICJ880105	0.7229938227	0.1264487935	0.2021650033	1167	0.9	0.8
2	RICJ880106	0.2142562117	1.5649597969	0.2513643569	875	0.6	0.6
2	RICJ880108	0.0818777978	3.0977011494	0.2532154647	1153	1.1	1.1
2	RICJ880110	0.3847162824	0.7631691649	0.378226149	1121	1	1
2	RICJ880111	0.2902275191	1.1321531494	0.0972708528	1215.5	1.3	0.8
2	RICJ880113	0.4113982502	0.681221651	0.474521538	1101	0.8	0.8
2	RICJ880115	0.7571185752	0.0962434796	0.8081737142	982.5	0.8	0.8
2	RICJ880116	0.7451470724	0.1063203202	0.1459971599	1191	1	0.9
2	RICJ880117	0.1511287223	2.0971386163	0.5771183031	943.5	0.7	0.9
2	ROBB760101	0.6820820607	0.1689093145	0.1553855685	1188	1.4	0.6
2	ROBB760102	0.6580549402	0.197231605	0.3545533132	899	-4	-3.5
2	ROBB760104	0.1313147133	2.3198898382	0.407173019	1115	1.4	0.1
2	ROBB760105	0.5675548743	0.3292667746	0.4057107354	1115.5	0.4	-1.7
2	ROBB760106	0.6470583627	0.2110759139	0.2281202905	1161.5	0.4	-2.4
2	ROBB760108	0.2074390844	1.6128529051	0.0892471019	803	-1.8	1
2	ROBB760110	0.7503165147	0.101901161	0.0629833129	784	-3.7	2.1
2	ROBB760111	0.382143756	0.7715052489	0.0940507872	806	-0.5	1.3
2	ROBB760113	0.3525291655	0.873560405	0.137181064	829	-2.4	1
2	ROBB790101	0.3497876303	0.8836111891	0.4764607623	1100	0.3	0.3
2	ROSG850101	0.6171190189	0.2517207028	0.5269282157	1091	158	115.5
2	ROSG850102	0.1446429351	2.1661871721	0.16062785	1185.5	0.74	0.72
2	ROSM880101	0.4851972214	0.4913001364	0.5028624369	929.5	-0.67	0
2	ROSM880102	0.2130178164	1.5735226523	0.4874805935	926.5	-0.67	0
2	SIMZ760101	0.0659462503	3.466896783	0.4977619848	1096	0.73	0.73
2	SNEP660101	0.8388904486	0.0415814876	0.7757372563	1048	0.234	0.234
2	SNEP660102	0.925735936	0.0087382286	0.1838564739	1176.5	0.149	0.022
2	SNEP660103	0.4002997278	0.7143375404	0.2992537551	1141	-0.008	-0.067
2	SUEM840102	0.1667046893	1.9443753941	0.4599532514	925.5	8	1
2	SWER830101	0.6041237366	0.2707680488	0.190419377	1174	-0.4	-0.55
2	TANS770101	0.1866748074	1.771174581	0.0684219611	1237	1.06	1.01
2	TANS770103	0.5286698731	0.4001153511	0.8700266586	1033	1.087	1.052
2	TANS770104	0.2943493434	1.1128575425	0.1969681779	853	0.795	1.015
2	TANS770106	0.1308970427	2.3249896903	0.9083381405	1027	0.901	0.901

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	TANS770107	0.6477360152	0.2102063396	0.5971335865	1078	0.289	0.393
2	TANS770108	0.2498280015	1.3419299955	0.6319726603	1072	0.5	0.835
2	VASM830103	0.3750419384	0.7949400419	0.7445867128	972	0.159	0.159
2	VELV850101	0.7969893979	0.0665773619	0.4673401375	1102.5	0.03731	0.0371
2	VENT840101	0.823974479	0.0497737557	1	1035	0	0
2	VHEG790101	0.0760083238	3.2237827637	0.3680793159	901	-12.04	-4.15
2	WEBA780101	0.0826060315	3.0827524576	0.4166457731	912	0.88	0.88
2	WERD780101	0.1999380617	1.6677796655	0.0708481687	1234.5	0.52	0.49
2	WERD780102	0.1472973286	2.1375072929	0.6787301747	1064	-0.2	-0.18
2	WERD780103	0.473680861	0.5178176975	0.304921762	885.5	0.06	0.13
2	WERD780104	0.1898978784	1.7452659046	0.1775886088	1179	-0.07	-0.11
2	WOEC730101	0.2131130133	1.5728622909	0.47481636	924	7	7.5
2	WOLR810101	0.4266152782	0.637920184	0.5163984795	1093	1.94	-3.68
2	WOLS870101	0.9148552823	0.0114970053	0.7634022698	975	0.07	1.96
2	WOLS870102	0.7247683852	0.1247642438	0.7145808657	967	-1.68	-1.03
2	ZIMJ680101	0.0652607202	3.4849083202	0.4875048882	1098	0.83	0.83
2	ZIMJ680102	0.7707574636	0.0854299549	0.83494506	1038.5	14.28	14.45
2	ZIMJ680103	0.4476027978	0.5819148583	0.1472119906	836	0.13	1.58
2	ZIMJ680105	0.69339876	0.1564550277	0.7819570557	1047	9.9	8.2
2	AURR980102	0.2863062258	1.1508410327	0.972576165	1008	0.98	1.05
2	AURR980103	0.3876262008	0.753835675	0.3143568802	888.5	0.96	0.96
2	AURR980106	0.4183881632	0.6610375325	0.3204566118	1135.5	1.05	0.84
2	AURR980107	0.4314354555	0.6246875777	0.1094836582	1210	0.95	0.8
2	AURR980116	0.6245833668	0.2411721001	0.4797428222	1100	1.19	1.19
2	AURR980117	0.8167242374	0.0540354263	0.819132016	1041	0.77	0.92
2	AURR980120	0.520755525	0.4157244598	0.9147291065	1026	1.05	0.95
2	VINM940103	0.4335102816	0.6190609593	0.399078635	908	0.994	1.018
2	MUNV940103	0.6328920271	0.2297584855	0.465495139	922	0.976	0.987
2	MUNV940104	0.7114107848	0.1377593914	0.9531642162	1020	0.784	0.784
2	MUNV940105	0.4856529869	0.4902721304	0.9661383227	1007	1.33	1.33
2	WIMW960101	0.1745626626	1.8734653175	0.8509395792	1036	4.11	4.12
2	KIMC930101	0.8881581515	0.0198931949	0.8287085989	985.5	-0.44	-0.41
2	MONM990101	0.943434423	0.0050632911	0.6283747783	953	0.5	0.7

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	KUMS000102	0.3943138314	0.7327632979	0.2394398331	1158	7.7	6
2	TAKK010101	0.0833192419	3.0682517674	0.5922386657	1079	9.8	9.8
2	FODM020101	0.3771740707	0.7878384622	0.6493349879	956	0.95	0.91
2	NADH010101	0.0853825094	3.0270582997	0.3896141038	1119	58	-11
2	NADH010102	0.73661567	0.1138416077	0.2714268057	1148.5	51	-13
2	NADH010103	0.778100239	0.0798974392	0.2843708607	1145	41	-18
2	NADH010104	0.8142229985	0.0555484855	0.3221557977	1135	32	-22
2	NADH010105	0.5118942112	0.4337085414	0.637777573	1071	24	-9
2	NADH010106	0.9395815468	0.0057779291	0.8731182325	1032.5	5	-38
2	MONM990201	0.41890344	0.6595695966	0.3630951475	900.5	0.4	0.9
2	CEDJ970101	0.619364462	0.2485177641	0.2543718425	1153.5	7.8	6.7
2	CEDJ970102	0.4562738486	0.5599545037	0.3068512956	1139	6.9	6.7
2	CEDJ970103	0.3377199954	0.9291705565	0.0541412309	1249.5	7	7
2	CEDJ970104	0.9768756969	0.0008449839	0.159986992	1186	7.1	6.6
2	CEDJ970105	0.1494887434	2.1142724833	0.0537339937	1250	7.4	6.3
2	FUKS010101	0.2203941372	1.5233833626	0.1351258592	1197	5.06	4.47
2	FUKS010102	0.0806012248	3.1242612809	0.6377748041	1071	6.77	5.41
2	FUKS010108	0.6695635048	0.1833413283	0.0915906539	1220.5	7.32	5.62
2	FUKS010111	0.2396774237	1.4013985825	0.8891417663	1030	6.54	6.54
2	FUKS010112	0.3220193067	0.9918581423	0.5756124641	1082	5.72	5.72
2	MITS020101	0.6507640947	0.2063471811	0.8845751769	1001.5	0	0
2	TSAJ990101	0.2550759095	1.3123475008	0.8509704379	1036	163	138.2
2	TSAJ990102	0.2212325049	1.5178133197	0.6668911267	1066	163.9	139
2	COSI940101	0.79809222	0.0658397645	0.4673401375	1102.5	0.0373	0.0371
2	WILM950101	0.9924785566	8.93722866e-05	0.7911974982	979.5	0.21	0.21
2	WILM950104	0.3627408642	0.8370540141	0.5867040276	945	1.09	1.09
2	KUHL950101	0.3578187803	0.8544697287	0.7787398506	977.5	0.69	0.78
2	GUOD860101	0.8927550377	0.0182816063	0.5806391955	1080.5	25	0
2	JURD980101	0.2059123448	1.6238380693	0.5700962644	1083	1.1	-0.64
2	BASU050101	0.7889303728	0.0721010249	0.0601891338	1244	0.1366	0.0019
2	BASU050102	0.6966653897	0.1529626597	0.126781376	1201	0.0728	-0.0053
2	BASU050103	0.8668954719	0.0282554135	0.355271629	1127	0.151	0.0381
2	PUNT030101	0.3823158593	0.7709450551	0.4788019653	925	-0.17	0.01

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
2	PUNT030102	0.407542208	0.6925762914	0.4754344656	924	-0.15	0.08
2	GEOR030103	0.5661047066	0.3317429289	0.3680618617	1124	1	0.956
2	GEOR030104	0.6344897069	0.2276028106	0.142737823	1193.5	1.042	0.984
2	GEOR030106	0.9884733492	0.0002099058	0.0833454884	1226	0.99	0.999
2	GEOR030107	0.2408264553	1.3945120577	0.882770154	994	0.994	0.986
2	GEOR030109	0.9743047834	0.0010433853	0.7695701855	976	0.978	0.978
2	ZHOH040101	0.6829546409	0.1679292009	0.1074326669	1211	2.71	2.12
2	ZHOH040102	0.7939222127	0.068651865	0.8541281751	1035.5	3.2	2.5
2	ZHOH040103	0.9100352002	0.0128413604	0.100383517	1215	13.4	8.5
2	BAEK050101	0.1483654175	2.1261335521	0.0601822192	1244	0.0166	-0.0442
2	HARY940101	0.3992272181	0.7176092524	0.8194214135	1041	164.6	139.1
2	PONJ960101	0.2184905161	1.5361266865	0.6493430299	1069	162.6	138.4
2	DIGM050101	0.3597648925	0.8475442092	0.0537557702	1250	1.076	1.076
2	WOLR790101	0.4204952807	0.6550518032	0.5163933761	1093	1.12	0.54
2	OLSK800101	0.8401765232	0.0409110864	0.3856050095	1120	1.38	0.89
2	KIDA850101	0.7579943352	0.0955279437	0.8257178128	985	-0.27	-0.16
2	GUYH850102	0.2019289711	1.6529640558	0.0567305919	778.5	0.05	0.12
2	GUYH850104	0.7701005442	0.0859347195	0.4174429735	912	-0.31	0.1
2	GUYH850105	0.2331410189	1.4413609724	0.372419646	902	-0.27	0.17
2	JACR890101	0.0836893016	3.0607815048	0.460546701	1104	0.18	-0.08
2	COWR900101	0.8651094059	0.0290263048	0.3637795653	1125	0.42	0
2	BLAS910101	0.5640167557	0.3353298735	0.741586994	1053.5	0.616	0.501
2	CORJ870102	0.5994833966	0.2777841878	0.2330478379	1160	-0.414	-0.563
2	MIYS990105	0.067387755	3.429677797	0.0736747816	794.5	-0.02	-0.02
2	ENGD860101	0.3103233695	1.0412703178	0.4410931186	917	-1.6	-1
2	FASG890101	0.6583493382	0.1968687161	0.3595074616	899	-0.21	0
3	ANDN920101	0.2213469417	1.5170549959	0.0650316779	785	4.38	4.5
3	ARGP820102	0.1615475708	1.9930594604	0.6848993646	1063	0.72	0.72
3	ARGP820103	0.3998102619	0.7158290727	0.2884532634	1144	0.62	0.68
3	BEGF750101	0.0608310344	3.606481812	0.1523803339	1187.5	0.6	0.44
3	BEGF750102	0.1129647495	2.5631456008	0.0888503884	1219.5	0.72	0.72
3	BEGF750103	0.3742563865	0.7975709621	0.3999398111	910.5	0.75	0.75
3	BHAR880101	0.2393672533	1.4032645071	0.1945332405	1173	0.466	0.463

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	BIOV880101	0.0536226848	3.8264058364	0.1046935038	1212	-13	-38
3	BIOV880102	0.1218336054	2.4404224659	0.2565421209	1153	-8	-54
3	BROC820101	0.0870170738	2.9951938074	0.0560512312	1248	-1.2	-1.2
3	BULH740101	0.0685858459	3.3993966509	0.2698313135	876	-0.35	-0.3
3	BUNA790101	0.2501213284	1.3402561814	0.129168629	825	8.274	8.391
3	BUNA790103	0.4002395868	0.714520662	0.7666865063	976	6.9	6.9
3	CHAM810101	0.9625102634	0.0022219662	0.7570335271	1050.5	0.68	0.7
3	CHAM820102	0.3696975191	0.8129948523	0.0621424396	783	0	0.656
3	CHAM830101	0.083623041	3.0621164117	0.0601520026	781	0.99	1.07
3	CHAM830103	0.2495512525	1.3435114504	0.540342047	1056	1	1
3	CHAM830108	0.2893380216	1.1363636364	0.1371299315	1192.5	1	0
3	CHOP780203	0.3658227042	0.8263165445	0.1133962831	817	1.01	0.96
3	CHOP780207	0.0575261035	3.7036085632	0.2561628249	1153	1.04	0.93
3	CHOP780209	0.6045161485	0.2701799674	0.1138160433	1207.5	0.9	0.9
3	CHOP780211	0.6610951044	0.1935034042	0.0699174458	789	0.82	0.96
3	CHOP780213	0.280180158	1.1807039629	0.5814600079	944	0.085	0.082
3	CHOP780215	0.3878536735	0.7531103106	0.3599646285	1126	0.085	0.079
3	CHOP780216	0.2161345019	1.5520856224	0.10621287	813	1.05	0.92
3	CIDH920101	0.8264728938	0.0483477705	0.2002115006	1171	-0.24	-0.2
3	CRAJ730102	0.8732322128	0.0256063368	0.2398913226	1158	1.18	0.89
3	DAYM780101	0.1545603855	2.0619674671	0.559515631	1085	4.9	5.5
3	DESM900101	0.2474338079	1.3556838718	0.76361294	1050	0.96	1
3	DESM900102	0.1473791847	2.1366322928	0.8637875585	1034	1.06	0.98
3	FASG760102	0.4909399705	0.4784637396	0.6205146526	1074	282	270
3	FASG760103	0.2388982338	1.4060916956	0.1026942374	1214	0	-5.6
3	FASG760104	0.107161518	2.6494435306	0.3690050997	901.5	9.18	9.21
3	FASG760105	0.1120402144	2.5765575863	0.9498945069	1020.5	2.16	2.16
3	FAUJ880101	0.4264922573	0.6382608842	0.0709072365	1235	2.34	1.6
3	FAUJ880105	0.7149849642	0.1342106308	0.7460191672	972	1.52	1.52
3	FAUJ880107	0.2296808733	1.4630755607	0.2596882702	1152	11.1	10.4
3	FAUJ880110	0.5779890376	0.3118110236	0.2263873115	871	0	2
3	GEIM800108	0.0513135203	3.9036773	0.0549999704	776	0.82	0.97
3	GRAR740102	0.6571154705	0.1983923208	0.0769235663	795	9	9.2

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	HOPA770101	0.1009323802	2.7480763466	0.9114858247	998.5	1.7	1.7
3	HUTJ700101	0.3107246678	1.0395343594	0.4414854802	917	45	41.84
3	ISOY800103	0.0901426704	2.9360587422	0.1134166227	817	1.01	0.97
3	ISOY800104	0.6592400049	0.1957732756	0.2647609729	874.5	0.97	0.92
3	ISOY800106	0.3951728758	0.7300940207	0.7022721825	1060	1.07	1.07
3	ISOY800107	0.054911492	3.7848230817	0.1374348824	1196	1.05	0.92
3	JOND920101	0.496518217	0.466234548	0.4755535043	1101	0.051	0.052
3	JUKT750101	0.635008929	0.2269049377	0.9019608235	997	2.6	3.3
3	JUNJ780101	0.5481992715	0.3633547689	0.7574613867	974	382	400
3	KARP850101	0.3857427343	0.7598652966	0.0865875037	801	1.038	1.041
3	KARP850102	0.5145275761	0.4283072311	0.8701502124	992	1.028	1.006
3	KARP850103	0.9461758545	0.0045836618	0.0957668086	807	0.914	0.923
3	KRIW710101	0.1383935465	2.2361545112	0.2302810903	1161	6.1	5.25
3	KRIW790101	0.1181182215	2.4905720305	0.3229058986	1135	6.09	5.37
3	MANP780101	0.2741696329	1.2108256184	0.1756538851	1180	12.43	11.89
3	MAXF760104	0.4373195958	0.6088375645	0.96292042	1018.5	0.76	0.57
3	MAXF760105	0.5398457121	0.3787792841	0.3071763623	886	0.48	0.81
3	MAXF760106	0.2246267111	1.495521362	0.5648877364	1084	1.05	1.01
3	MEIH800101	0.1721154673	1.895143086	0.1796374175	847.5	0.98	0.98
3	MEIH800103	0.2388050131	1.4066544317	0.1318778401	1198.5	90	83
3	MIYS850101	0.128972844	2.3487250742	0.0821642204	1227	2.06	2.04
3	NAGK730101	0.0535100939	3.8300896135	0.9490742426	1004.5	1.1	1
3	NAGK730102	0.2497467467	1.3423940924	0.6670943653	1066	0.9	0.9
3	NAGK730103	0.6709927012	0.1816583299	0.2270682313	863	0.84	0.84
3	NAKH900101	0.2543965784	1.3161339755	0.1419932132	1194	5.86	5.14
3	NAKH900107	0.1028248643	2.7174096295	0.8070506504	1043	6.34	5.42
3	NAKH900108	0.4569053079	0.5583810535	0.9790551527	1016	-0.12	0.14
3	NAKH900109	0.465688599	0.5368492892	0.0615450965	1243	6.26	4.8
3	NAKH900112	0.0554784818	3.7668617972	0.7147976309	967	4.88	3.51
3	NAKH900113	0.3858598673	0.7594890801	0.1191292137	820	1.24	0.75
3	NAKH920103	0.3832113516	0.7680360546	0.5925904388	946	4.38	5.15
3	NAKH920104	0.4279237421	0.6343056758	0.9468979952	1004	3.73	5.26
3	NAKH920105	0.0543260686	3.803579541	0.8828581899	1031	4.17	2.38

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	NAKH920107	0.839364819	0.0413335485	0.6208482787	951	4.75	5.75
3	NISK800101	0.368501734	0.817084978	0.057113197	1247	-0.07	-0.36
3	NOZY710101	0.1242076376	2.4092716277	0.3502409248	1118	0	0
3	OOBM770101	0.0785111591	3.1687586401	0.3728014332	1123	-1.864	-1.753
3	OOBM770104	0.0724695063	3.3050597611	0.2030595019	855	-16.225	-12.48
3	OOBM770105	0.143020296	2.1840174683	0.5012123235	929	-10.131	-9.424
3	OOBM850103	0.1976040724	1.6853737733	0.1419934573	831	0.43	0.43
3	OOBM850105	0.5426402503	0.3735701846	0.2271514532	1162	5.97	4.55
3	PALJ810104	0.3127514095	1.0308118945	0.3361190484	1131	0.83	0.86
3	PALJ810108	0.1162809149	2.5160311397	0.9790695152	1009	1.02	1.05
3	PALJ810111	0.2391856515	1.4043583697	0.3587407905	1126	0.99	0.98
3	PALJ810114	0.4666431891	0.5345484227	0.7057781542	965.5	0.67	0.91
3	PLIV810101	0.1022365097	2.7268759312	0.3131034502	1137	-3.25	-2.94
3	PONP800101	0.7716606529	0.084738597	0.1557515502	1188	12.01	11.65
3	PONP800102	0.7171881703	0.1320492664	0.3431752462	1130	7.31	7.08
3	PONP800103	0.6371095725	0.224094936	0.4848237632	1099	2.55	2.57
3	PONP800105	0.438388915	0.605992367	0.4927214022	927.5	13.9	14.1
3	PONP800106	0.9754648922	0.0009512674	0.3033407269	1140	11.18	11.18
3	PONP800108	0.6681569815	0.1850066133	0.0674746836	1238	6.16	5.8
3	PTIO830101	0.0628692825	3.549388523	0.0808293156	1226.5	1	1
3	PTIO830102	0.1894596342	1.7487581014	0.1884829295	1174	0.9	0.9
3	QIAN880102	0.3047071821	1.065879038	0.0952767672	807	-0.02	0
3	QIAN880104	0.9020938101	0.0152206697	0.4463177285	1107	-0.03	0
3	QIAN880105	0.9705279552	0.0013728052	0.4560046154	920	-0.08	-0.04
3	QIAN880116	0.4962693328	0.4667752152	0.6177942996	950.5	-0.08	-0.06
3	QIAN880122	0.7018533664	0.1475094164	0.1775361648	1179	0.09	-0.1
3	QIAN880125	0.7140704543	0.1351136259	0.1605632976	839	-0.03	-0.02
3	QIAN880127	0.4408467166	0.5994932536	0.5273526876	1091	0.06	0
3	QIAN880135	0.6377131875	0.2232914581	0.164927982	841	-0.14	0.06
3	QIAN880139	0.4626488702	0.5442268883	0.4658519154	1103	-0.01	0.04
3	RACS770101	0.2487470532	1.3481193552	0.0991328057	809	0.962	0.986
3	RACS770103	0.4762819037	0.5117363291	0.368002697	901	1.63	1.61
3	RACS820101	0.0895956387	2.946243599	0.5751446177	1082	1.03	1.34

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	RACS820103	0.0505256849	3.9309007566	0.8857086556	1030	1.63	1.23
3	RACS820104	0.8537309786	0.034188938	0.4461964476	1107	1.07	1.13
3	RACS820105	0.6873650292	0.1630261757	0.2398767451	867	0.99	1.09
3	RACS820108	0.1973291834	1.6874622132	0.8892538271	1030	1.3	1.19
3	RACS820109	0.1899190179	1.745097693	0.0860752186	834	0	0
3	RACS820112	0.0527809734	3.8541492789	0.2775367904	1147	1.25	1.18
3	RACS820113	0.4687376009	0.5295267694	0.0604139974	781	16.46	19.95
3	RACS820114	0.2190632032	1.5322788862	0.1522280537	835.5	17.82	18.56
3	RICJ880101	0.5909625745	0.2909688527	0.0738187469	793.5	0.9	1
3	RICJ880102	0.5909625745	0.2909688527	0.0738187469	793.5	0.9	1
3	RICJ880114	0.3381356531	0.9275645964	0.1002034713	812.5	0.7	1
3	RICJ880115	0.891541205	0.0187004156	0.6323764167	1071.5	0.9	0.8
3	ROBB760107	0.2928596846	1.1197905846	0.2847205622	880	0.7	0
3	ROBB760108	0.2387780529	1.4068172303	0.3859717782	905	1	0.4
3	ROBB760110	0.7663279299	0.0888648698	0.5540828483	939	2	0.6
3	ROBB760113	0.1627715581	1.9813433325	0.3727345336	902	1	0.2
3	SNEP660101	0.6708154043	0.1818666128	0.8008090792	1044	0.234	0.236
3	SUEM840102	0.3197961649	1.0010653281	0.438544122	919.5	1	8
3	TANS770107	0.6931489635	0.1567239645	0.5860405454	1080	1.288	0.393
3	VASM830101	0.2874263323	1.1454694312	0.1079907118	1211	0.215	0.184
3	VASM830102	0.5650269466	0.3335912227	0.6208499717	1074	0.459	0.431
3	VASM830103	0.6548038606	0.2012657466	0.299541277	884	0.194	0.206
3	VELV850101	0.826595458	0.0482783757	0.1516351585	1189.5	0.08292	0.03731
3	WARP780101	0.7386164057	0.1120520912	0.7821529932	1047	7.98	7.08
3	WEBA780101	0.1322692022	2.3083040211	0.3306812461	892	0.88	0.85
3	WERD780101	0.4105586603	0.6836804587	0.0663438247	1238	0.49	0.49
3	WERD780104	0.2690730845	1.2370318327	0.1044516627	1213	-0.17	-0.17
3	WOEC730101	0.3994022092	0.7170745611	0.1699235844	843	7.9	7.5
3	WOLS870101	0.1020758331	2.7294716647	0.1079916749	814	2.18	1.96
3	ZIMJ680103	0.0918021061	2.9055750819	0.5697912429	1083	1.67	1.67
3	ZIMJ680105	0.0598134003	3.6357675708	0.3493536167	1128.5	6.9	8.2
3	AURR980102	0.5129302811	0.4315776603	0.055288902	1248	1.12	1.04
3	AURR980112	0.054246206	3.8061551641	0.2178844106	1165	1.4	0.96

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	AURR980115	0.4658637612	0.5364265226	0.1260801968	1201.5	1.65	0.89
3	AURR980117	0.7428507395	0.1083167093	0.3012099765	1140.5	1.04	0.92
3	AURR980119	0.8748098332	0.0249676052	0.4437739204	1107.5	0.97	0.97
3	VINM940103	0.6736863275	0.178511279	0.8700984261	1033	1.018	1.018
3	VINM940104	0.3095907552	1.0444471358	0.3446099305	895.5	0.796	0.799
3	MUNV940104	0.0894828914	2.9483512991	0.10839698	815	0.784	0.784
3	WIMW960101	0.8733819935	0.0255453395	0.0549841111	1249	4.24	4.08
3	MONM990101	0.6200391131	0.2475604396	0.955791153	1005.5	1.3	0.7
3	KUMS000101	0.3611553654	0.8426278481	0.4086177467	1115	4.6	4.5
3	KUMS000102	0.7329204076	0.1171883942	0.76963797	976	3.6	5.5
3	KUMS000104	0.6692850181	0.1836703403	0.1740946384	1180.5	4.5	4.5
3	FODM020101	0.2781515585	1.1907774562	0.6322372032	1072	0.95	0.96
3	NADH010106	0.540374377	0.3777900199	0.6380186357	1071	-38	-8
3	MONM990201	0.1113132465	2.5871907962	0.3088257801	887.5	1.1	0.9
3	CEDJ970101	0.8607468001	0.0309542191	0.9790609613	1016	4.2	4.9
3	CEDJ970102	0.5211125019	0.4150113004	0.6208195186	1074	5	5.2
3	CEDJ970104	0.2963609568	1.103567071	0.4706676453	1102	4.9	5.3
3	FUKS010102	0.7206183496	0.1287237369	0.9212264024	1025	5.41	5.41
3	FUKS010104	0.5087047738	0.4403157367	0.7945747533	1045	5.81	5.81
3	FUKS010105	0.3732599792	0.8009193361	0.0911985393	1221	4.1	4.02
3	FUKS010106	0.3015761464	1.0798579447	0.364198975	1125	4.93	3.51
3	FUKS010107	0.2766836594	1.1981252655	0.7883573643	1046	5.41	4.84
3	FUKS010108	0.0905575623	2.9283794872	0.0953400087	1218	5.62	2.86
3	FUKS010110	0.2943932852	1.1126537223	0.6010168982	1077.5	4.9	5.15
3	FUKS010111	0.1607392041	2.0008537201	0.0638626397	784	3.71	6.38
3	FUKS010112	0.9046585966	0.0144298228	0.2059780332	1169	5.17	5.17
3	COSI940101	0.8252667252	0.0490334826	0.1516351585	1189.5	0.0829	0.0373
3	BASU050103	0.1571289247	2.0362258555	0.8956224089	996	0.0248	0.0381
3	SUYM030101	0.2883012041	1.1412925767	0.4511870137	1106	0.06	0.027
3	GEOR030103	0.3044984436	1.0668051245	0.0519481381	1252	1	0.986
3	GEOR030104	0.7709697974	0.085267147	0.1291298834	1200	1.069	0.992
3	GEOR030107	0.3199942128	1.0002416563	0.0821642714	1227	1.058	0.994
3	GEOR030109	0.8044576984	0.0616677754	0.1027169135	1214	1.003	0.956

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
3	ZHOH040103	0.1630894784	1.9783168587	0.1240720832	1202	8.5	8.5
3	BAEK050101	0.2854129463	1.1551442545	0.0738621886	1233	-0.0442	-0.0701
3	DIGM050101	0.1206331061	2.4564362393	0.0993140897	1216	1.105	1.056
3	GUYH850102	0.5290532138	0.3993699511	0.082346369	799	0.12	0.12
3	CASG920101	0.3309887245	0.9555561027	0.5114914602	1093.5	-0.1	-0.4
3	CORJ870101	0.4018795621	0.7095415403	0.0549988538	1249	50.27	49.26
3	MIYS990101	0.09848223	2.7887424279	0.0539681968	775	0.38	0.38
3	MIYS990102	0.0946859656	2.8540198631	0.0517762163	773	0.06	0.06
3	MIYS990103	0.3459030474	0.8980394029	0.0535955287	775.5	0.09	0.09
3	MIYS990105	0.1138133165	2.5509433962	0.0635811718	784	-0.02	0.05
4	ANDN920101	0.4885775609	0.4837136958	0.2762597464	877.5	4.36	4.44
4	ARGP820101	0.521561296	0.414115942	0.751546294	1052	1.15	1.15
4	ARGP820103	0.1052076467	2.6796789735	0.3955473584	907	0.62	0.81
4	BEGF750101	0.0505443479	3.9302505924	0.3166369291	1135.5	0.6	0.6
4	BEGF750103	0.225556222	1.4894879493	0.9164131059	999.5	0.75	0.75
4	BHAR880101	0.5430113033	0.372882264	0.300268496	1141	0.466	0.466
4	BIOV880101	0.4755198379	0.5135124349	0.9499368598	1004.5	-20	-20
4	BIOV880102	0.5538516276	0.3531651553	0.5121159033	931	-36	-36
4	BROC820102	0.1868370539	1.7698577784	0.0684247026	1237.5	3.2	2.1
4	BULH740101	0.1516048439	2.09220426	0.4437774357	1107.5	-0.35	-0.39
4	BULH740102	0.9223278771	0.0095612941	0.0916079233	1221	0.73	0.709
4	BUNA790101	0.2397573184	1.4009184334	0.4138031948	1114	8.391	8.38
4	BUNA790102	0.4256785079	0.6405182848	0.4615389612	921	4.385	4.471
4	BUNA790103	0.0524514647	3.8651402583	0.374267921	1120	6.5	6.5
4	BURA740101	0.6070533842	0.2663971718	0.3395053032	1131	0.318	0.318
4	BURA740102	0.2710597778	1.2267421433	0.3275251253	891	0.34	0.354
4	CHAM810101	0.7587292665	0.0949297299	0.3531799171	1126.5	0.68	0.68
4	CHAM820102	0.4958662094	0.4676519231	0.8763990841	1032	0	0
4	CHAM830101	0.0515931898	3.8941214563	0.8322997695	986	0.99	1.06
4	CHAM830105	0.3456220163	0.8990918264	0.0556163693	1225	1	0
4	CHAM830107	0.3740131749	0.7983870968	0.5501728066	1080	0	0
4	CHAM830108	0.3989793047	0.7183673469	0.2038751543	1170	1	0
4	CHOC760103	0.2263729992	1.4842111973	0.1534870727	836	0.18	0.22

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	CHOC760104	0.3550484948	0.8644188502	0.0971843246	809	0.04	0.08
4	CHOP780201	0.1566072738	2.0414143316	0.3074536087	1139	1.13	1.01
4	CHOP780202	0.2305118311	1.4578244692	0.7360976839	1054.5	0.93	0.83
4	CHOP780203	0.1812416531	1.8160654693	0.5926790285	946	1.01	1.01
4	CHOP780205	0.2712050307	1.2259935807	0.2268712589	1162	1.13	1.11
4	CHOP780206	0.8332625262	0.0445821004	0.5173179635	932	0.93	0.98
4	CHOP780208	0.0866397114	3.0024913362	0.8892357607	1030	0.9	0.87
4	CHOP780209	0.5271972247	0.4029879744	0.6404146457	1070.5	0.9	0.85
4	CHOP780210	0.1588549528	2.0191990235	0.1610071321	839	0.89	1.09
4	CHOP780211	0.2912881018	1.1271545287	0.2632484644	874	0.85	0.85
4	CHOP780212	0.0513206693	3.9034323334	0.1232470878	822	0.061	0.07
4	CHOP780213	0.3196176802	1.0018082185	0.0716099383	790	0.085	0.085
4	CHOP780214	0.2303788793	1.4586630796	0.1022637449	1214.5	0.072	0.065
4	CHOP780215	0.2341220154	1.4352762534	0.331553221	1133	0.081	0.07
4	CHOP780216	0.0988196102	2.7830765327	0.5438086029	937	1.05	1.05
4	CIDH920101	0.869233328	0.0272624791	0.446899198	1107	-0.24	-0.2
4	CIDH920102	0.6614065015	0.1931239343	0.9079705696	1027	-0.09	-0.08
4	CIDH920103	0.3821804846	0.7713856674	0.4914364301	927	-0.52	-0.06
4	CIDH920104	0.4126240723	0.677645099	0.7272077608	969	-0.57	-0.21
4	CIDH920105	0.5141107326	0.4291589919	0.810424624	1042.5	-0.41	-0.09
4	COHE430101	0.2295719151	1.4637658228	0.3813964771	1121	0.76	0.75
4	CRAJ730101	0.3259106638	0.9759443474	0.6529657142	1068.5	1.03	0.96
4	CRAJ730102	0.9018968278	0.0152822945	0.9919554379	1014	0.89	0.99
4	CRAJ730103	0.0865773462	3.0037007403	0.7791975465	977.5	0.79	1.1
4	DAYM780101	0.5215146108	0.4142090167	0.3073478966	1139	6	5.5
4	DESM900101	0.3874058698	0.7545388507	0.2537552964	871	0.96	0.96
4	DESM900102	0.1558761648	2.0487197771	0.3520959486	897	0.98	0.98
4	EISD860103	0.4722342005	0.5212237389	0.2518834441	870.5	-0.67	0
4	FASG760102	0.1290425276	2.3478585334	0.0817268285	1227.5	270	249
4	FASG760103	0.6699678368	0.1828642612	0.5764115699	1082	0	-7.5
4	FASG760104	0.3581476627	0.8532956835	0.1465990997	833.5	9.18	9.21
4	FASG760105	0.2828365857	1.1676528734	0.5474203018	938	2.16	2.19
4	FAUJ830101	0.0755130955	3.2349050439	0.2964826263	883	0	0.31

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	FAUJ880101	0.0925541641	2.8919596369	0.8734797404	992.5	2.34	2.34
4	FAUJ880102	0.6094858243	0.2628023251	0.2180595086	1165	0.71	0.66
4	FAUJ880108	0.0505196123	3.9311123659	0.1073941196	816	0	0.03
4	FAUJ880110	0.3879348162	0.752851711	0.4453355979	1099.5	1	0
4	FAUJ880112	1	3.60115292e-33	1	1012.5	0	0
4	FAUJ880113	0.1522958243	2.0850747793	0.2781585355	1147	4.3	4.27
4	FINA910101	0.1175118717	2.4989246156	0.4660534212	928	1	1
4	FINA910102	0.287968827	1.142877503	0.1111432878	832.5	1	1
4	FINA910103	0.3930894911	0.7365822785	0.2217521423	1159.5	1	1
4	GARJ730101	0.5353730577	0.3872205916	0.7885377108	979	0.21	0.28
4	GEIM800101	0.1081254759	2.6347523073	0.1170553264	1206	1.13	1
4	GEIM800102	0.0713948428	3.3306022176	0.898300401	996.5	1.09	1.08
4	GEIM800103	0.4327498545	0.6211183253	0.0986755554	808.5	1.2	1.2
4	GEIM800104	0.1084487945	2.6298574736	0.311634799	1138	1.02	1
4	GEIM800106	0.9010239749	0.0155568834	0.0754874689	1231.5	1.15	1.01
4	GEIM800107	0.5875937322	0.2962908807	0.5093775429	1094.5	0.99	0.93
4	GEIM800108	0.3184407765	1.0067205808	0.4966453594	928	0.9	0.93
4	GEIM800109	0.0597325262	3.6381180905	0.5178411199	1092.5	0.96	0.94
4	GEIM800110	0.501027151	0.4565187713	0.0825516623	798	0.93	0.94
4	GEIM800111	0.2544252106	1.3159741305	0.4714092163	923	0.96	0.93
4	GOLD730101	0.6672453089	0.1860907832	0.755629546	1051	1.5	1.3
4	GRAR740102	0.1995637985	1.6705844304	0.8417675046	1037.5	9.2	8.1
4	GUYH850101	0.1143069031	2.5438925244	0.2470452288	1156	0.48	0.48
4	HOPA770101	0.5888467047	0.2943041566	0.1332022665	1198	1.7	1.1
4	HUTJ700101	0.6905065794	0.1595852257	0.6912264552	963	48.03	38.3
4	ISOY800103	0.1028435703	2.7171096492	0.7031811681	1060	1.01	1.01
4	ISOY800104	0.2185939142	1.535431068	0.263159144	874	0.97	0.97
4	ISOY800106	0.662331332	0.1919995532	0.4164526208	1113	1.09	1.09
4	ISOY800108	0.3266529411	0.9729376277	0.2007168806	854	0.61	0.68
4	JANJ790101	0.8791625956	0.0232481469	0.2305798045	864	0.8	0.8
4	JOND750101	0.5295866016	0.3983344069	0.751546294	1052	1.64	1.64
4	JOND920101	0.5070026825	0.44387131	0.2945645077	1142	0.059	0.053
4	JUKT750101	0.4530250481	0.5681046884	0.3355930039	1132	3.6	3.6

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	JUNJ780101	0.4674431749	0.5326260471	0.4814278826	1100	427	400
4	KANM800101	0.0628609891	3.5496167409	0.1841360392	1177	1.05	1
4	KANM800102	0.152850821	2.0793752519	0.8291432951	985.5	0.85	0.92
4	KANM800104	0.6236471824	0.2424796905	0.9340979614	1023	0.79	0.82
4	KARP850101	0.548969486	0.361954658	0.4277284545	914	1.038	1.041
4	KARP850103	0.3492135209	0.8857296752	0.643770823	1070	0.921	0.923
4	KRIW710101	0.5845943163	0.3010821303	0.1841297896	1177	5.9	5.25
4	KRIW790101	0.2297120719	1.4628779881	0.0854866499	1225	6.09	5.37
4	KRIW790102	0.683964851	0.1667986598	0.3498227961	1128.5	0.28	0.27
4	KYTJ820101	0.083695448	3.0606577362	0.179982798	846.5	-1.6	-0.8
4	LAWE840101	0.6842094498	0.1665255965	0.8766354518	1032	-0.06	0.05
4	LEVM760103	0.4115119919	0.6808891227	0.3905960198	1119	118.2	118.2
4	LEVM760106	0.7631047153	0.0914107641	0.4040732211	1114.5	6	6
4	LEVM780101	0.7064084527	0.1428149646	0.1167086449	1206.5	1.07	0.97
4	LEVM780103	0.0707978543	3.3449731575	0.8260956617	985	0.88	0.88
4	LEVM780106	0.069066599	3.3874073471	0.9501696353	1004.5	0.9	0.9
4	LEWP710101	0.088047596	2.9754415156	0.0686509081	788	0.27	0.28
4	LIFS790102	0.0579421556	3.6910538026	0.5815082824	944	0.68	0.7
4	LIFS790103	0.1183456398	2.4874517135	0.9823103976	1015.5	1.02	0.9
4	MANP780101	0.285880486	1.1528898071	0.5331376988	1090	11.89	11.72
4	MAXF760101	0.1656189705	1.9544782263	0.0728577497	1234	1.19	0.98
4	MAXF760104	0.1953847697	1.7023341154	0.0571953371	1247	0.57	0.28
4	MAXF760105	0.5455456497	0.3682069223	0.2453299733	1156	0.65	0.48
4	MCMT640101	0.3605352369	0.8448172039	0.0607368144	1244	21.29	13.92
4	MEEJ800101	0.5389751376	0.3804122442	0.760633875	974.5	0.8	1.2
4	MEEJ800102	0.3229569742	0.9880001301	0.8323677504	986	-0.5	-0.1
4	MEEJ810101	0.5444020459	0.3703116192	0.3777963299	1122	-0.2	1.1
4	MEEJ810102	0.0935043286	2.8749313999	0.7270902885	1056	0.2	1
4	MEIH800101	0.6612056021	0.1933687003	1	1012.5	0.98	0.98
4	MEIH800102	0.4067493042	0.6949308401	0.5057416938	1095	1.04	1.04
4	MEIH800103	0.1581187906	2.0264348235	0.8606459013	1034.5	83	83
4	MIYS850101	0.8773311383	0.0239639657	0.6500692848	1069	1.92	1.92
4	NAGK730101	0.6931813414	0.1566890906	0.1756364261	1177	1.23	0.94

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	NAGK730102	0.7871678513	0.0733405268	0.258719563	872.5	0.87	0.9
4	NAGK730103	0.4918530827	0.4764459014	0.2184756287	860	0.84	0.87
4	NAKH900101	0.9547064062	0.0032444111	0.0869857187	1224	6.01	5.86
4	NAKH900104	0.0805501025	3.1253345141	0.4812787071	925	-0.37	-0.05
4	NAKH900106	0.1076891102	2.6413845682	0.4864407103	926	-0.48	0.11
4	NAKH900107	0.0520787966	3.8776608502	0.6155876593	1075	4.67	4.67
4	NAKH900108	0.0900277979	2.9381918178	0.985526469	1015	-0.12	-0.19
4	NAKH900109	0.0866010378	3.0032411931	0.1635454573	1185	4.8	4.36
4	NAKH900113	0.2347614504	1.4313268914	0.4300662213	914.5	0.75	0.73
4	NAKH920101	0.0747403636	3.252420018	0.1235777095	1203	6.75	6.25
4	NAKH920102	0.100990639	2.7471227874	0.1084412658	1211	6.11	6.11
4	NAKH920103	0.5745337263	0.3175220315	0.6099267786	1076	5.25	5.25
4	NAKH920104	0.2225636514	1.5090216594	0.234044159	1160	6.22	6.07
4	NAKH920107	0.2388327382	1.4064870383	0.5331361688	1090	4.93	4.95
4	NAKH920108	0.0505063147	3.9315758314	0.6240726629	951.5	1.96	3.93
4	NISK800101	0.1063777962	2.6614969328	0.8957617625	996	-0.26	-0.26
4	NISK860101	0.7418945055	0.1091541184	0.5491776194	1087	-0.93	-0.93
4	NOZY710101	0.5425581743	0.3737224688	0.6321542135	1065.5	0	0
4	OOBM770101	0.0591808291	3.654245041	0.2215156026	1164	-1.753	-1.753
4	OOBM770103	0.2760961372	1.2010801295	0.9085321606	998	-0.554	-0.491
4	OOBM850101	0.6474552408	0.2105663697	0.8703065441	992	1.47	1.75
4	OOBM850102	0.4019042185	0.7094669084	0.1317306714	826	0.95	0.95
4	OOBM850103	0.9980897043	5.76486907e-06	0.4615050554	921	0.43	-0.33
4	OOBM850104	0.3462966015	0.8965675617	0.8893842346	995	-1.6	-2.49
4	OOBM850105	0.6175450185	0.2511110746	0.1923204475	1174	5.16	5.14
4	PALJ810101	0.7778726264	0.0800659285	0.0900926106	1221.5	1.09	0.93
4	PALJ810102	0.088101426	2.9744167495	0.0840087814	1226	1.1	1.02
4	PALJ810104	0.2567948696	1.3028222781	0.7647183212	1049.5	0.82	0.82
4	PALJ810105	0.3190657448	1.0041089829	0.4590195548	920.5	0.91	0.91
4	PALJ810107	0.1139213238	2.5493975904	0.2470688187	869	0.96	1.01
4	PALJ810108	0.4482148922	0.5803427479	0.2340531841	1160	1.05	1.05
4	PALJ810109	0.4139897406	0.673679109	0.0885051058	1223	1.06	0.99
4	PALJ810114	0.6660054767	0.187571246	0.7001435237	1060.5	0.91	0.91

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	PALJ810115	0.2199856492	1.5261066293	0.6594885401	958	0.77	0.91
4	PALJ810116	0.2279129296	1.4743253121	0.6113321352	949.5	0.9	0.9
4	PARJ860101	0.3208340125	0.9967564911	0.4565458477	920.5	4.2	2.1
4	PLIV810101	0.4816775482	0.4992933849	0.7290997332	1055.5	-3.25	-2.89
4	PONP800101	0.2665496952	1.2502405321	0.5817876241	944	11.49	11.49
4	PONP800102	0.3138023833	1.0263181419	0.2571443568	872	6.93	6.93
4	PONP800103	0.407163631	0.6936996405	0.0563443411	777	2.55	2.6
4	PONP800104	0.2341211295	1.4352817343	0.2710172206	1149	12.24	12.06
4	PONP800105	0.596121517	0.2829392433	0.6879112729	962.5	14.1	14.1
4	PONP800106	0.253970616	1.3185146562	0.6912391833	963	11.05	11.18
4	PONP800107	0.791076743	0.0706068717	0.8576089184	1035	3.13	2.6
4	PONP800108	0.7793479148	0.0789772574	0.9855267698	1010	5.7	5.7
4	PRAM820102	0.1379509715	2.241245596	0.6522328106	956.5	0.104	0.104
4	PRAM900102	0.7064084527	0.1428149646	0.1167086449	1206.5	1.07	0.97
4	PRAM900104	0.0612685798	3.5940527284	0.857630668	990	0.88	0.88
4	PTIO830101	0.4835409597	0.495049505	0.092057813	1219.5	1	0.95
4	PTIO830102	0.1936243482	1.7159511125	0.3685128561	902	0.7	0.9
4	QIAN880102	0.150398786	2.1047382078	0.4339431879	915.5	-0.02	0.01
4	QIAN880103	0.6900198745	0.1601155235	0.4521322137	1105.5	-0.09	-0.1
4	QIAN880105	0.3756166446	0.7930202148	0.3691037496	1124	0.03	-0.08
4	QIAN880106	0.2996863091	1.0883873807	0.1465962628	1192	0.2	-0.03
4	QIAN880107	0.1021476308	2.7283112097	0.0885053952	1223	0.23	0
4	QIAN880108	0.1414675998	2.2012969556	0.0630210421	1242	0.23	0.1
4	QIAN880116	0.5276973561	0.402010774	0.1306706163	825.5	-0.09	-0.08
4	QIAN880117	0.1236485612	2.4165467547	0.647156991	955.5	0.04	0.02
4	QIAN880118	0.1655033267	1.9555588453	0.7853868621	978.5	0.09	0.11
4	QIAN880119	0.1162414985	2.5165822623	0.7854841579	978.5	0.19	-0.1
4	QIAN880120	0.0752412845	3.2410435693	0.9213271216	1000	-0.02	-0.25
4	QIAN880121	0.086834671	2.9987167987	0.9597797333	1006	-0.09	-0.26
4	QIAN880122	0.492806822	0.4743450109	0.9052145369	997.5	0.02	0.03
4	QIAN880124	0.7206937949	0.1286511323	0.1519026751	836	-0.1	-0.1
4	QIAN880125	0.534378428	0.389115353	0.6326955409	953	-0.03	-0.03
4	QIAN880126	0.5951464706	0.2844457647	0.067822657	787	-0.04	0

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	QIAN880127	0.0887245286	2.9626046891	0.2001600456	854	0.04	0.04
4	QIAN880129	0.2049768061	1.6306177374	0.4758040575	1101	0.06	-0.01
4	QIAN880131	0.2351707483	1.4288058476	0.1084398951	814	0.12	0.14
4	QIAN880132	0.2053847404	1.6276569727	0.8703074227	992	-0.4	0
4	QIAN880134	0.1902240967	1.7426725786	0.5491825518	938	-0.04	-0.04
4	QIAN880135	0.3879348162	0.752851711	0.3188337971	889	-0.14	-0.09
4	QIAN880137	0.1582789552	2.0248572457	0.6164923763	950.5	-0.09	-0.09
4	QIAN880138	0.0545473753	3.7964635449	0.054271281	775	-0.2	-0.15
4	QIAN880139	0.9339426283	0.0069094121	0.3930048582	1118.5	-0.01	0.08
4	RACS770101	0.325828558	0.9762774939	0.3451525614	895.5	0.962	0.962
4	RACS770102	0.6359715097	0.225614639	0.486446494	1099	1.055	1.055
4	RACS770103	0.6558750448	0.1999311341	0.6384901338	1071	1.63	1.61
4	RACS820101	0.302830851	1.0742334841	0.1930069179	1173.5	1.03	1.17
4	RACS820102	0.1832203204	1.7995357754	0.8766648861	993	1.14	1.09
4	RACS820104	0.1862654725	1.7745027757	0.6912199947	963	1.07	1.25
4	RACS820106	0.7467875346	0.1049067172	0.1710156236	1182	0.96	0.9
4	RACS820107	0.447109181	0.5831851258	0.4373025789	1109	1.2	0.92
4	RACS820108	0.3318791872	0.9520242308	0.1469808195	1192	1.3	1.02
4	RACS820109	0.2476194637	1.3546114449	0.4583944761	1085	0	0
4	RACS820110	0.7293454017	0.1204779553	0.2094483145	857	1	1.05
4	RACS820111	0.3865019436	0.7574297235	0.8010482162	981	1.03	1.08
4	RADA880108	0.5167609241	0.4237642696	0.2212832193	861	-0.41	-0.06
4	RICJ880101	0.2026821555	1.6474045054	0.530794691	1090	0.9	0.9
4	RICJ880102	0.2026821555	1.6474045054	0.530794691	1090	0.9	0.9
4	RICJ880104	0.1781492669	1.842328141	0.1241101911	824.5	0.7	0.8
4	RICJ880105	0.6540867777	0.2021621622	0.4305633623	915.5	0.9	0.8
4	RICJ880106	0.3286242092	0.9649969296	0.3752561264	905.5	0.6	0.6
4	RICJ880108	0.0785321119	3.168306142	0.7168332884	1057.5	1.1	1.1
4	RICJ880111	0.4244441145	0.6439551046	0.3177097992	1136	1	0.8
4	RICJ880113	0.0738552397	3.2727272727	0.2978254559	1141.5	0.9	1.3
4	RICJ880114	0.6962103243	0.153446442	0.5889740666	946	0.8	1
4	RICJ880115	0.1722967662	1.8935247961	0.0782549337	1227.5	0.9	0.8
4	RICJ880116	0.7336748013	0.1165007414	0.1401777847	1194	1.2	0.9

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	RICJ880117	0.1267496196	2.3766530886	0.8479826547	988.5	1.1	1
4	ROBB760101	0.4378068768	0.6075396978	0.2964995313	1142	1.6	0.6
4	ROBB760102	0.4683477834	0.5304586604	0.2006953117	854.5	-2.8	-3.5
4	ROBB760103	0.1169298553	2.5069876598	0.5017854249	1096	0.5	0.3
4	ROBB760104	0.209726057	1.5965777871	0.6675162398	1066	1.4	1.4
4	ROBB760105	0.3077592773	1.0524323685	0.475806667	924	0.4	-1.7
4	ROBB760106	0.2214795975	1.5161765367	0.7387914517	971	0.4	-2.4
4	ROBB760108	0.7600741051	0.0938403965	0.4533794451	919.5	1	1
4	ROBB760110	0.9214246488	0.0097856978	0.7697387382	976	2	2
4	ROBB760111	0.3713353412	0.8074228979	0.4207507535	1112.5	1	-0.5
4	ROBB760113	0.64288837	0.2164749868	0.4714137505	923	1	1
4	ROBB790101	0.1413581447	2.2025232106	0.3260573279	1134	0.3	0.3
4	ROSG850102	0.8599055408	0.0313333382	0.3881482567	906	0.66	0.66
4	ROSM880103	0.0852611187	3.0294514388	0.8644639541	1033.5	0.6	0.4
4	SIMZ760101	0.749207715	0.1028403058	0.5896182617	1079.5	1.5	1.3
4	SNEP660101	0.4550298275	0.5630644536	0.5384572674	936	0.228	0.236
4	SNEP660102	0.947859146	0.0043010373	0.1619074405	839.5	-0.011	0.022
4	SUEM840101	0.3760349459	0.7916254816	0.1946721614	1173	1.033	0.939
4	SUEM840102	0.0617297369	3.5810565936	0.2964210284	887	1	6
4	SWER830101	0.4526468921	0.5690593699	0.4757011938	924	-0.59	-0.49
4	TANS770105	0.6599173378	0.1949426657	0.1138672097	1208	0.677	0.677
4	TANS770106	0.6336469565	0.2287383361	0.1854309792	1176.5	0.901	0.937
4	TANS770107	0.0549836409	3.7825264383	0.0566958427	1246	1.061	0.218
4	TANS770108	0.4926552352	0.4746784666	0.8323595225	986	0.835	0.982
4	TANS770109	0.260677968	1.2815954855	0.1402888597	830	0.758	0.945
4	VASM830101	0.7721497421	0.0843654899	0.158672879	1187	0.215	0.17
4	VASM830102	0.1990398006	1.6745218652	0.6043011624	948	0.459	0.507
4	VASM830103	0.6518638173	0.2049563198	0.0928079477	805	0.194	0.198
4	VELV850101	0.575916868	0.315227751	0.6380216756	954	0.0371	0.03731
4	VENT840101	0.2441191741	1.375	0.3467030288	1125	0	0
4	WARP780101	0.6329405258	0.2296928643	0.1196454419	820	7.49	7.79
4	WEBA780101	0.4671878963	0.5332389067	0.3561684208	898	0.88	0.85
4	WERD780101	0.8001926648	0.0644470411	0.8984986344	1028.5	0.49	0.49

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	WERD780102	0.0502586074	3.940233373	0.6211250099	951	-0.2	-0.19
4	WERD780104	0.6309842957	0.2323489145	0.4513985265	919	-0.4	-0.11
4	WOEC730101	0.2132279633	1.5720653789	0.5273515567	1091	7.9	7
4	WOLS870101	0.4475964003	0.5819313075	0.4000627342	1117	2.23	0.71
4	ZASB820101	0.4136054151	0.6747932414	0.0541733026	1250	-0.089	-0.125
4	ZIMJ680101	0.5706821998	0.3239688041	0.5192002659	1092.5	1.6	1.4
4	ZIMJ680102	0.7176440729	0.131604505	0.6731742448	1065	15.71	15.71
4	ZIMJ680105	0.4373905175	0.6086485267	0.876644045	1032	5.6	9.5
4	AURR980102	0.6637035868	0.1903384217	0.2618770547	1151.5	1.06	1.02
4	AURR980103	0.6280281634	0.2363984081	0.7352477324	970.5	0.96	0.97
4	AURR980106	0.5183331267	0.4205868686	0.9919554979	1011	0.9	0.9
4	AURR980107	0.1568637209	2.0388611544	0.2537488172	1154	0.94	0.91
4	AURR980108	0.3591378173	0.8497700495	0.123559771	1203	1.33	0.9
4	AURR980112	0.3367951708	0.9327533997	0.1922461958	1174	1.4	0.94
4	AURR980119	0.2245588091	1.4959633001	0.1684154995	1183	1.03	0.97
4	VINM940101	0.6946363397	0.1551265619	0.9790957512	1009	1.008	1.008
4	VINM940102	0.4038765627	0.7035185731	0.512123628	931	1.31	1.315
4	VINM940103	0.5175469004	0.4221736919	0.5147039582	1093.5	1.022	1.022
4	VINM940104	0.1247997746	2.4016064637	0.4535484387	919.5	0.799	0.806
4	MUNV940101	0.277650809	1.1932784756	0.0795456534	796	0.503	0.706
4	MUNV940102	0.4716215209	0.5226713744	0.1102490243	815	0.753	0.939
4	MUNV940104	0.182667209	1.8041352341	0.4719368739	1101.5	0.784	0.784
4	MUNV940105	0.1907822043	1.738247984	0.0690649179	1237	1.33	1.2
4	WIMW960101	0.0855204296	3.0243438288	0.7001312372	1060.5	4.08	4.12
4	KIMC930101	0.8069259811	0.0600890524	0.4776559605	924.5	-0.44	-0.41
4	MONM990101	0.6453029607	0.2133386107	0.8078589718	1042.5	1.6	0.7
4	BLAM930101	0.3126548611	1.0312257159	0.1325401029	1198.5	0.73	0.59
4	PARS000101	0.2860212583	1.1522119396	0.1535086081	836	0.353	0.353
4	PARS000102	0.3443237042	0.90396907	0.1979341766	1172	0.34	0.339
4	KUMS000101	0.3912798829	0.7422585333	0.1196494485	1205	6.1	4.6
4	KUMS000102	0.5263272597	0.4046917646	0.4862785023	1099	6	5.5
4	TAKK010101	0.4695098834	0.527684268	0.4420000091	1108	10.5	10.5
4	NADH010101	0.1759690657	1.8611671498	0.2817386017	879	-24	-24

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	NADH010102	0.1291426241	2.346614724	0.2034078635	855	-55	-26
4	NADH010103	0.4393912651	0.6033350788	0.2437523216	868	-35	-31
4	NADH010104	0.6770101972	0.1746725347	0.3002596247	884	-29	-29
4	NADH010105	0.8916571443	0.0186602043	0.7577442934	974	-9	-9
4	NADH010106	0.1574179283	2.0333598996	0.2500638605	870	-38	-8
4	NADH010107	0.0600758798	3.6281625204	0.991959303	1014	36	28
4	MONM990201	0.6436506803	0.2154817924	0.8034009382	981.5	1.3	0.9
4	CEDJ970101	0.3683889516	0.8174717083	0.3953160014	1118	5.1	4.9
4	CEDJ970102	0.5767334255	0.3138784326	0.4914562102	1098	5.8	5.4
4	CEDJ970103	0.4860351701	0.4894113321	0.4200977303	1112.5	4.7	4.7
4	CEDJ970104	0.6360677591	0.2254858692	0.1561341544	1188	6.7	5.5
4	FUKS010102	0.8371267583	0.042510095	0.0665959183	1239	6.87	5.36
4	FUKS010103	0.1332473516	2.2965289787	0.7793007767	977.5	4.51	5.6
4	FUKS010104	0.912445586	0.0121596833	0.1923203049	1174	5.81	5.7
4	FUKS010105	0.5175218493	0.4222243221	0.2710187032	1149	4.02	3.8
4	FUKS010106	0.4247903121	0.6429896882	0.4420638683	1108	3.51	3.42
4	FUKS010107	0.4604490506	0.5496146856	0.8135487976	1042	4.46	4.46
4	FUKS010108	0.0648757221	3.4951140663	0.7481358115	1052.5	3.3	3.3
4	FUKS010110	0.5697145256	0.325602033	0.1814572067	1178	6.01	5.15
4	FUKS010111	0.218722958	1.5345634745	0.5930832671	946	5.45	5.45
4	FUKS010112	0.2600106377	1.2852151264	0.1102475068	1210	5.72	5.5
4	COSI940101	0.5767277783	0.3138877511	0.6380216756	954	0.0371	0.0373
4	PONP930101	0.068557535	3.4001055341	0.9855267313	1015	0.2	0.2
4	GUOD860101	0.172710411	1.8898399916	0.5631646773	941	-2	25
4	JURD980101	0.0591924526	3.6539035925	0.1157669181	818	-1.9	-0.5
4	BASU050101	0.6700808308	0.1827310723	0.5384599854	1089	0.0363	0.0363
4	BASU050102	0.5759207957	0.3152212516	0.3195842591	1136	0.0394	0.0394
4	BASU050103	0.6469979268	0.2111535719	0.2745555002	877	0.0248	0.0844
4	SUYM030101	0.2083426037	1.606397457	0.0596158185	1245	0.016	-0.052
4	PUNT030102	0.3658986214	0.8260536398	0.6442130884	1070	0.15	0.15
4	GEOR030103	0.8238718519	0.049832796	0.1586800644	1187	1	0.959
4	GEOR030104	0.7871556954	0.0733491149	0.1610874502	1186	1.042	0.992
4	GEOR030105	0.59149769	0.2901292292	0.9469680072	1021	1.131	1.097

Continued on next page...

Table C.1: Cleavage Site Dataset: Indistinguishable Features for N-Termini Positions 1-4

N-term. Res. Pos.	AAIndex Entry	Lavene P-Value	Lavene Statistic	M-W-Wilcoxon P-Value	U-Statistic	Cath. Fea. Median	Cleavage Site Fea. Median
4	GEOR030109	0.3649153717	0.8294646077	0.7638933034	1050	0.978	0.978
4	ZHOH040101	0.28222425	1.1706473455	0.1672541581	1183.5	2.71	2.18
4	ZHOH040103	0.6106195941	0.2611373008	0.8543804279	989.5	8.5	9.9
4	BAEK050101	0.6926351536	0.1572779854	0.4615148983	1104	-0.0762	-0.0762
4	DIGM050101	0.3552273645	0.8637732196	0.1492675749	1191	1.105	1.076
4	OLSK800101	0.4123462142	0.6784544121	0.0653914638	785	0.85	0.86
4	GUYH850102	0.9196511505	0.0102339667	0.8665367674	1033.5	0.12	0.12
4	CASG920101	0.335769072	0.9367442177	0.5993305994	947.5	-0.5	-0.5
4	CORJ870101	0.1277824861	2.3636099505	0.8830177686	994	48.66	48.66
4	CORJ870102	0.4501060034	0.5755068079	0.4517409469	919	-0.584	-0.503
4	CORJ870103	0.809240283	0.0586285262	0.2007673392	1171	0.75	-0.96
4	CORJ870104	0.6206066982	0.2467568282	0.237016437	1159	0.08	-0.26
4	CORJ870105	0.6265699606	0.2384119146	0.3078995701	1139	-1.03	-1.03
4	CORJ870106	0.8665019853	0.0284243311	0.3396632427	1131	-3.89	-3.89
4	CORJ870107	0.9275582897	0.0083134493	0.1841273373	1177	-0.26	-0.56
4	CORJ870108	0.9156706582	0.0112770259	0.1276004532	824	1.33	1.37
4	MIYS990101	0.3568541247	0.8579220133	0.3396551068	894	0.44	0.44
4	MIYS990102	0.3686167507	0.8166907577	0.3232521411	890	0.07	0.07
4	MIYS990103	0.8533546735	0.034367134	0.385462421	905	0.09	0.09
4	MIYS990104	0.941578121	0.0054016819	0.5014181506	929	0.07	0.07
4	MIYS990105	0.588036554	0.2955877474	0.3278686566	891.5	0.08	0.08
4	FASG890101	0.4581031216	0.5554058603	0.5764089976	1082	0.75	0.75

Appendix D: Physicochemical Profiles

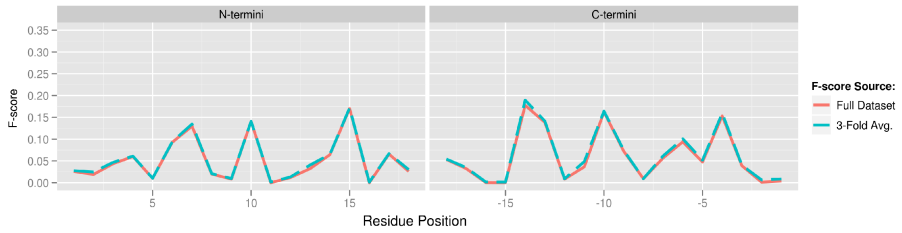


Figure D.1: Termini Profile for AAIndex ID: ARGP820101: Hydrophobicity index (Argos et al., 1982)

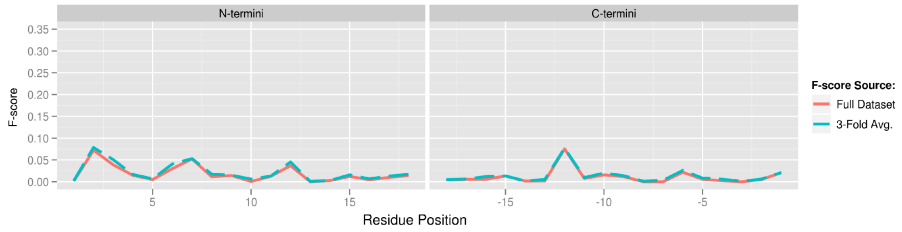


Figure D.2: Termini Profile for AAIndex ID: ARGP820102: Signal sequence helical potential (Argos et al., 1982)

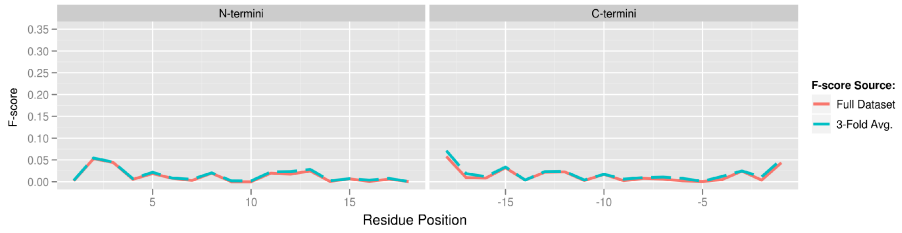


Figure D.3: Termini Profile for AAIndex ID: AURR980110: Normalized positional residue frequency at helix termini N5 (Aurora-Rose, 1998)

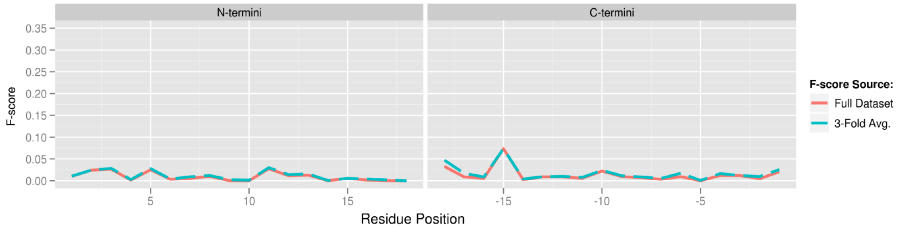


Figure D.4: Termini Profile for AAIndex ID: AURR980112: Normalized positional residue frequency at helix termini C4 (Aurora-Rose, 1998)

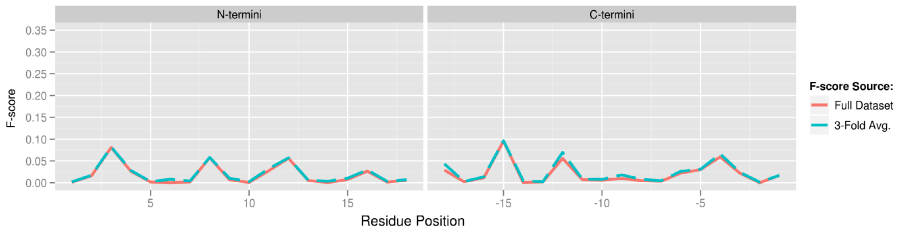


Figure D.5: Termini Profile for AAIndex ID: AURR980116: Normalized positional residue frequency at helix termini Cc (Aurora-Rose, 1998)

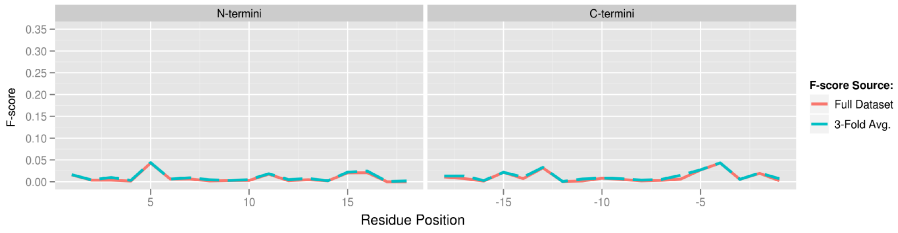


Figure D.6: Termini Profile for AAIndex ID: AURR980117: Normalized positional residue frequency at helix termini C' (Aurora-Rose, 1998)

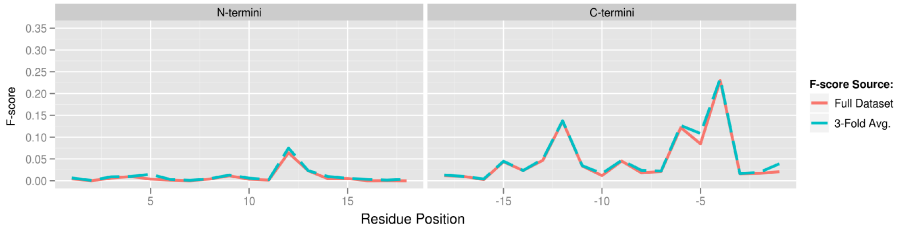


Figure D.7: Termini Profile for AAIndex ID: AURR980119: Normalized positional residue frequency at helix termini C"' (Aurora-Rose, 1998)

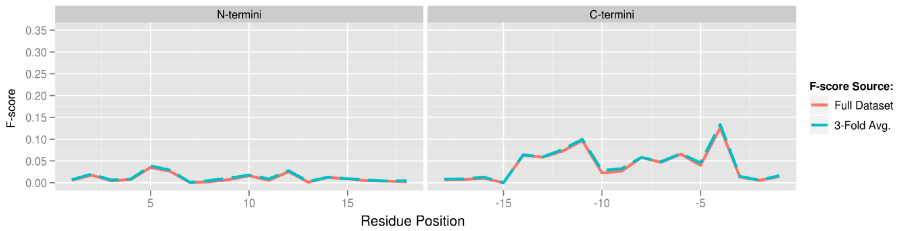


Figure D.8: Termini Profile for AAIndex ID: AURR980120: Normalized positional residue frequency at helix termini C4' (Aurora-Rose, 1998)

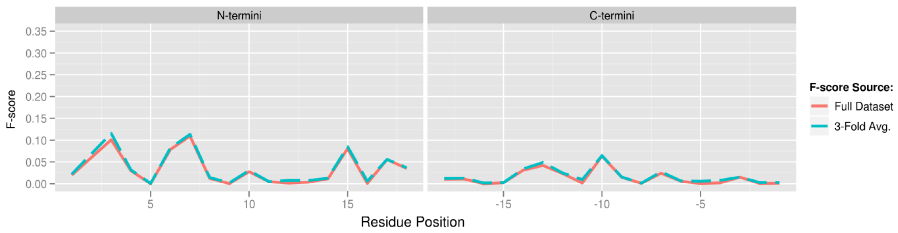


Figure D.9: Termini Profile for AAIndex ID: BASU050102: Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)

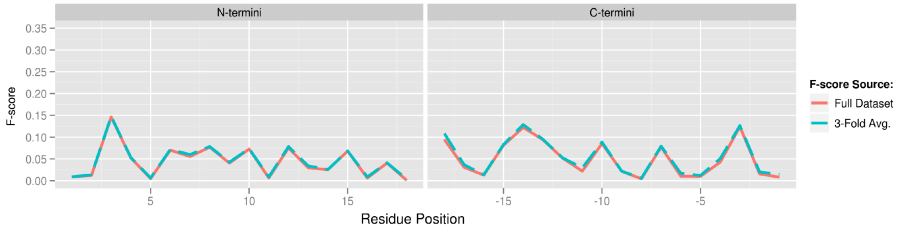


Figure D.10: Termini Profile for AAIndex ID: BIGC670101: Residue volume (Bigelow, 1967)

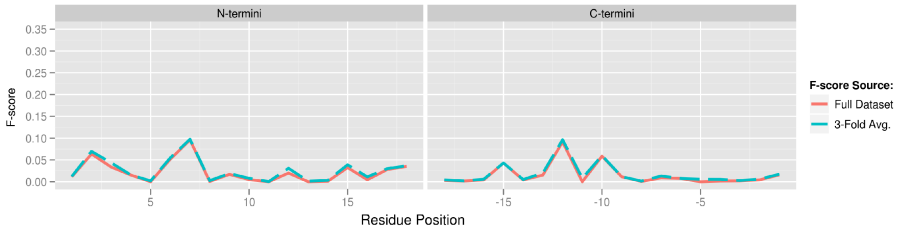


Figure D.11: Termini Profile for AAIndex ID: BIOV880101: Information value for accessibility; average fraction 35% (Biou et al., 1988)

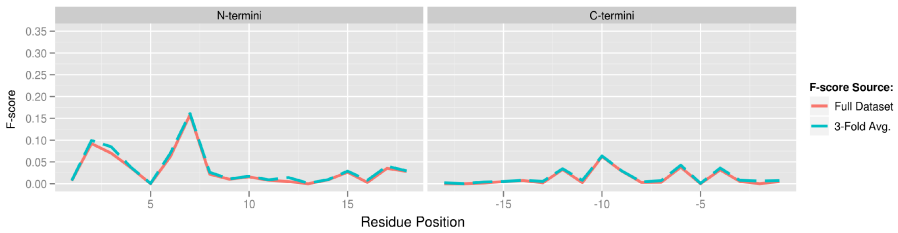


Figure D.12: Termini Profile for AAIndex ID: BROC820101: Retention coefficient in TFA (Browne et al., 1982)

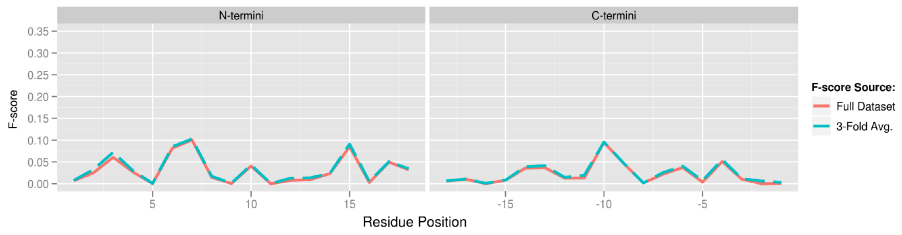


Figure D.13: Termini Profile for AAIndex ID: BULH740101: Transfer free energy to surface (Bull-Breese, 1974)

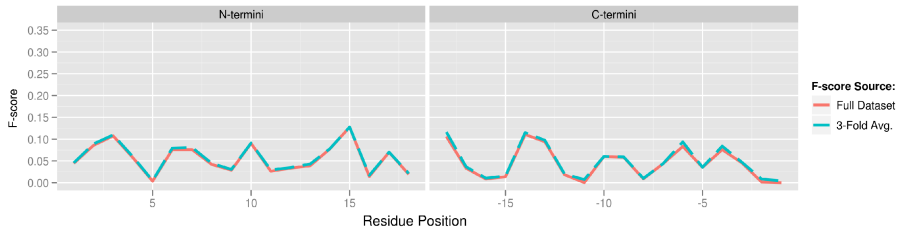


Figure D.14: Termini Profile for AAIndex ID: BULH740102: Apparent partial specific volume (Bull-Breese, 1974)

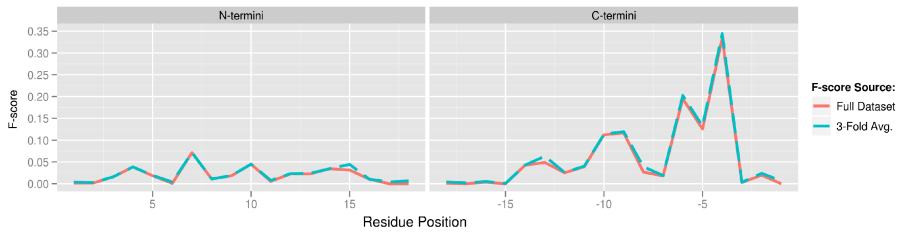


Figure D.15: Termini Profile for AAIndex ID: BUNA790101: alpha-NH chemical shifts (Bundi-Wuthrich, 1979)

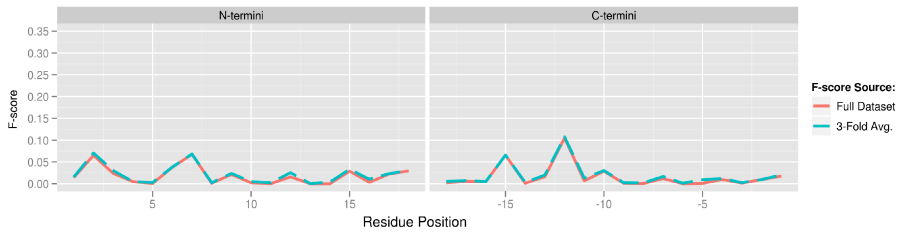


Figure D.16: Termini Profile for AAIndex ID: CASG920101: Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)

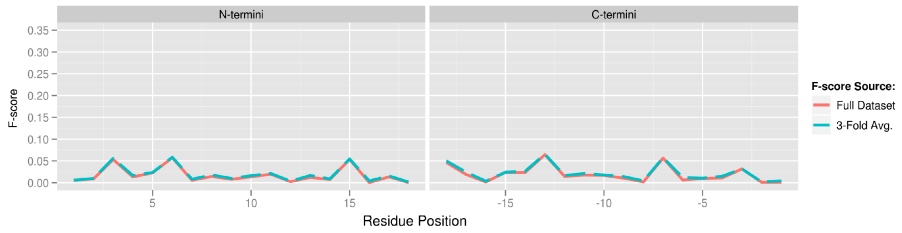


Figure D.17: Termini Profile for AAIndex ID: CHAM810101: Steric parameter (Charton, 1981)

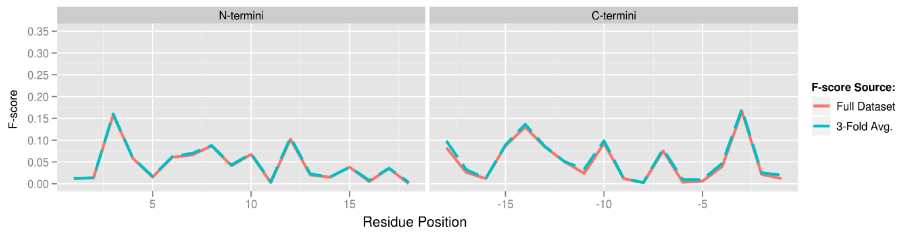


Figure D.18: Termini Profile for AAIndex ID: CHAM820101: Polarizability parameter (Charton-Charton, 1982)

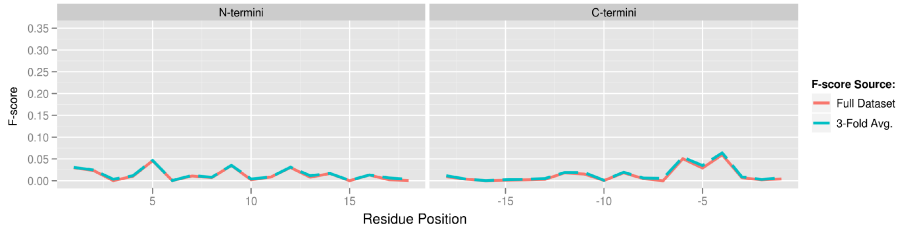


Figure D.19: Termini Profile for AAIndex ID: CHAM820102: Free energy of solution in water, kcal/mole (Charton-Charton, 1982)

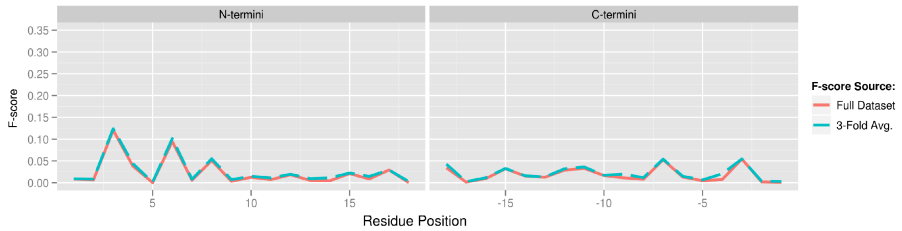


Figure D.20: Termini Profile for AAIndex ID: CHAM830104: The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)

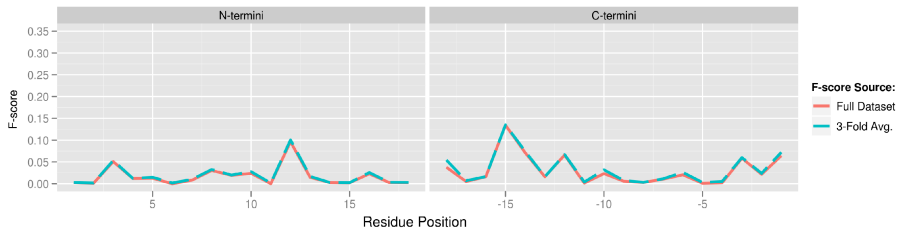


Figure D.21: Termini Profile for AAIndex ID: CHAM830105: The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)

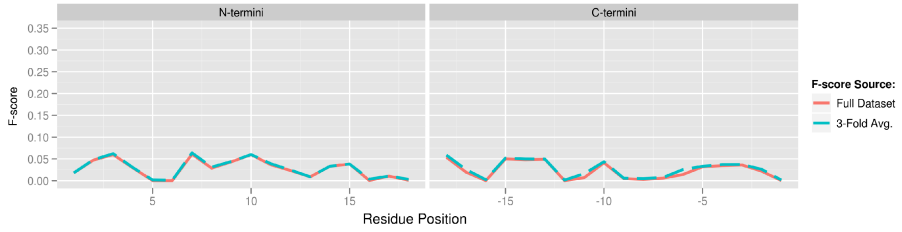


Figure D.22: Termini Profile for AAIndex ID: CHAM830107: A parameter of charge transfer capability (Charton-Charton, 1983)

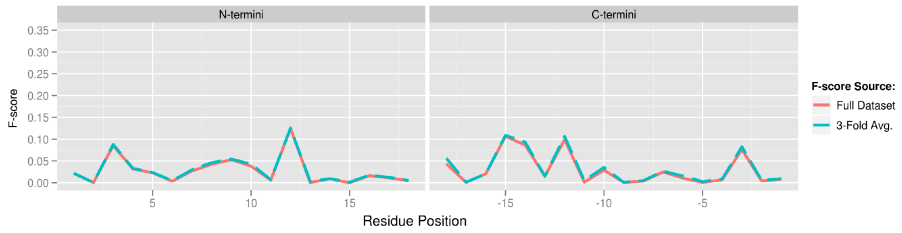


Figure D.23: Termini Profile for AAIndex ID: CHAM830108: A parameter of charge transfer donor capability (Charton-Charton, 1983)

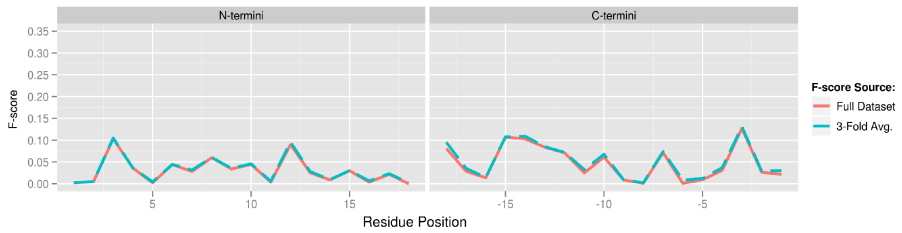


Figure D.24: Termini Profile for AAIndex ID: CHOC760101: Residue accessible surface area in tripeptide (Chothia, 1976)

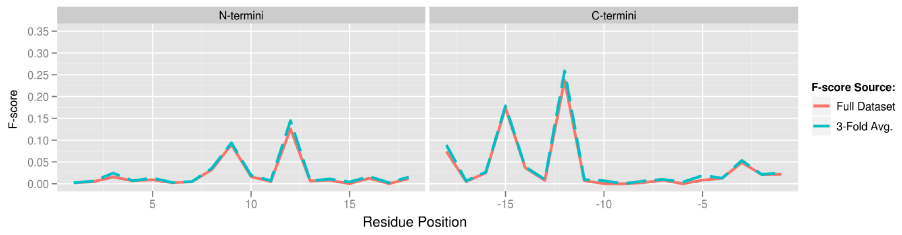


Figure D.25: Termini Profile for AAIndex ID: CHOC760102: Residue accessible surface area in folded protein (Chothia, 1976)

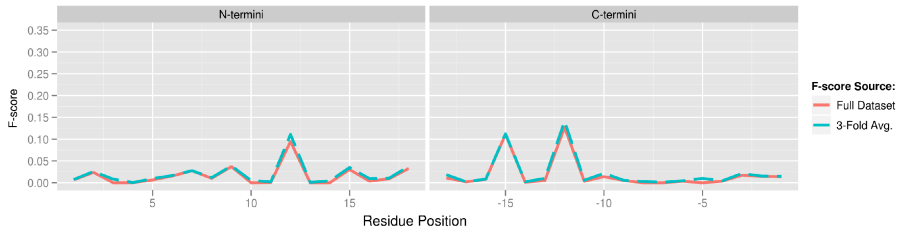


Figure D.26: Termini Profile for AAIndex ID: CHOC760103: Proportion of residues 95% buried (Chothia, 1976)

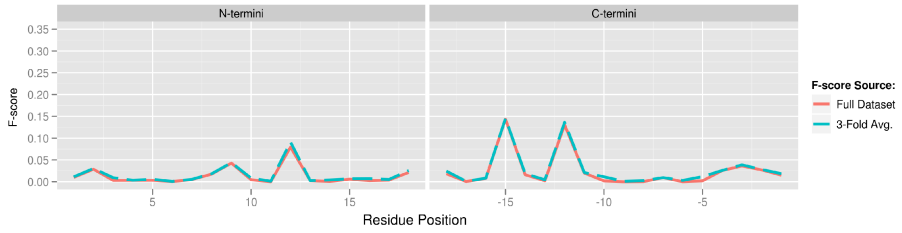


Figure D.27: Termini Profile for AAIndex ID: CHOC760104: Proportion of residues 100% buried (Chothia, 1976)

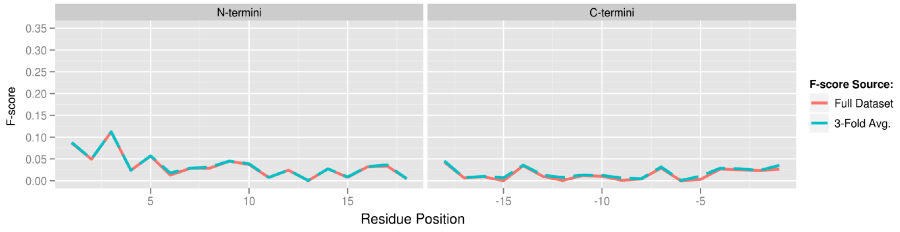


Figure D.28: Termini Profile for AAIndex ID: CHOP780204: Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)

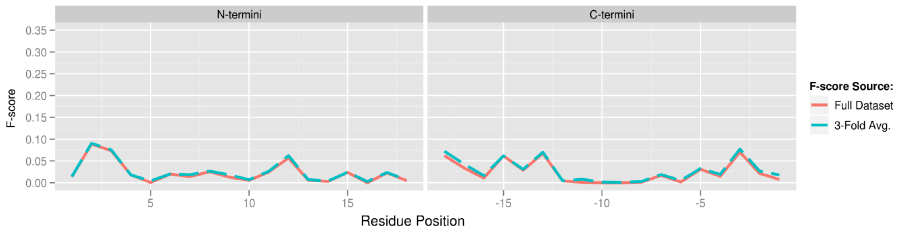


Figure D.29: Termini Profile for AAIndex ID: CHOP780206: Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)

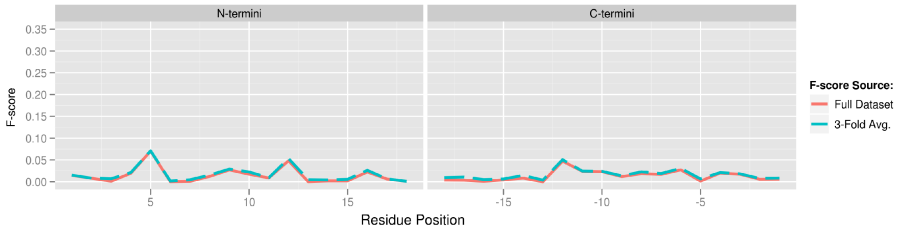


Figure D.30: Termini Profile for AAIndex ID: CHOP780207: Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)

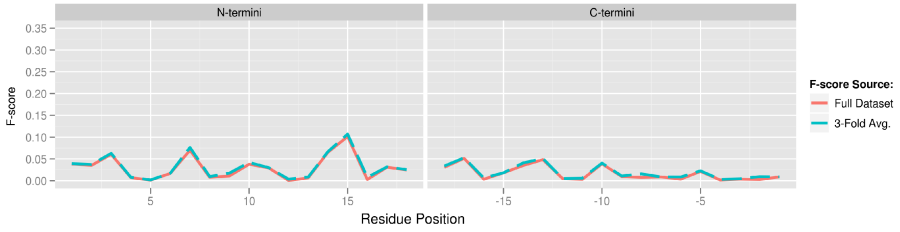


Figure D.31: Termini Profile for AAIndex ID: CHOP780208: Normalized frequency of N-terminal beta-sheet (Chou-Fasman, 1978b)

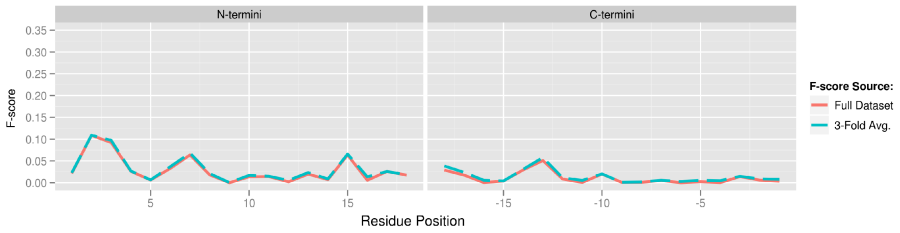


Figure D.32: Termini Profile for AAIndex ID: CHOP780210: Normalized frequency of N-terminal non beta region (Chou-Fasman, 1978b)

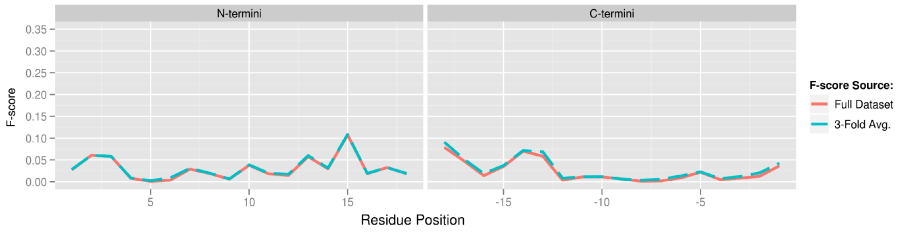


Figure D.33: Termini Profile for AAIndex ID: CHOP780212: Frequency of the 1st residue in turn (Chou-Fasman, 1978b)

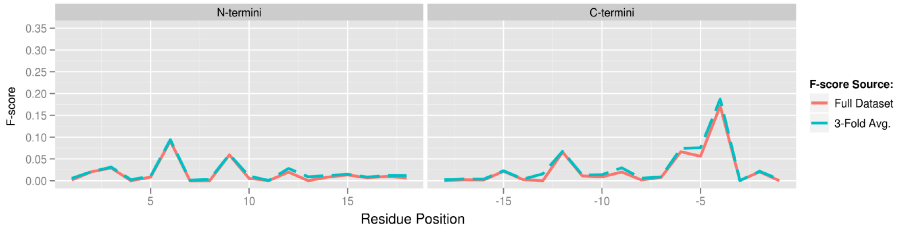


Figure D.34: Termini Profile for AAIndex ID: CHOP780213: Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)

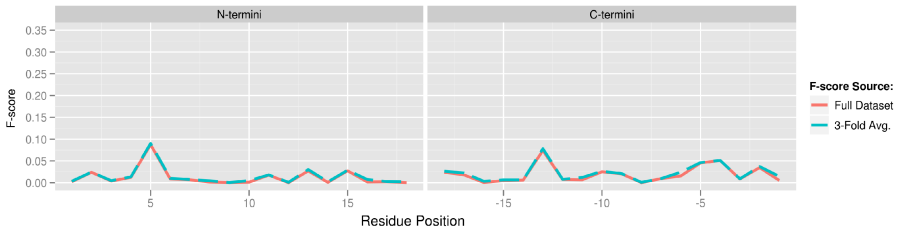


Figure D.35: Termini Profile for AAIndex ID: CHOP780215: Frequency of the 4th residue in turn (Chou-Fasman, 1978b)

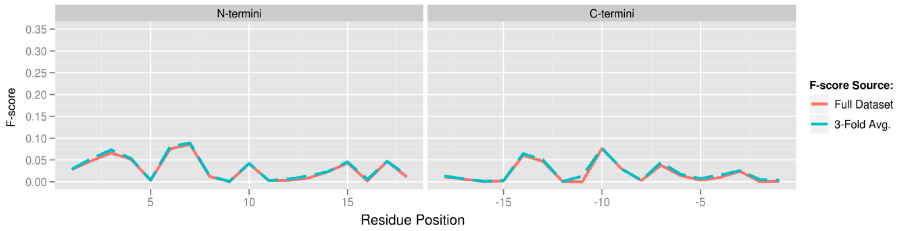


Figure D.36: Termini Profile for AAIndex ID: CIDH920101: Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)

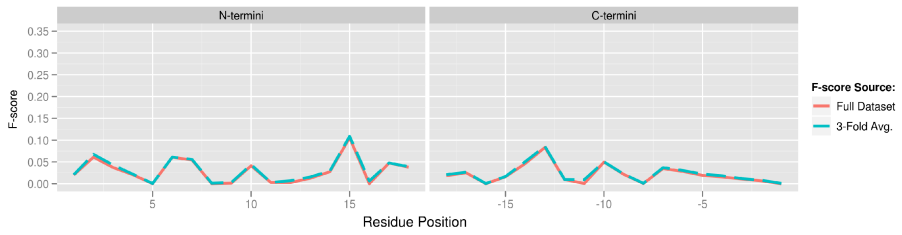


Figure D.37: Termini Profile for AAIndex ID: CIDH920103: Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)

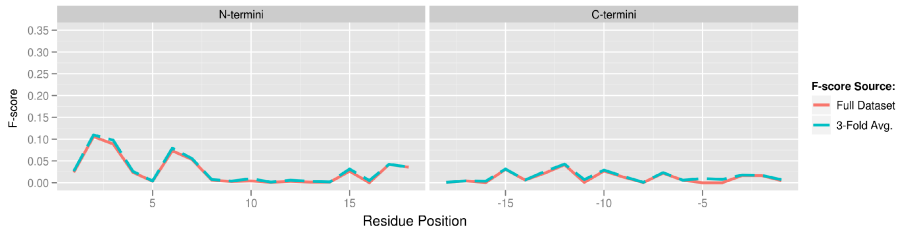


Figure D.38: Termini Profile for AAIndex ID: CORJ870103: PRIFT index (Cornette et al., 1987)

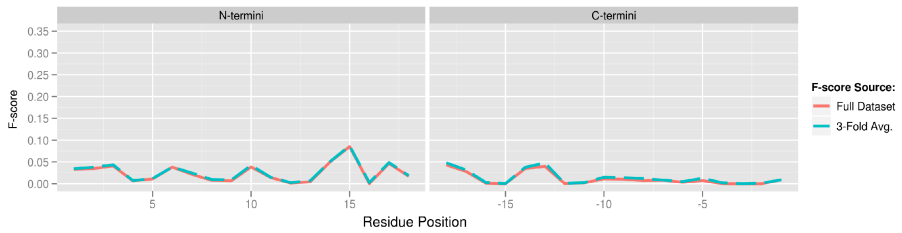


Figure D.39: Termini Profile for AAIndex ID: CRAJ730102: Normalized frequency of beta-sheet (Crawford et al., 1973)

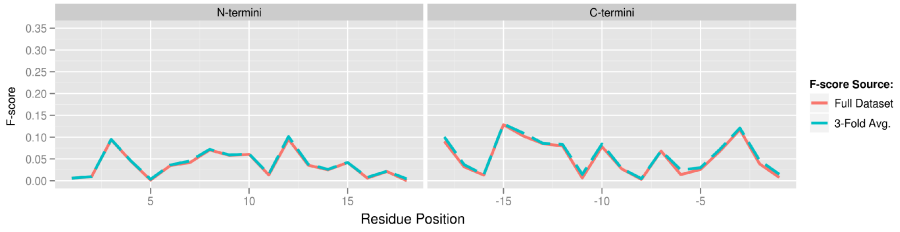


Figure D.40: Termini Profile for AAIndex ID: DAWD720101: Size (Dawson, 1972)

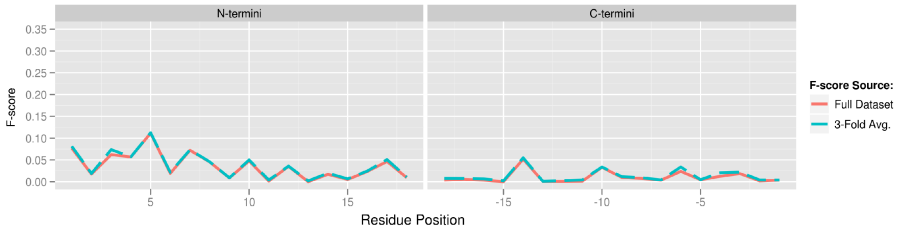


Figure D.41: Termini Profile for AAIndex ID: DAYM780201: Relative mutability (Dayhoff et al., 1978b)

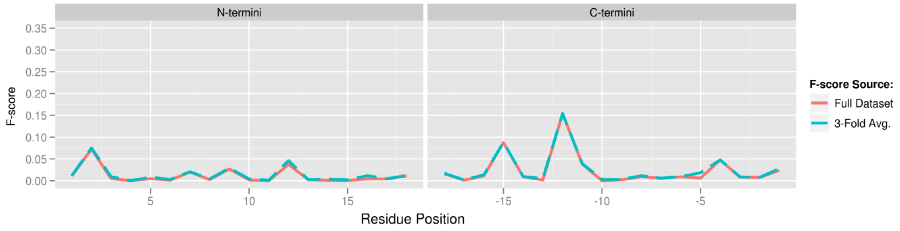


Figure D.42: Termini Profile for AAIndex ID: DESM900101: Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)

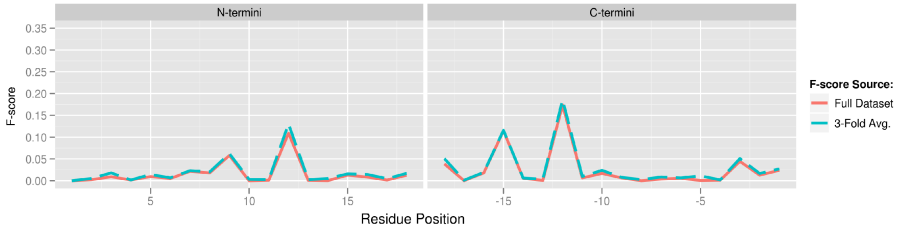


Figure D.43: Termini Profile for AAIndex ID: EISD840101: Consensus normalized hydrophobicity scale (Eisenberg, 1984)

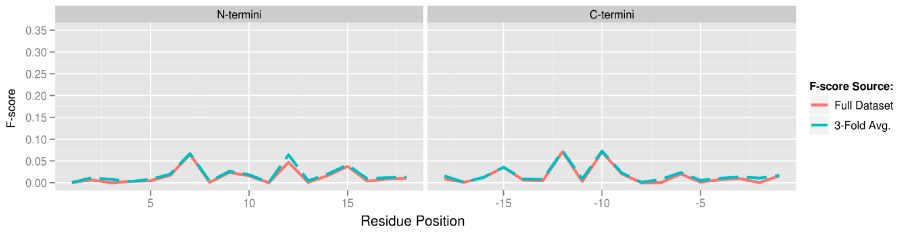


Figure D.44: Termini Profile for AAIndex ID: EISD860101: Solvation free energy (Eisenberg-McLachlan, 1986)

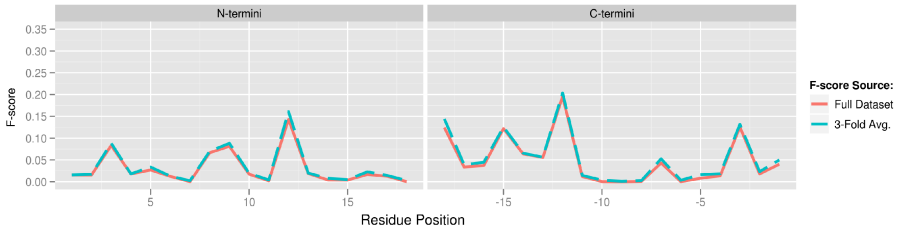


Figure D.45: Termini Profile for AAIndex ID: EISD860102: Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)

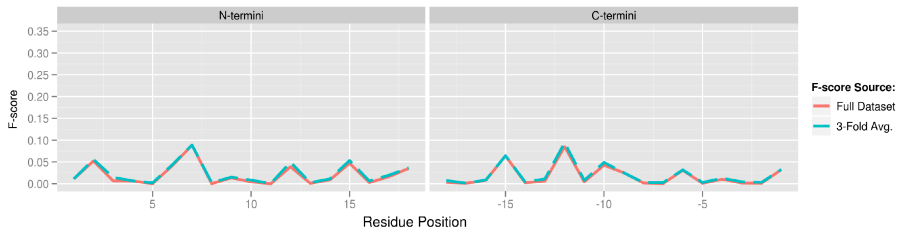


Figure D.46: Termini Profile for AAIndex ID: EISD860103: Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)

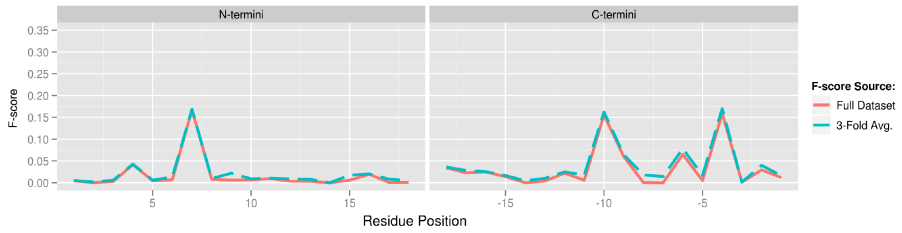


Figure D.47: Termini Profile for AAIndex ID: FASG760103: Optical rotation (Fasman, 1976)

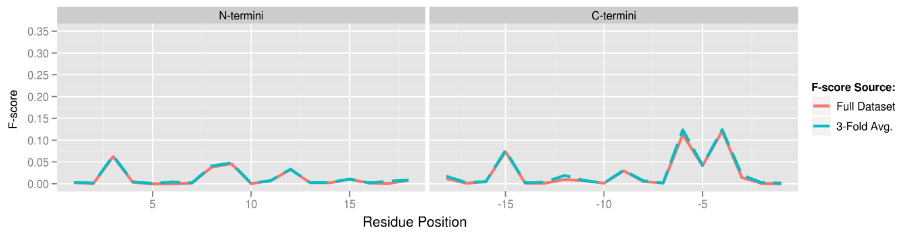


Figure D.48: Termini Profile for AAIndex ID: FASG760104: pK-N (Fasman, 1976)

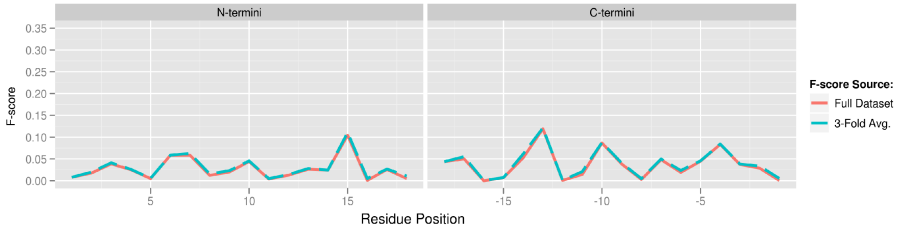


Figure D.49: Termini Profile for AAIndex ID: FAUJ880101: Graph shape index (Fauchere et al., 1988)

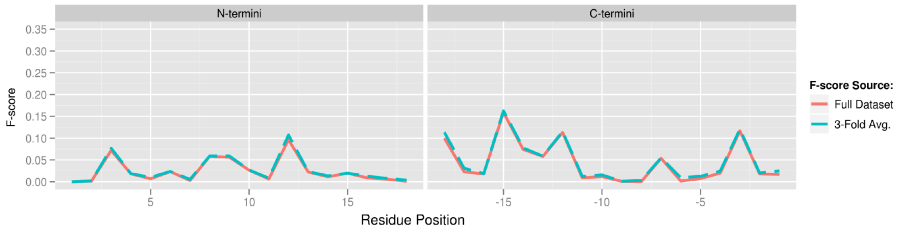


Figure D.50: Termini Profile for AAIndex ID: FAUJ880104: STERIMOL length of the side chain (Fauchere et al., 1988)

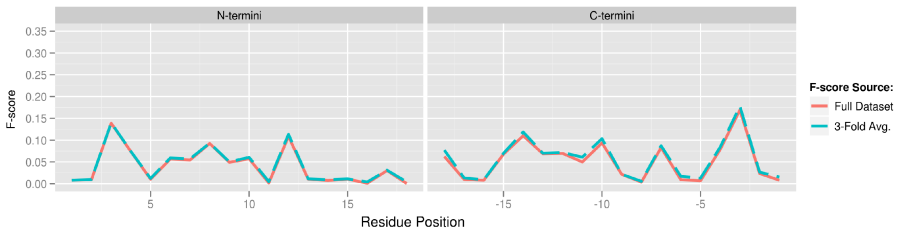


Figure D.51: Termini Profile for AAIndex ID: FAUJ880106: STERIMOL maximum width of the side chain (Fauchere et al., 1988)

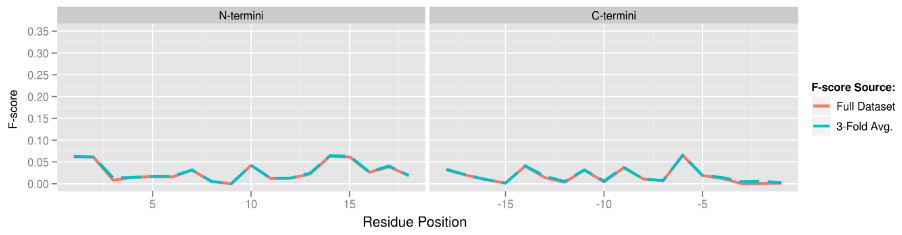


Figure D.52: Termini Profile for AAIndex ID: FAUJ880108: Localized electrical effect (Fauchere et al., 1988)

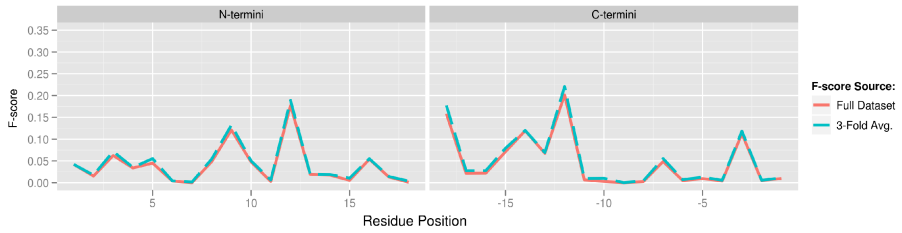


Figure D.53: Termini Profile for AAIndex ID: FAUJ880111: Positive charge (Fauchere et al., 1988)

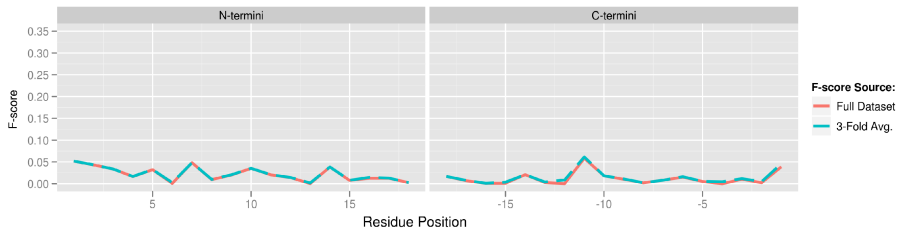


Figure D.54: Termini Profile for AAIndex ID: FAUJ880112: Negative charge (Fauchere et al., 1988)

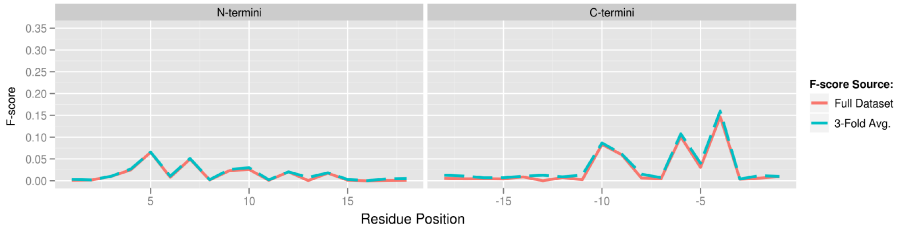


Figure D.55: Termini Profile for AAIndex ID: FAUJ880113: pK-a(RCOOH) (Fauchere et al., 1988)

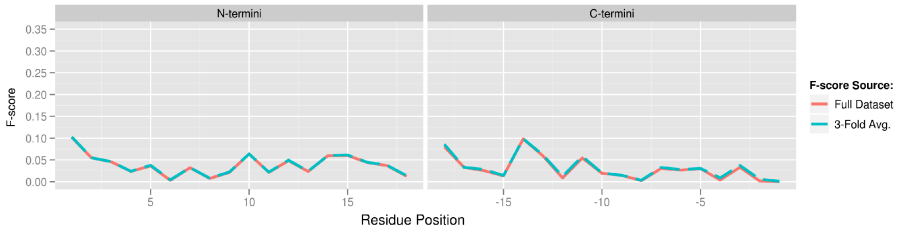


Figure D.56: Termini Profile for AAIndex ID: FINA910101: Helix initiation parameter at posision i-1 (Finkelstein et al., 1991)

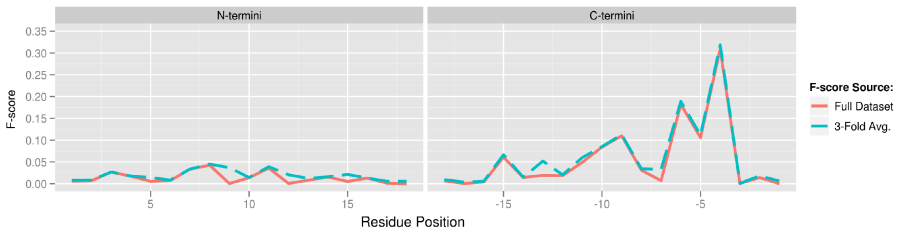


Figure D.57: Termini Profile for AAIndex ID: FINA910102: Helix initiation parameter at posision i,i+1,i+2 (Finkelstein et al., 1991)

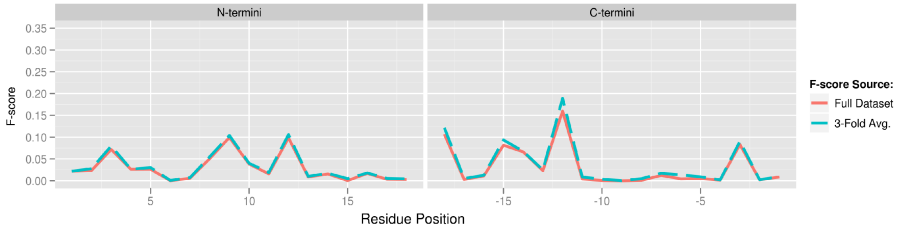


Figure D.58: Termini Profile for AAIndex ID: FINA910103: Helix termination parameter at posision j-2,j-1,j (Finkelstein et al., 1991)

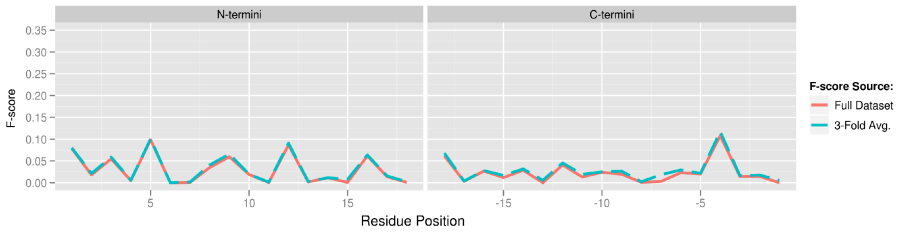


Figure D.59: Termini Profile for AAIndex ID: FINA910104: Helix termination parameter at posision j+1 (Finkelstein et al., 1991)

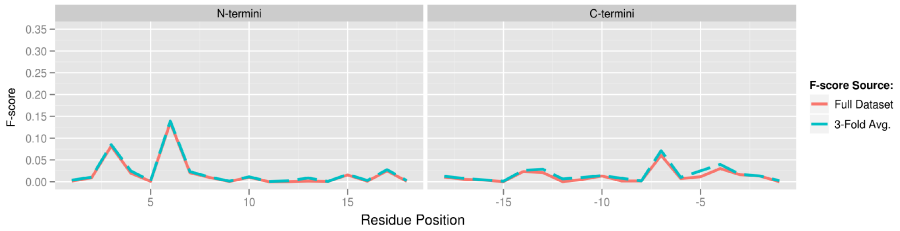


Figure D.60: Termini Profile for AAIndex ID: FODM020101: Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)

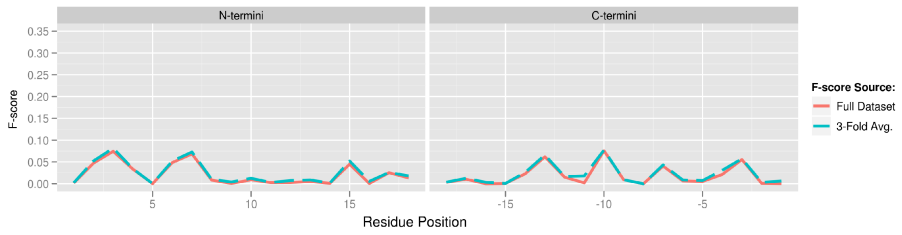


Figure D.61: Termini Profile for AAIndex ID: FUKS010103: Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)

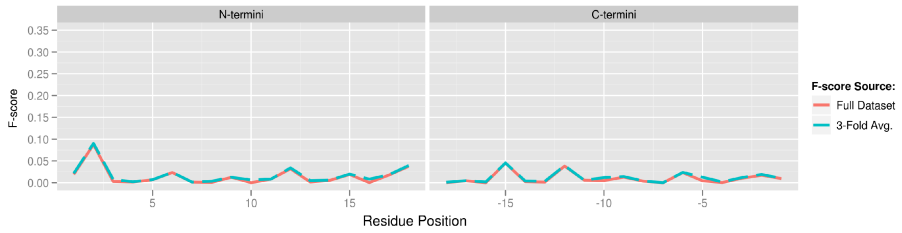


Figure D.62: Termini Profile for AAIndex ID: FUKS010105: Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)

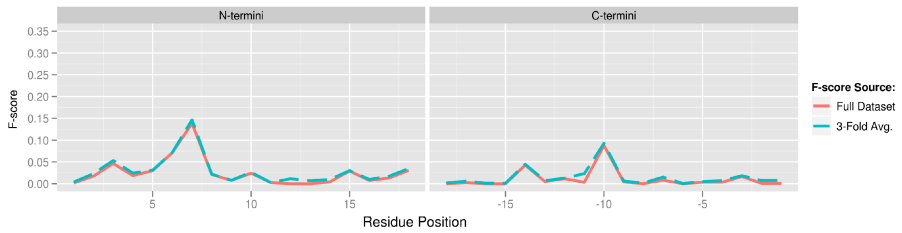


Figure D.63: Termini Profile for AAIndex ID: GARJ730101: Partition coefficient (Garel et al., 1973)

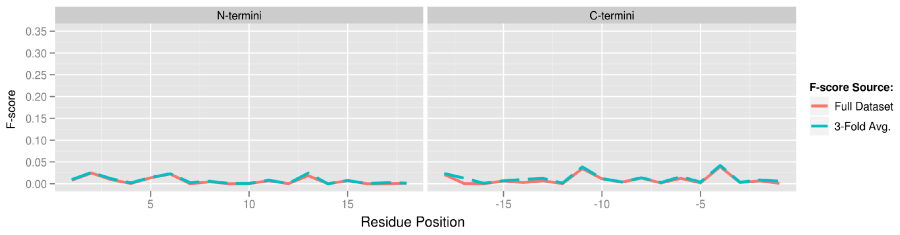


Figure D.64: Termini Profile for AAIndex ID: GEIM800102: Alpha-helix indices for alpha-proteins (Geisow-Roberts, 1980)

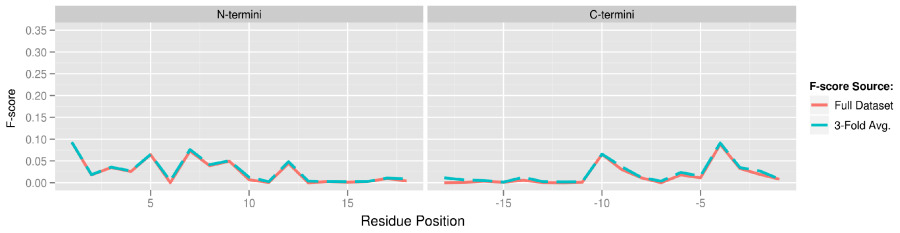


Figure D.65: Termini Profile for AAIndex ID: GEIM800103: Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)

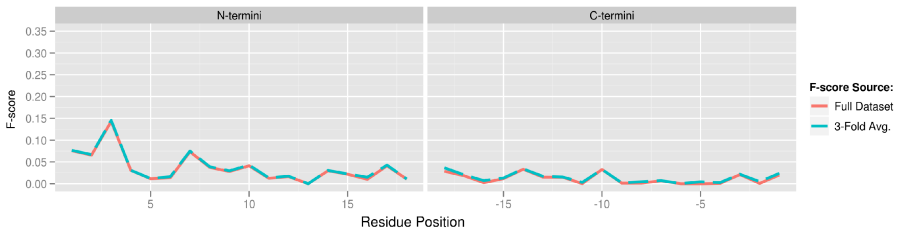


Figure D.66: Termini Profile for AAIndex ID: GEIM800106: Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)

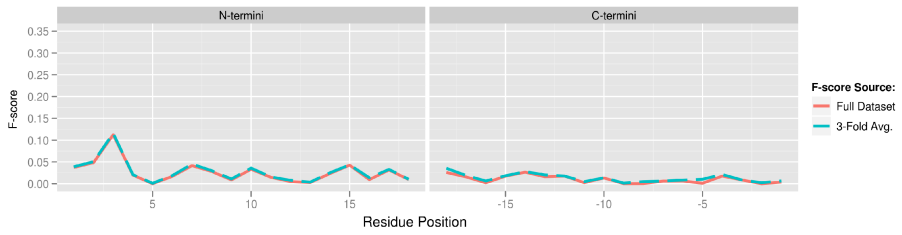


Figure D.67: Termini Profile for AAIndex ID: GEIM800110: Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)

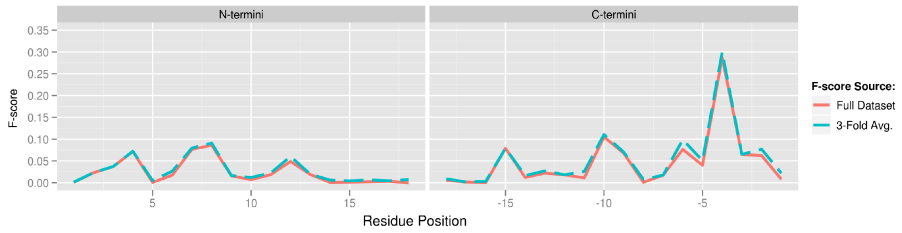


Figure D.68: Termini Profile for AAIndex ID: GEOR030101: Linker propensity from all dataset (George-Heringa, 2003)

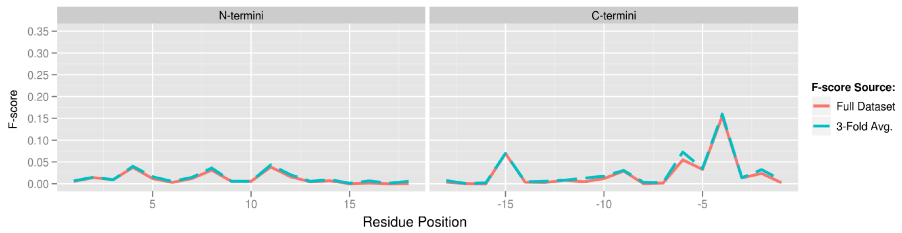


Figure D.69: Termini Profile for AAIndex ID: GEOR030104: Linker propensity from 3-linker dataset (George-Heringa, 2003)

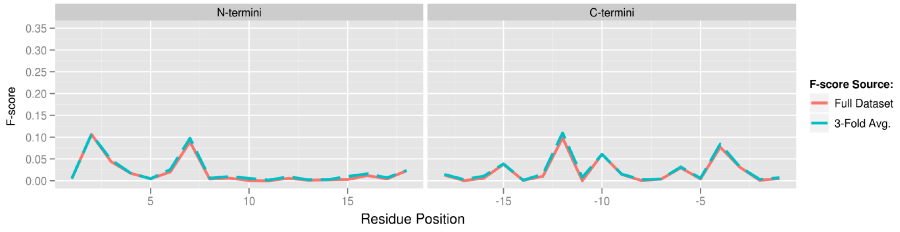


Figure D.70: Termini Profile for AAIndex ID: GEOR030105: Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)

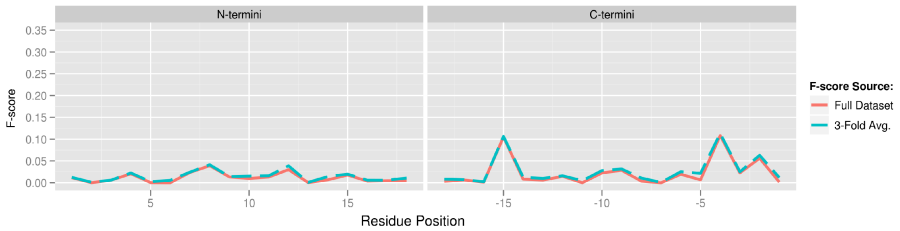


Figure D.71: Termini Profile for AAIndex ID: GEOR030107: Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)

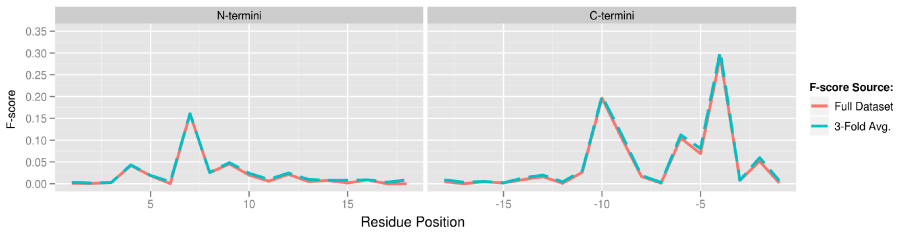


Figure D.72: Termini Profile for AAIndex ID: GEOR030109: Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)

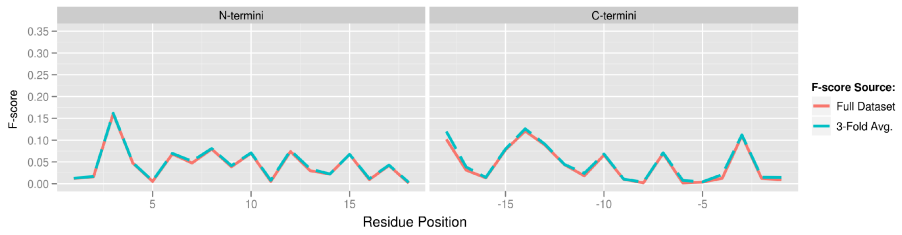


Figure D.73: Termini Profile for AAIndex ID: GRAR740103: Volume (Grantham, 1974)

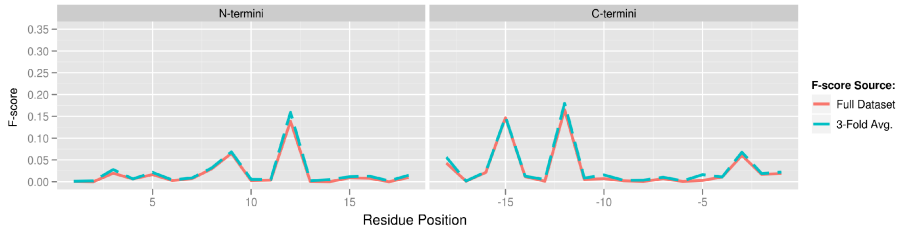


Figure D.74: Termini Profile for AAIndex ID: GUYH850105: Apparent partition energies calculated from Chothia index (Guy, 1985)

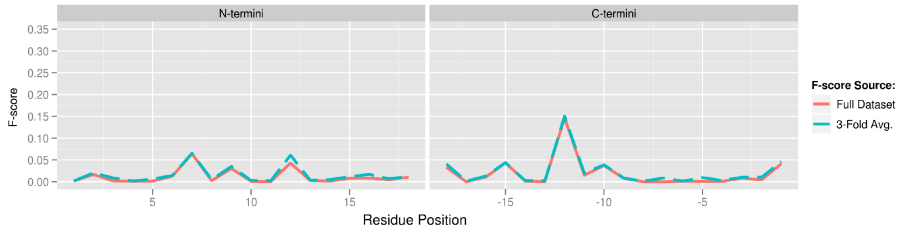


Figure D.75: Termini Profile for AAIndex ID: HOPT810101: Hydrophilicity value (Hopp-Woods, 1981)

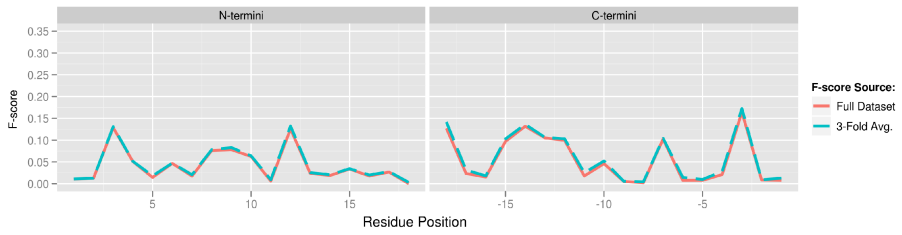


Figure D.76: Termini Profile for AAIndex ID: HUTJ700102: Absolute entropy (Hutchens, 1970)

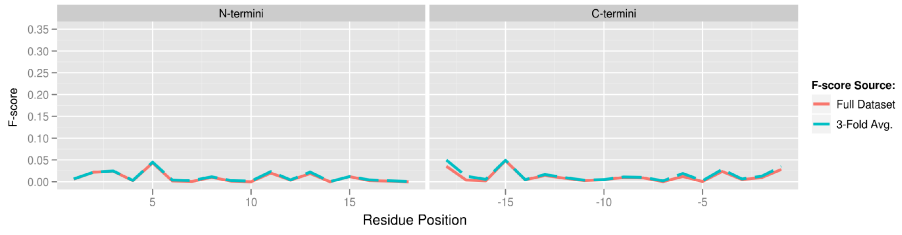


Figure D.77: Termini Profile for AAIndex ID: ISOY800101: Normalized relative frequency of alpha-helix (Isogai et al., 1980)

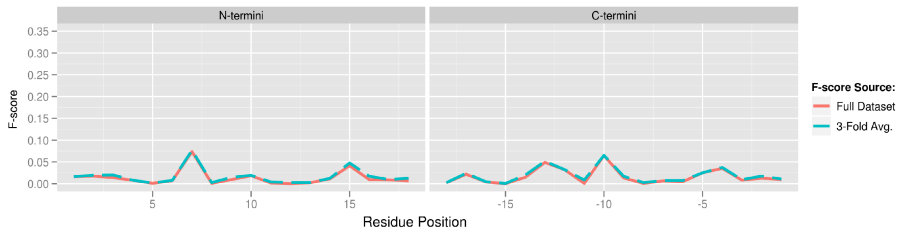


Figure D.78: Termini Profile for AAIndex ID: ISOY800102: Normalized relative frequency of extended structure (Isogai et al., 1980)

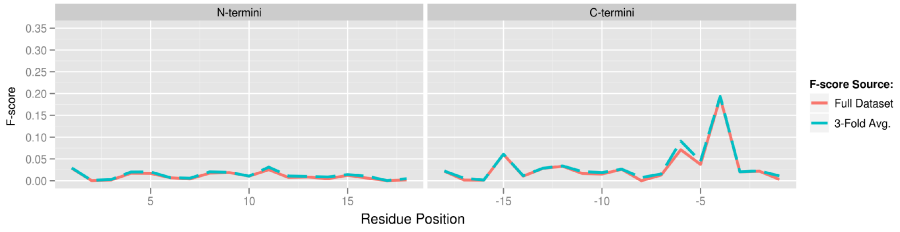


Figure D.79: Termini Profile for AAIndex ID: ISOY800106: Normalized relative frequency of helix end (Isogai et al., 1980)

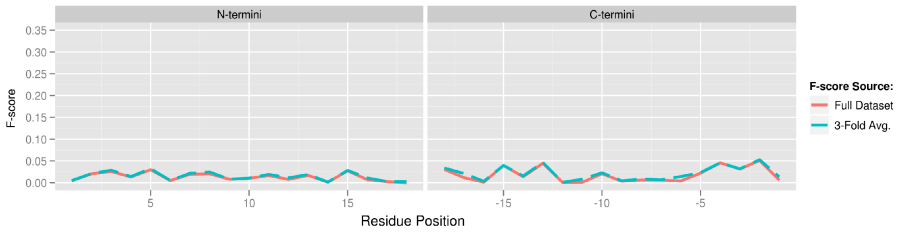


Figure D.80: Termini Profile for AAIndex ID: ISOY800108: Normalized relative frequency of coil (Isogai et al., 1980)

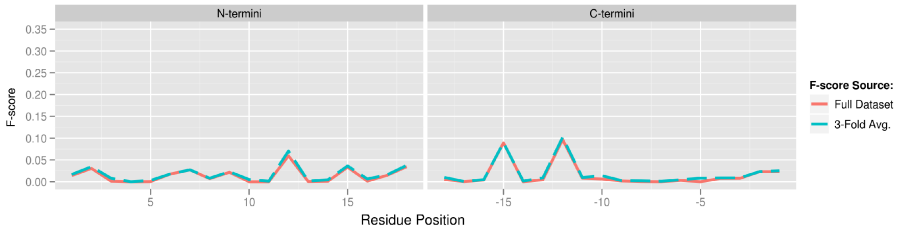


Figure D.81: Termini Profile for AAIndex ID: JANJ790101: Ratio of buried and accessible molar fractions (Janin, 1979)

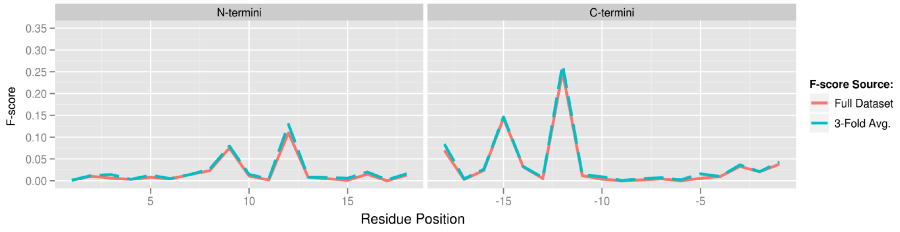


Figure D.82: Termini Profile for AAIndex ID: JANJ790102: Transfer free energy (Janin, 1979)

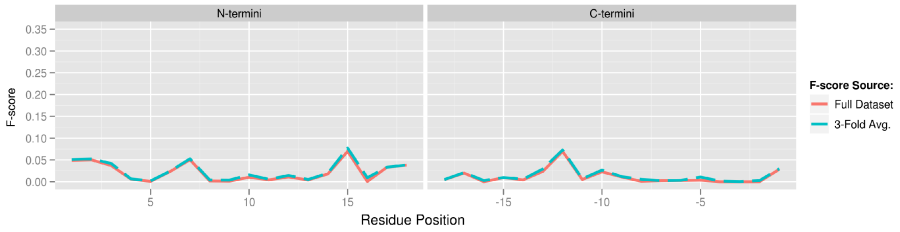


Figure D.83: Termini Profile for AAIndex ID: KANM800104: Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)

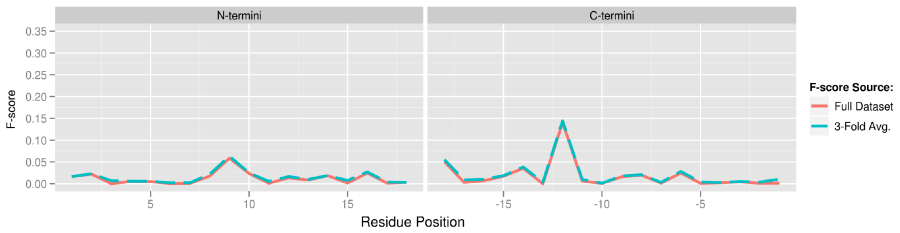


Figure D.84: Termini Profile for AAIndex ID: KARP850103: Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)

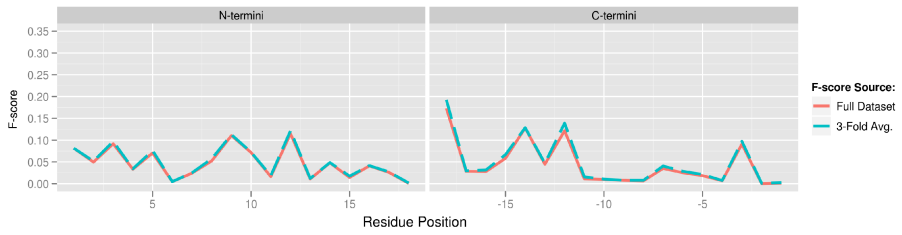


Figure D.85: Termini Profile for AAIndex ID: KLEP840101: Net charge (Klein et al., 1984)

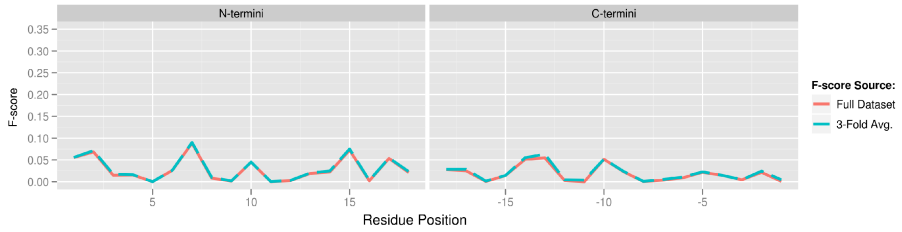


Figure D.86: Termini Profile for AAIndex ID: KOEP990102: Beta-sheet propensity derived from designed sequences (Koehl-Levitt, 1999)

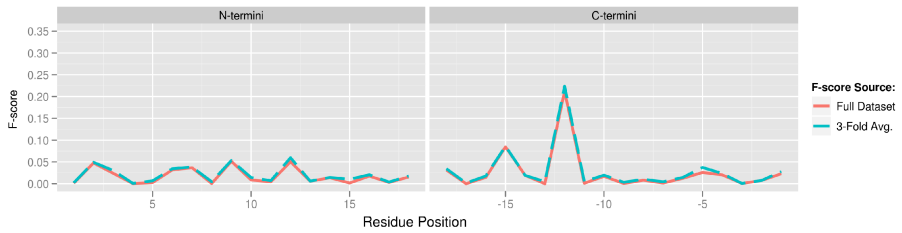


Figure D.87: Termini Profile for AAIndex ID: KRIW790102: Fraction of site occupied by water (Krigbaum-Komoriya, 1979)

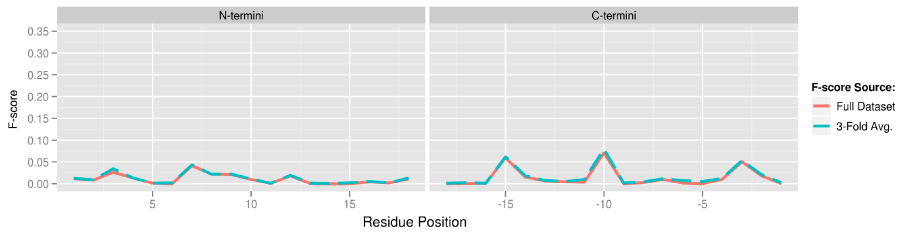


Figure D.88: Termini Profile for AAIndex ID: KUMS000101: Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)

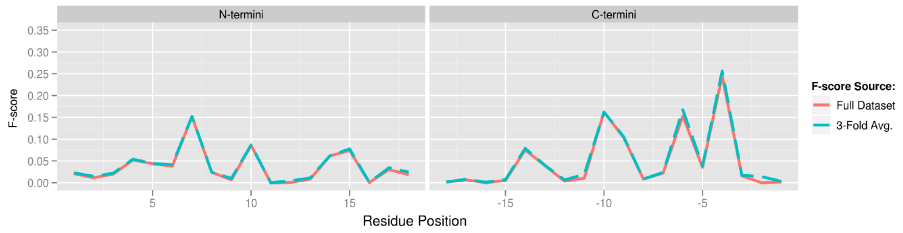


Figure D.89: Termini Profile for AAIndex ID: LAWE840101: Transfer free energy, CHP/water (Lawson et al., 1984)

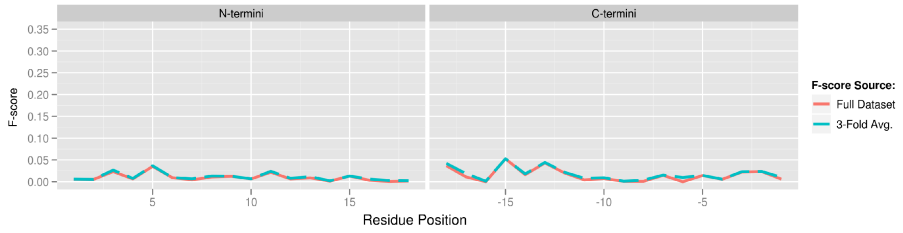


Figure D.90: Termini Profile for AAIndex ID: LEVM760103: Side chain angle theta(AAR) (Levitt, 1976)

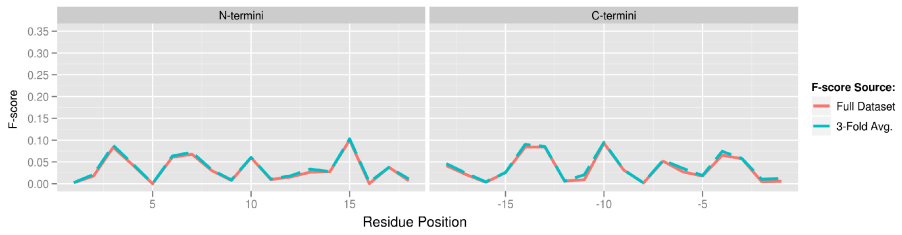


Figure D.91: Termini Profile for AAIndex ID: LEVM760106: van der Waals parameter R0 (Levitt, 1976)

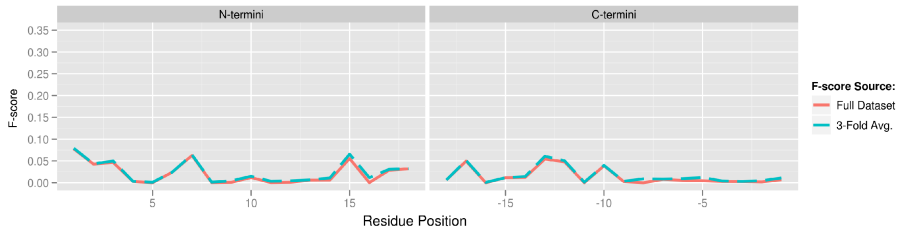


Figure D.92: Termini Profile for AAIndex ID: LEVM780102: Normalized frequency of beta-sheet, with weights (Levitt, 1978)

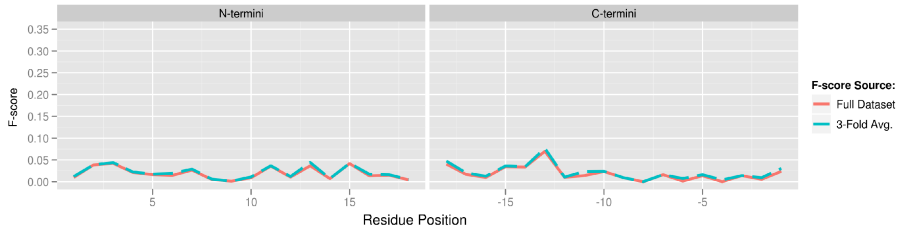


Figure D.93: Termini Profile for AAIndex ID: LEWP710101: Frequency of occurrence in beta-bends (Lewis et al., 1971)

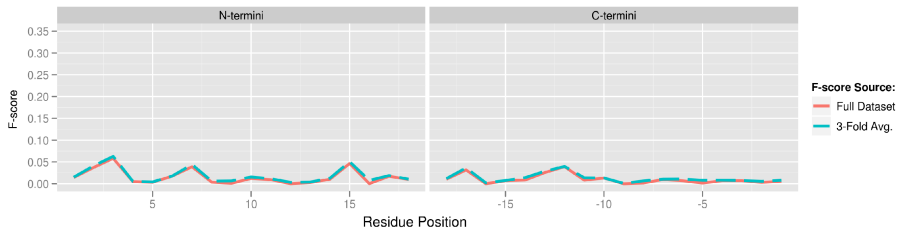


Figure D.94: Termini Profile for AAIndex ID: LIFS790103: Conformational preference for antiparallel beta-strands (Lifson-Sander, 1979)

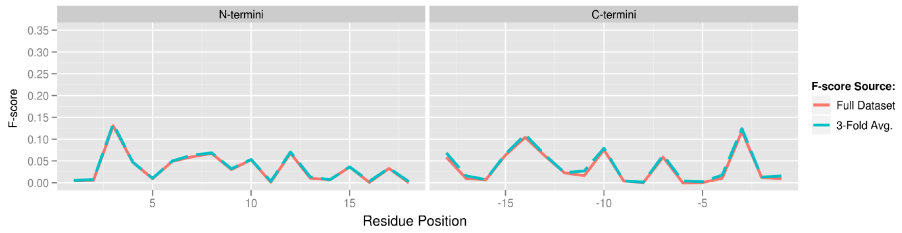


Figure D.95: Termini Profile for AAIndex ID: MCMT640101: Refractivity (McMeekin et al., 1964), Cited by Jones (1975)

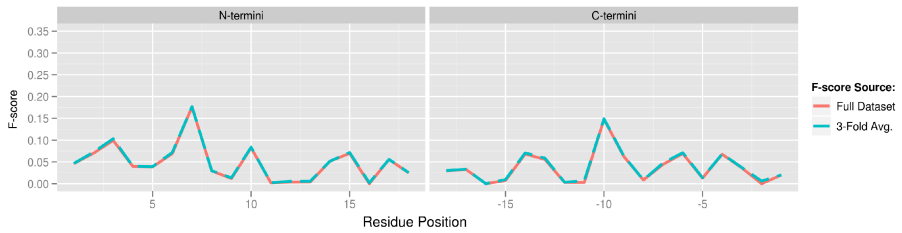


Figure D.96: Termini Profile for AAIndex ID: MEEJ800101: Retention coefficient in HPLC, pH7.4 (Meek, 1980)

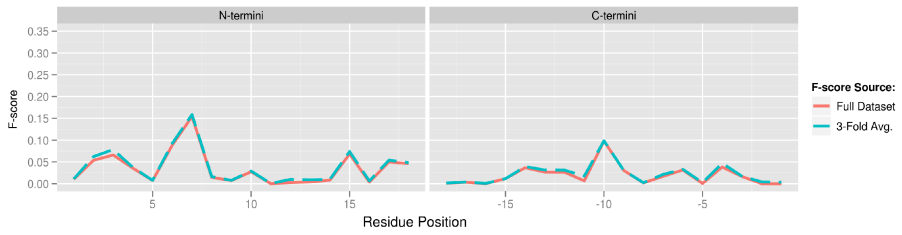


Figure D.97: Termini Profile for AAIndex ID: MEEJ810101: Retention coefficient in NaClO4 (Meek-Rossetti, 1981)

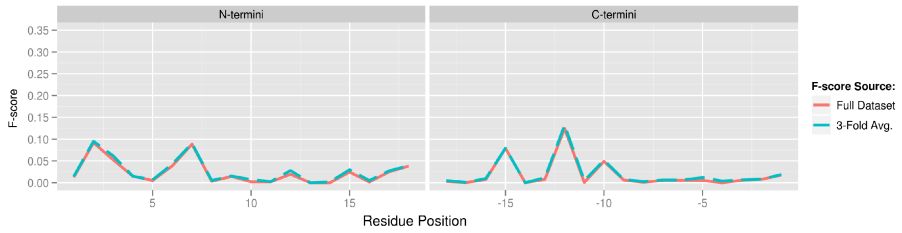


Figure D.98: Termini Profile for AAIndex ID: MEIH800103: Average side chain orientation angle (Meirovitch et al., 1980)

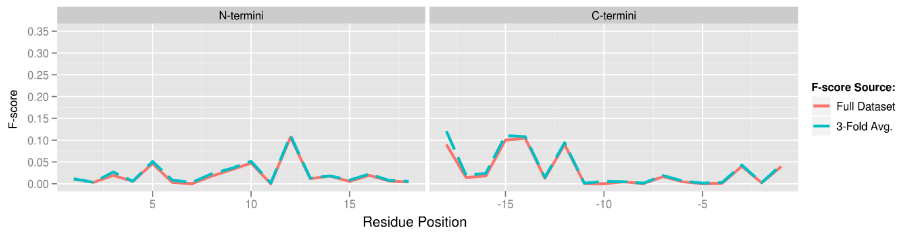


Figure D.99: Termini Profile for AAIndex ID: MITS020101: Amphiphilicity index (Mitaku et al., 2002)

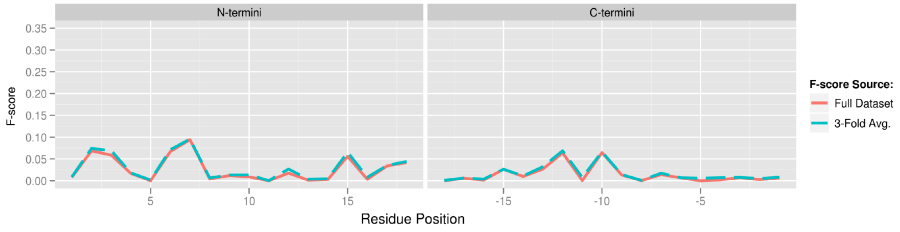


Figure D.100: Termini Profile for AAIndex ID: MIYS850101: Effective partition energy (Miyazawa-Jernigan, 1985)

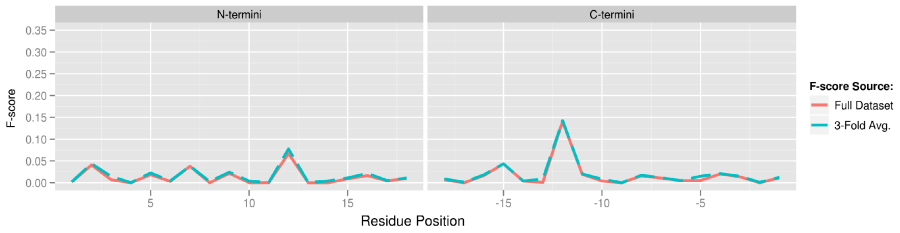


Figure D.101: Termini Profile for AAIndex ID: MONM990101: Turn propensity scale for transmembrane helices (Monne et al., 1999)

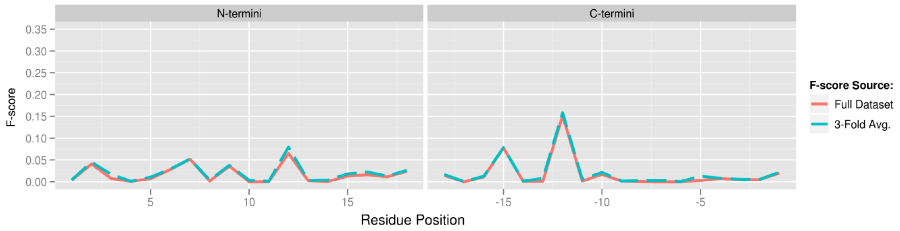


Figure D.102: Termini Profile for AAIndex ID: NADH010103: Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)

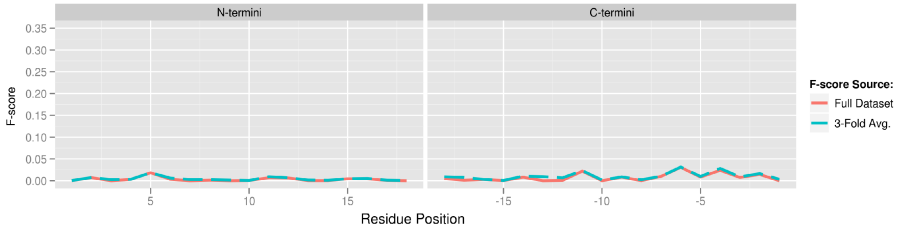


Figure D.103: Termini Profile for AAIndex ID: NADH010107: Hydropathy scale based on self-information values in the two-state model (50% accessibility) (Naderi-Manesh et al., 2001)

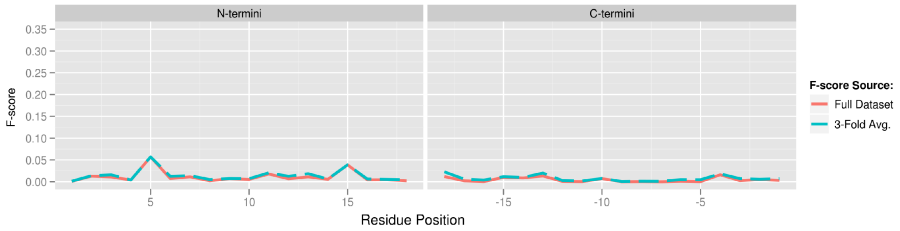


Figure D.104: Termini Profile for AAIndex ID: NAGK730103: Normalized frequency of coil (Nagano, 1973)

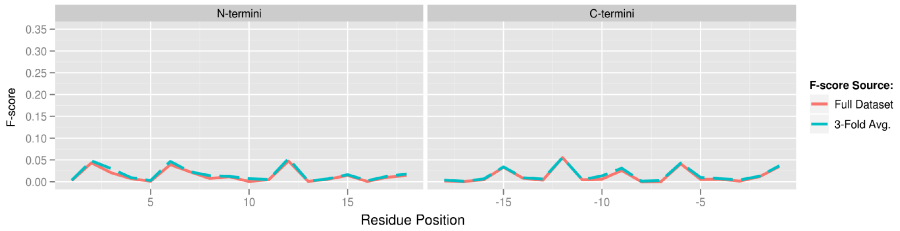


Figure D.105: Termini Profile for AAIndex ID: NAKH900103: AA composition of mt-proteins (Nakashima et al., 1990)

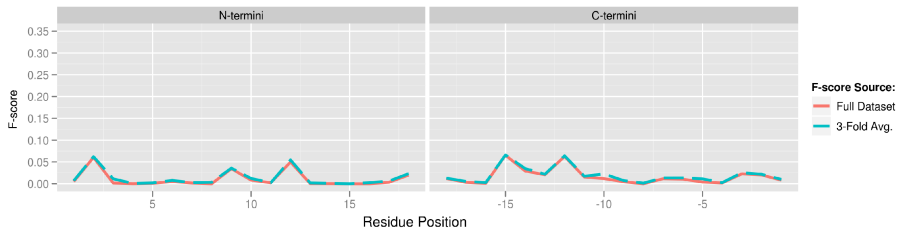


Figure D.106: Termini Profile for AAIndex ID: NAKH900109: AA composition of membrane proteins (Nakashima et al., 1990)

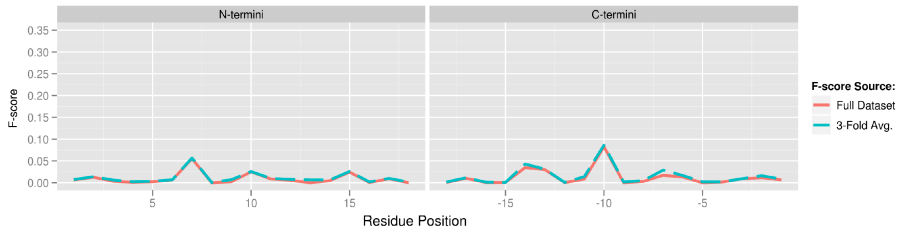


Figure D.107: Termini Profile for AAIndex ID: NAKH920101: AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)

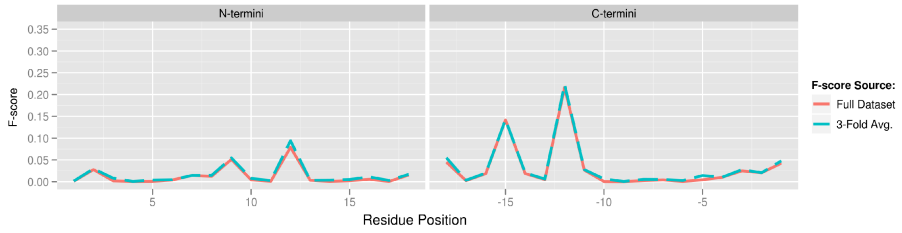


Figure D.108: Termini Profile for AAIndex ID: OOBM770101: Average non-bonded energy per atom (Oobatake-Ooi, 1977)

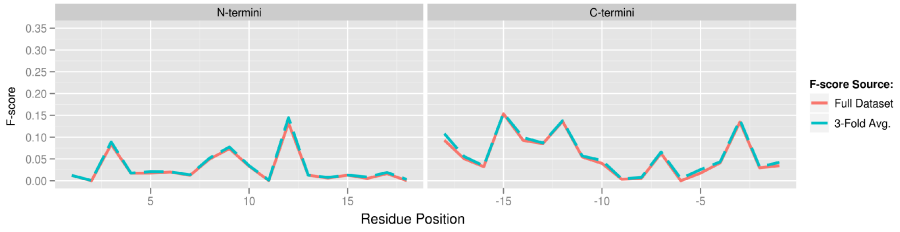


Figure D.109: Termini Profile for AAIndex ID: OOBM770102: Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)

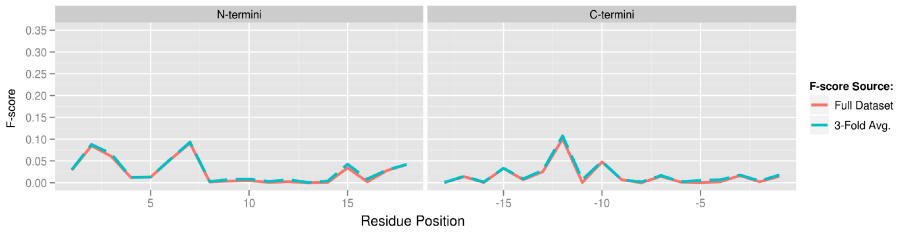


Figure D.110: Termini Profile for AAIndex ID: OOBM770103: Long range non-bonded energy per atom (Oobatake-Ooi, 1977)

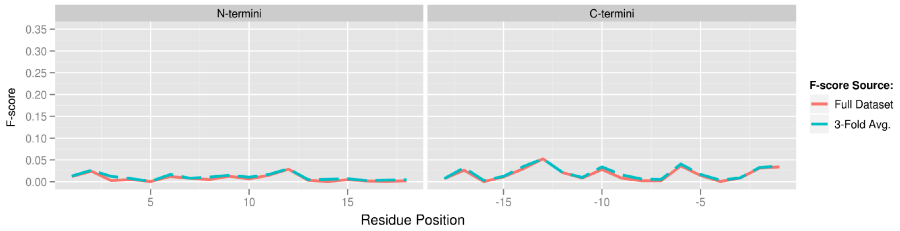


Figure D.111: Termini Profile for AAIndex ID: OOBM770104: Average non-bonded energy per residue (Oobatake-Ooi, 1977)

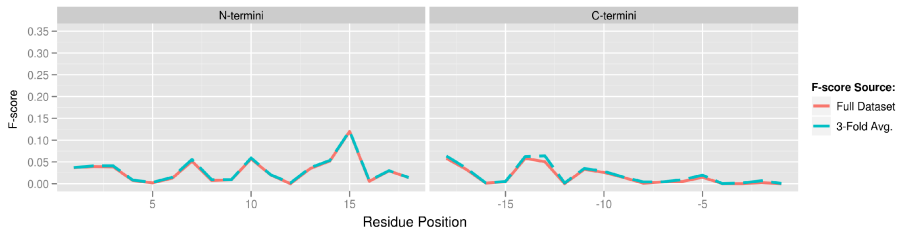


Figure D.112: Termini Profile for AAIndex ID: OOBM850101: Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)

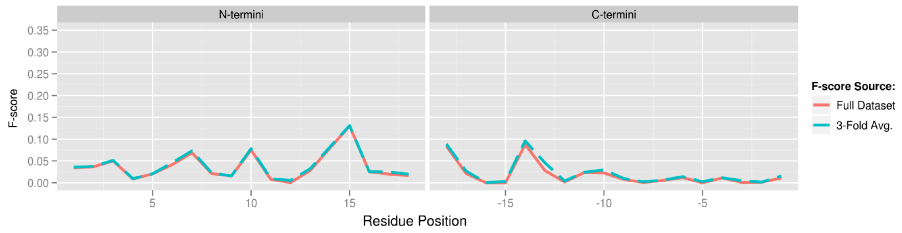


Figure D.113: Termini Profile for AAIndex ID: OOBM850104: Optimized average non-bonded energy per atom (Oobatake et al., 1985)

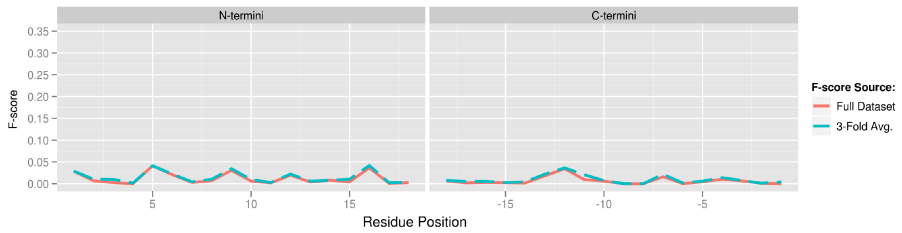


Figure D.114: Termini Profile for AAIndex ID: OOBM850105: Optimized side chain interaction parameter (Oobatake et al., 1985)

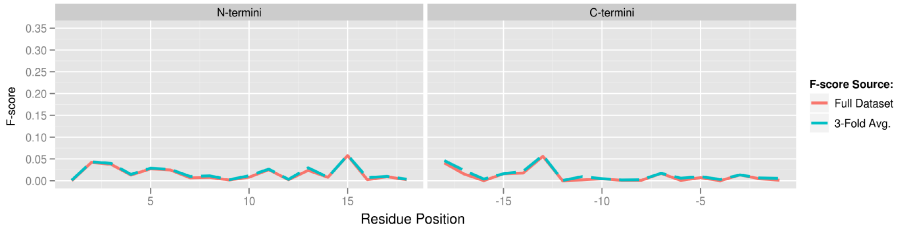


Figure D.115: Termini Profile for AAIndex ID: PALJ810105: Normalized frequency of turn from LG (Palau et al., 1981)

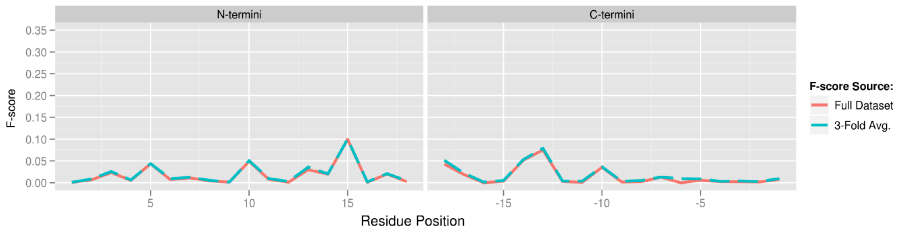


Figure D.116: Termini Profile for AAIndex ID: PALJ810114: Normalized frequency of turn in all-beta class (Palau et al., 1981)

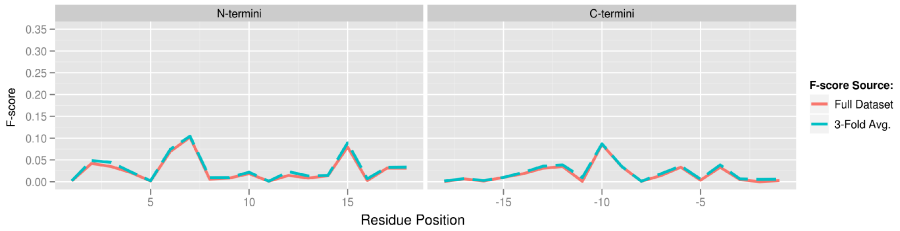


Figure D.117: Termini Profile for AAIndex ID: PLIV810101: Partition coefficient (Pliska et al., 1981)

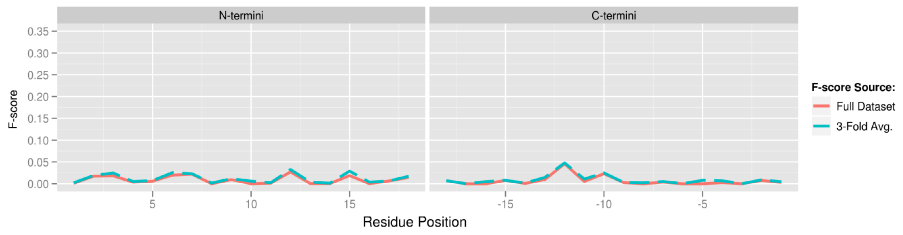


Figure D.118: Termini Profile for AAIndex ID: PONP800106: Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980)

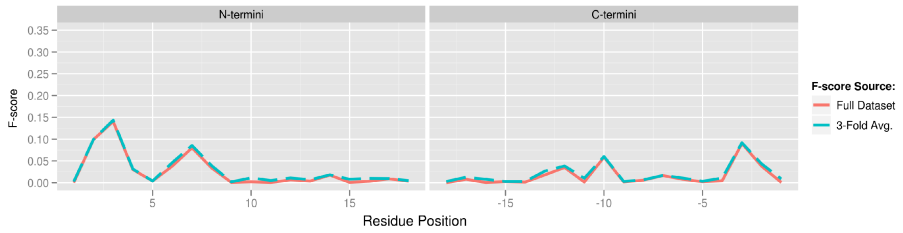


Figure D.119: Termini Profile for AAIndex ID: PRAM820101: Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)

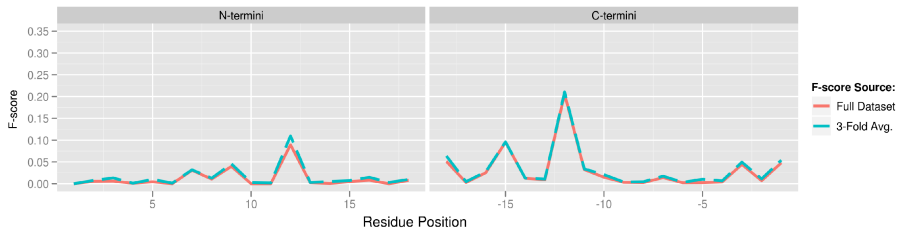


Figure D.120: Termini Profile for AAIndex ID: PRAM900101: Hydrophobicity (Prabhakaran, 1990)

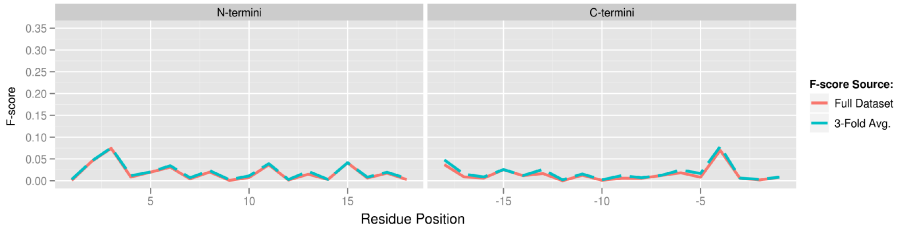


Figure D.121: Termini Profile for AAIndex ID: PTIO830101: Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)

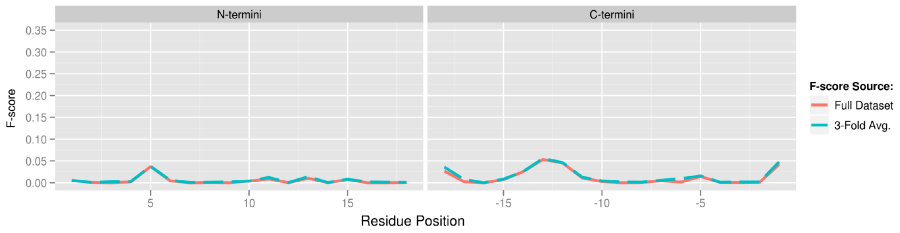


Figure D.122: Termini Profile for AAIndex ID: QIAN880102: Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)

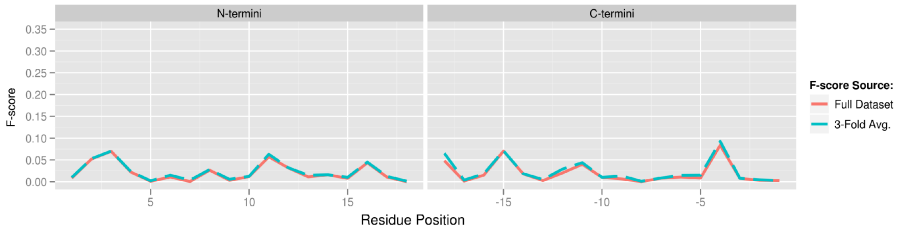


Figure D.123: Termini Profile for AAIndex ID: QIAN880110: Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)

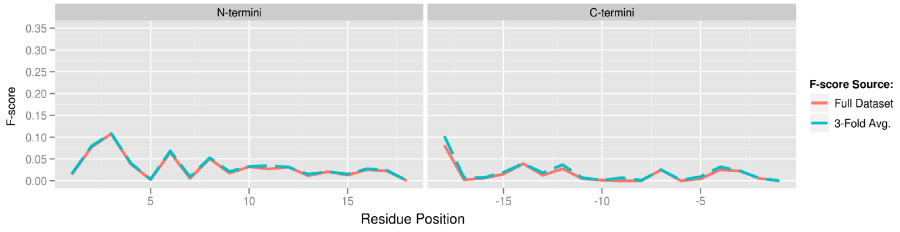


Figure D.124: Termini Profile for AAIndex ID: QIAN880112: Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)

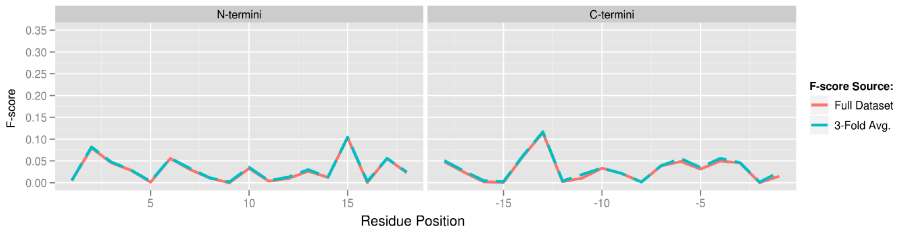


Figure D.125: Termini Profile for AAIndex ID: QIAN880114: Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)

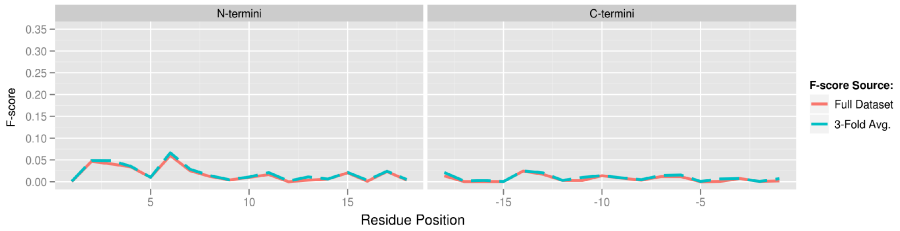


Figure D.126: Termini Profile for AAIndex ID: QIAN880116: Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)

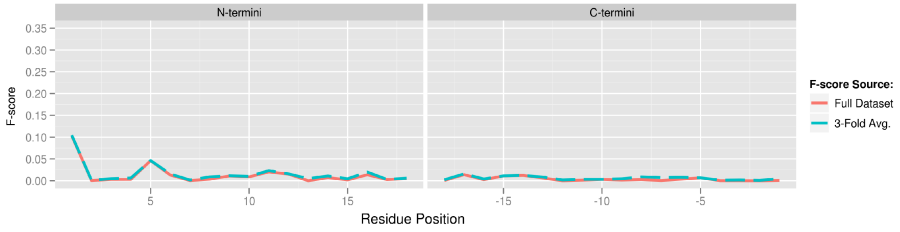


Figure D.127: Termini Profile for AAIndex ID: QIAN880117: Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)

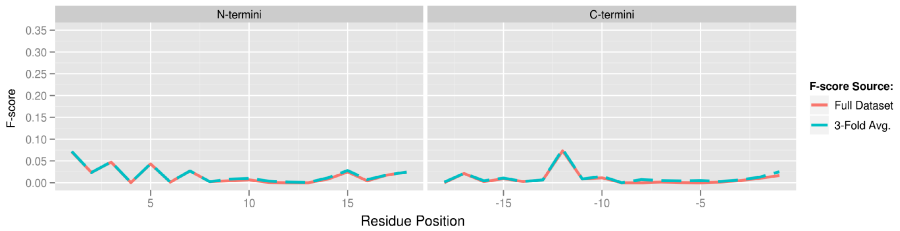


Figure D.128: Termini Profile for AAIndex ID: QIAN880118: Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)

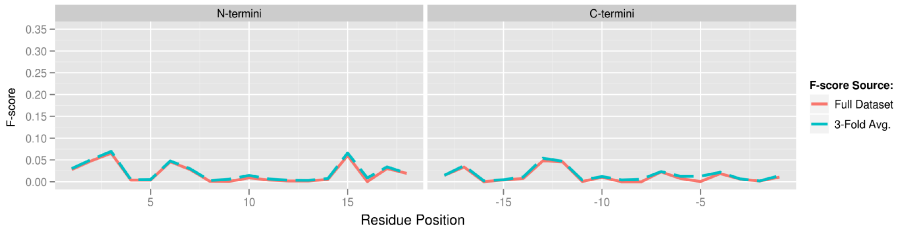


Figure D.129: Termini Profile for AAIndex ID: QIAN880121: Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)

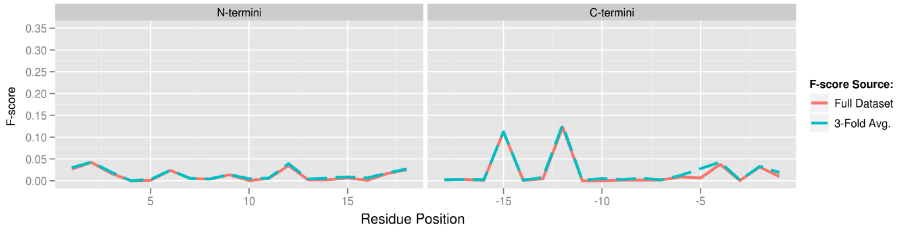


Figure D.130: Termini Profile for AAIndex ID: QIAN880122: Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)

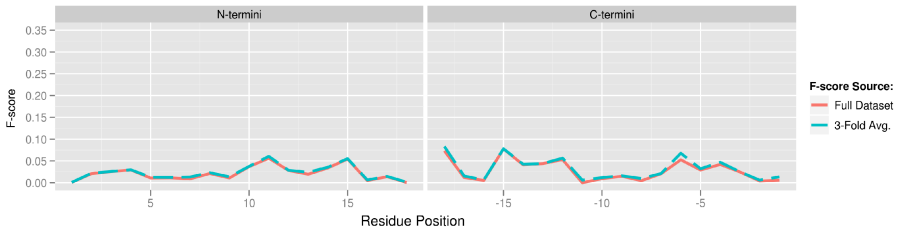


Figure D.131: Termini Profile for AAIndex ID: QIAN880124: Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)

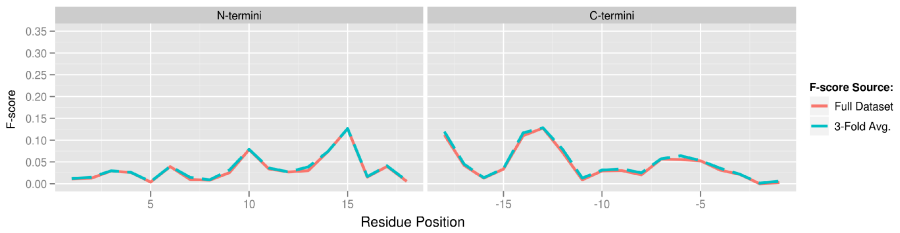


Figure D.132: Termini Profile for AAIndex ID: QIAN880125: Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)

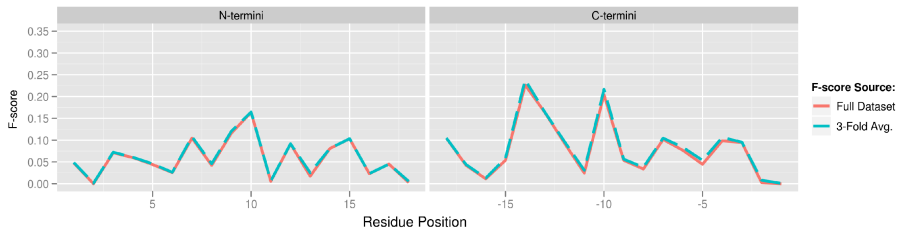


Figure D.133: Termini Profile for AAIndex ID: QIAN880129: Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)

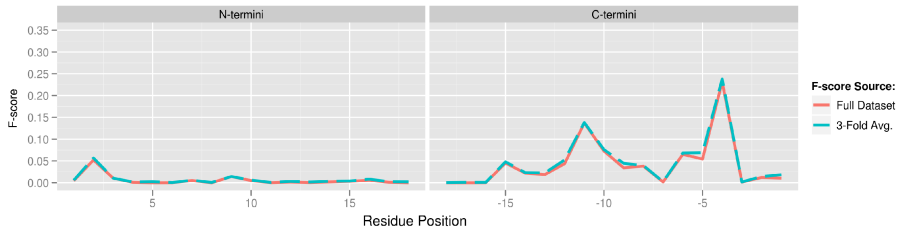


Figure D.134: Termini Profile for AAIndex ID: QIAN880137: Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)

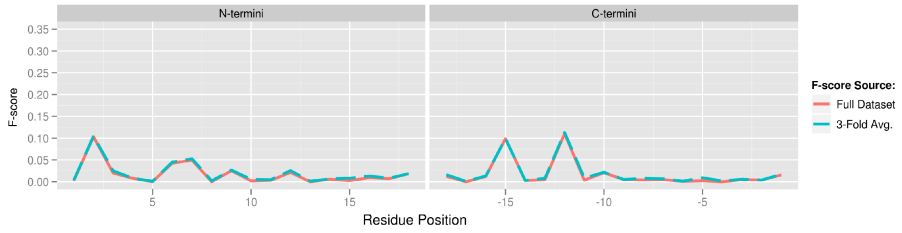


Figure D.135: Termini Profile for AAIndex ID: RACS770103: Side chain orientational preference (Rackovsky-Scheraga, 1977)

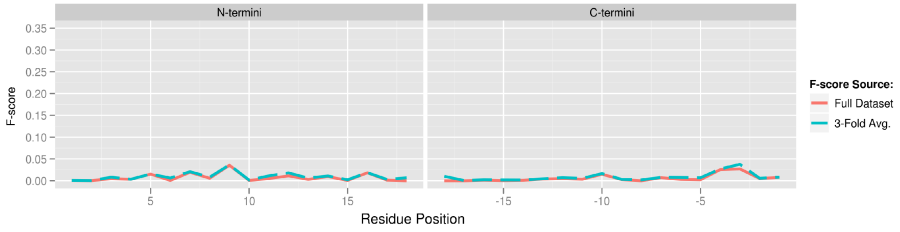


Figure D.136: Termini Profile for AAIndex ID: RACS820104: Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)

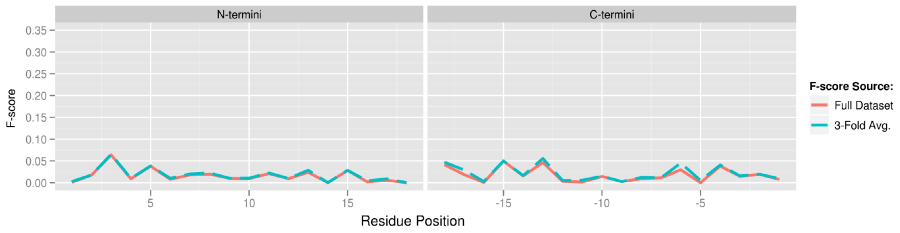


Figure D.137: Termini Profile for AAIndex ID: RACS820110: Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)

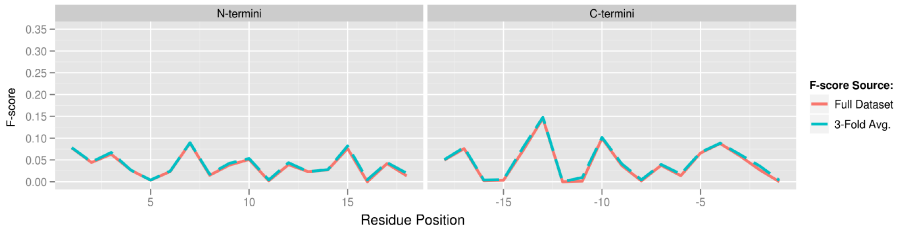


Figure D.138: Termini Profile for AAIndex ID: RACS820111: Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)

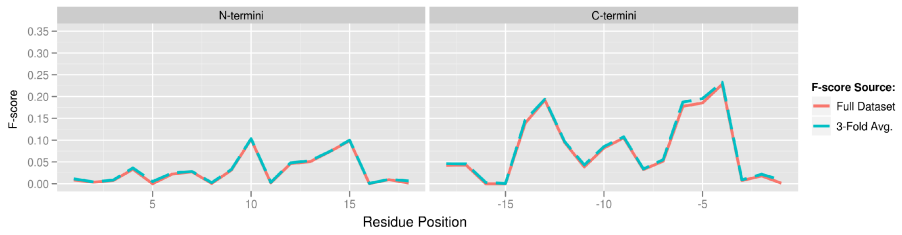


Figure D.139: Termini Profile for AAIndex ID: RACS820112: Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)

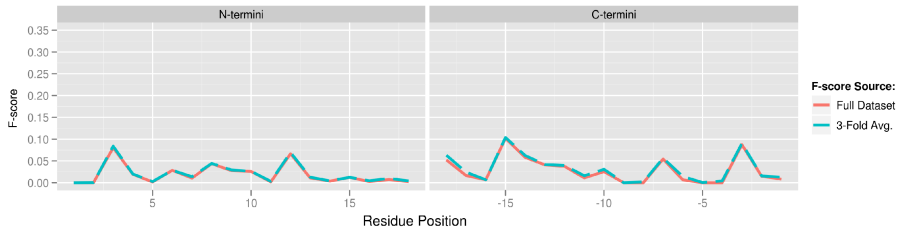


Figure D.140: Termini Profile for AAIndex ID: RADA880103: Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)

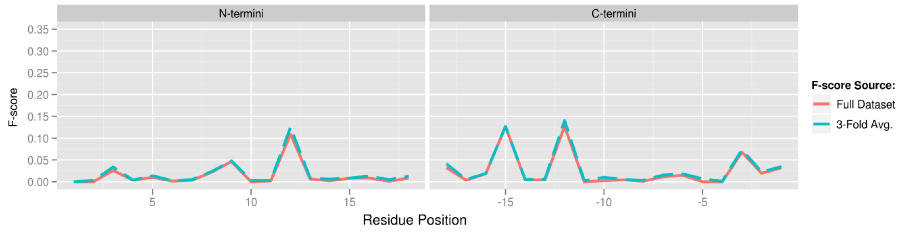


Figure D.141: Termini Profile for AAIndex ID: RADA880104: Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)

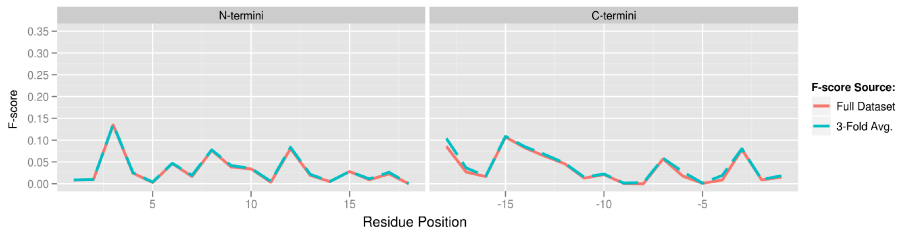


Figure D.142: Termini Profile for AAIndex ID: RADA880106: Accessible surface area (Radzicka-Wolfenden, 1988)

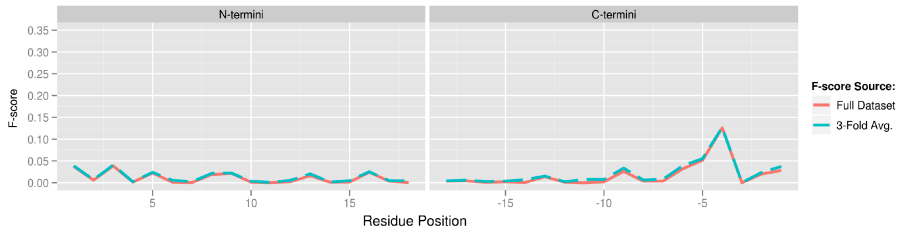


Figure D.143: Termini Profile for AAIndex ID: RICJ880104: Relative preference value at N1 (Richardson-Richardson, 1988)

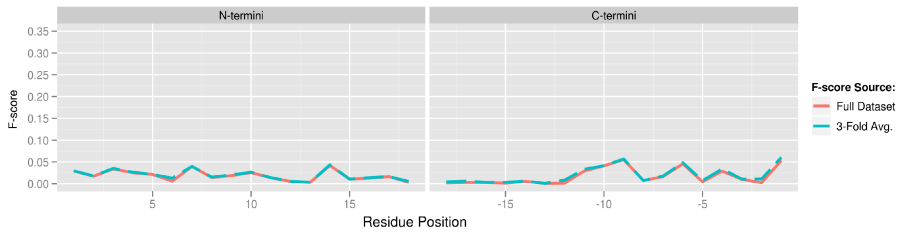


Figure D.144: Termini Profile for AAIndex ID: RICJ880105: Relative preference value at N2 (Richardson-Richardson, 1988)

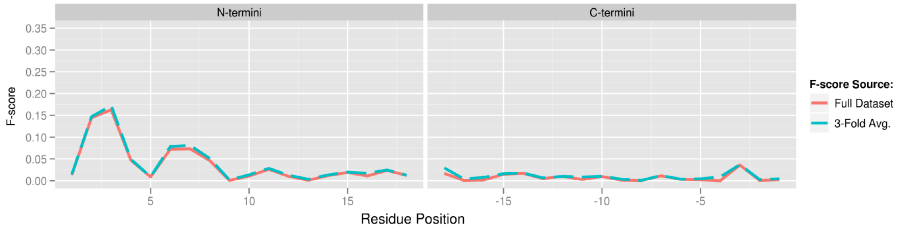


Figure D.145: Termini Profile for AAIndex ID: RICJ880107: Relative preference value at N4 (Richardson-Richardson, 1988)

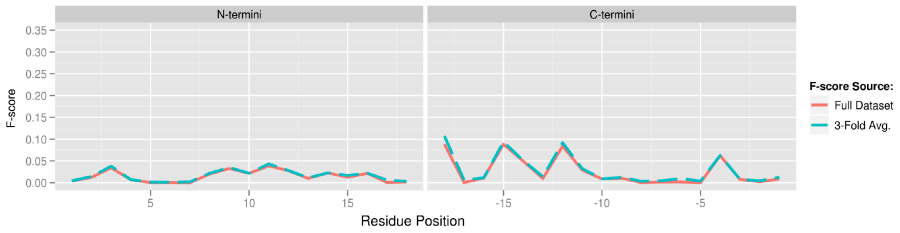


Figure D.146: Termini Profile for AAIndex ID: RICJ880108: Relative preference value at N5 (Richardson-Richardson, 1988)

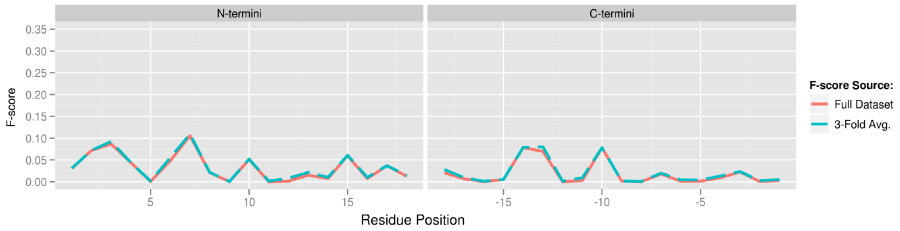


Figure D.147: Termini Profile for AAIndex ID: RICJ880111: Relative preference value at C4 (Richardson-Richardson, 1988)

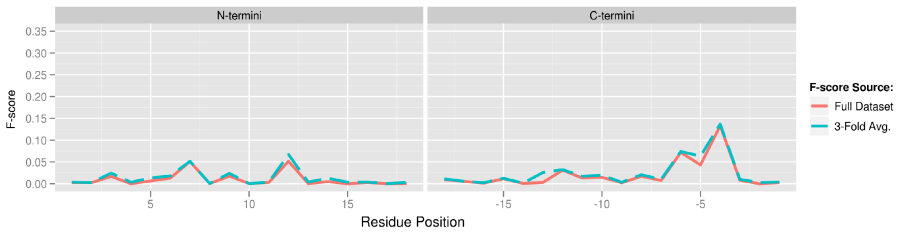


Figure D.148: Termini Profile for AAIndex ID: RICJ880116: Relative preference value at C' (Richardson-Richardson, 1988)

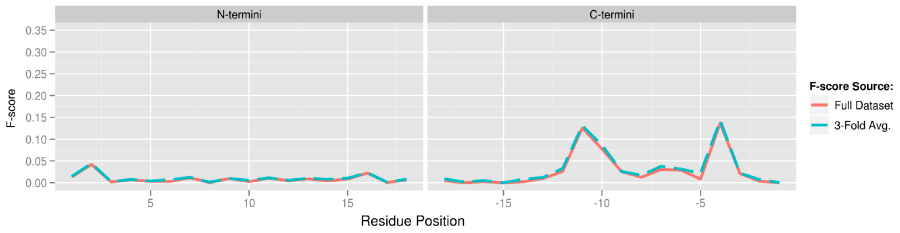


Figure D.149: Termini Profile for AAIndex ID: ROBB760109: Information measure for N-terminal turn (Robson-Suzuki, 1976)

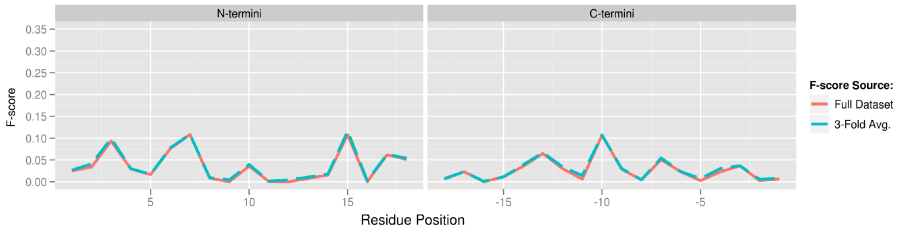


Figure D.150: Termini Profile for AAIndex ID: ROBB790101: Hydration free energy (Robson-Osguthorpe, 1979)

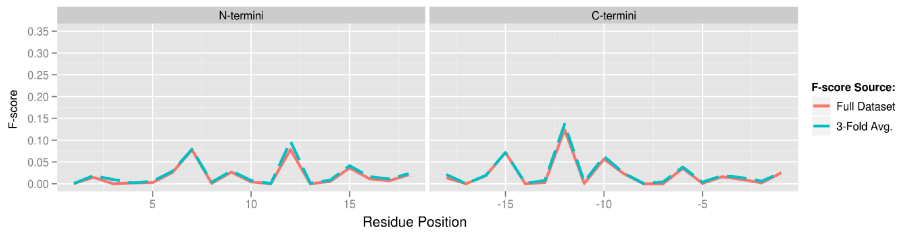


Figure D.151: Termini Profile for AAIndex ID: ROSM880102: Side chain hydropathy, corrected for solvation (Roseman, 1988)

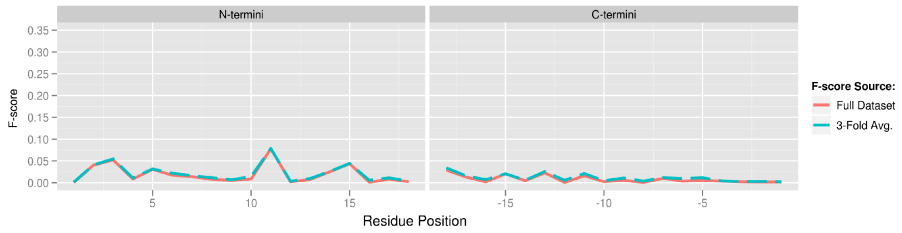


Figure D.152: Termini Profile for AAIndex ID: SNEP660101: Principal component I (Sneath, 1966)

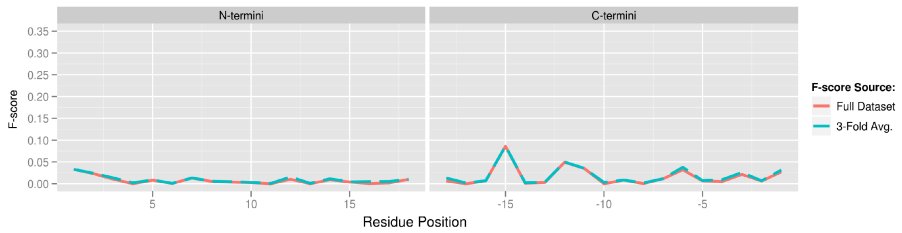


Figure D.153: Termini Profile for AAIndex ID: SNEP660102: Principal component II (Sneath, 1966)

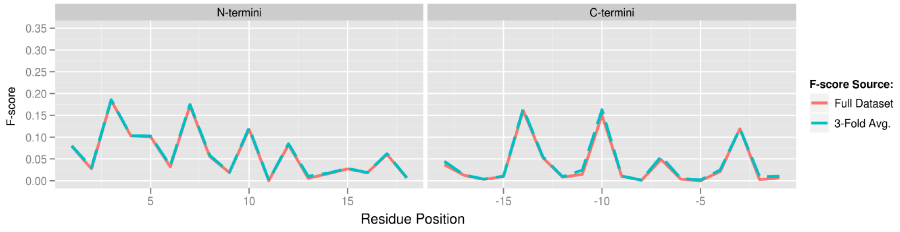


Figure D.154: Termini Profile for AAIndex ID: SNEP660103: Principal component III (Sneath, 1966)

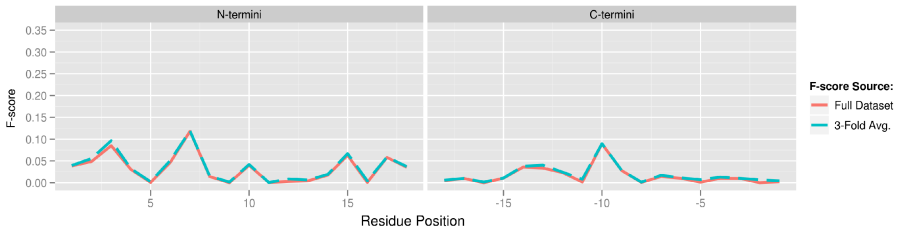


Figure D.155: Termini Profile for AAIndex ID: SWER830101: Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)

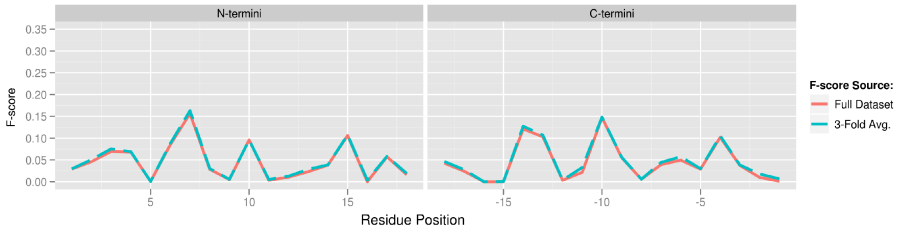


Figure D.156: Termini Profile for AAIndex ID: TAKK010101: Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)

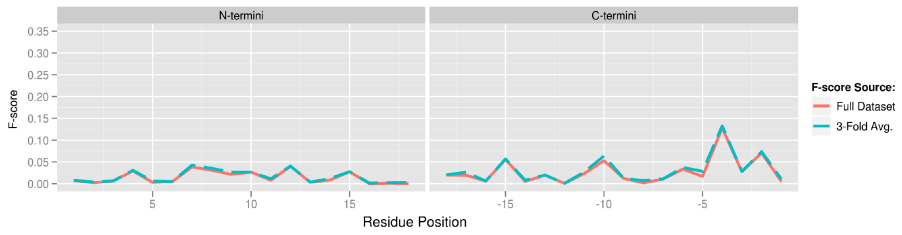


Figure D.157: Termini Profile for AAIndex ID: TANS770102: Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)

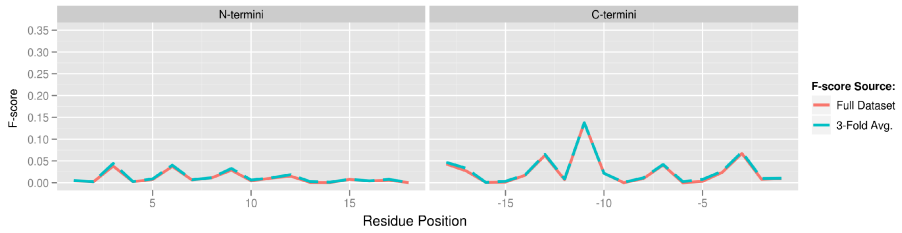


Figure D.158: Termini Profile for AAIndex ID: TANS770108: Normalized frequency of zeta R (Tanaka-Scheraga, 1977)

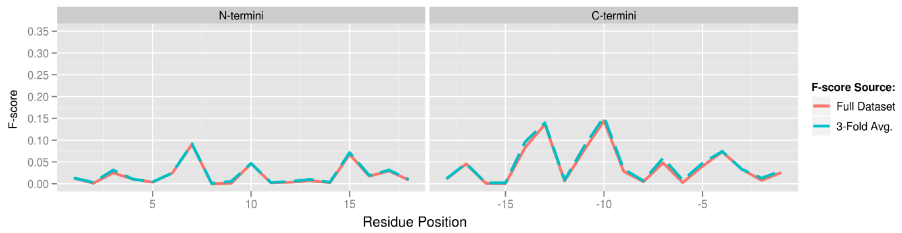


Figure D.159: Termini Profile for AAIndex ID: VASM830103: Relative population of conformational state E (Vasquez et al., 1983)

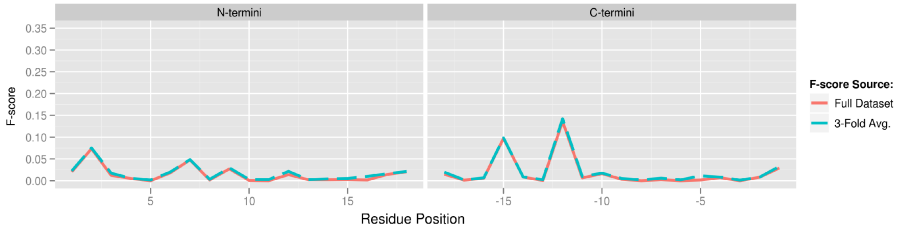


Figure D.160: Termini Profile for AAIndex ID: VINM940104: Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)

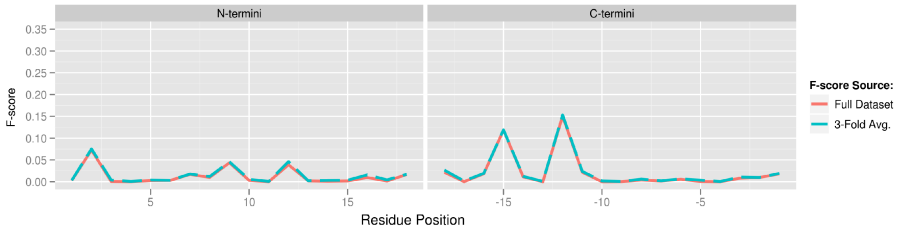


Figure D.161: Termini Profile for AAIndex ID: WARP780101: Average interactions per side chain atom (Warme-Morgan, 1978)

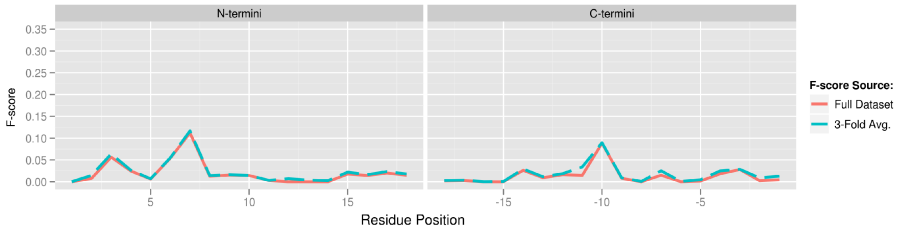


Figure D.162: Termini Profile for AAIndex ID: WEBA780101: RF value in high salt chromatography (Weber-Lacey, 1978)

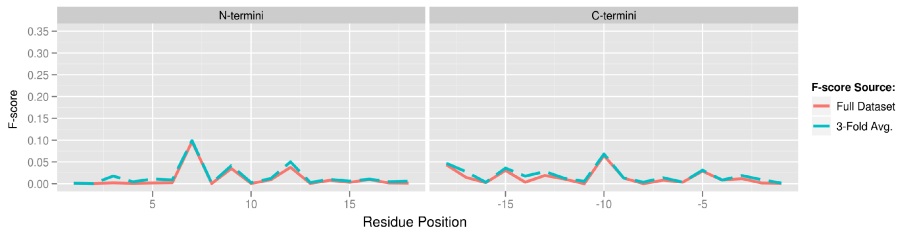


Figure D.163: Termini Profile for AAIndex ID: WERD780103: Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)

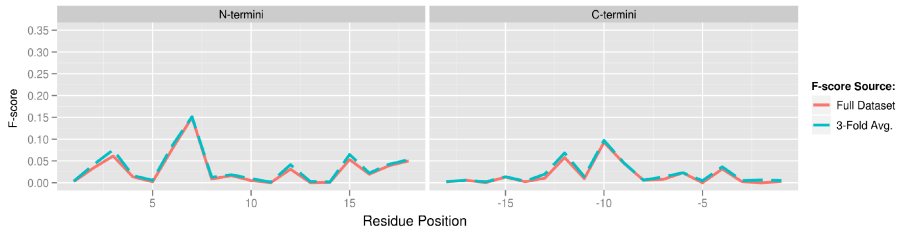


Figure D.164: Termini Profile for AAIndex ID: WILM950101: Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)

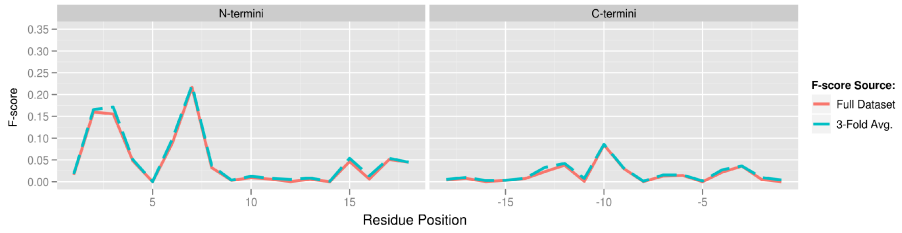


Figure D.165: Termini Profile for AAIndex ID: WILM950102: Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)

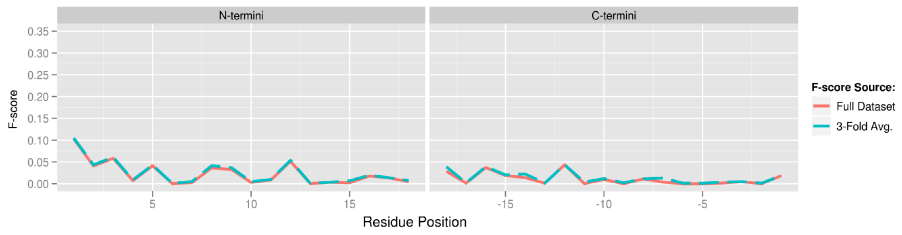


Figure D.166: Termini Profile for AAIndex ID: WILM950103: Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)

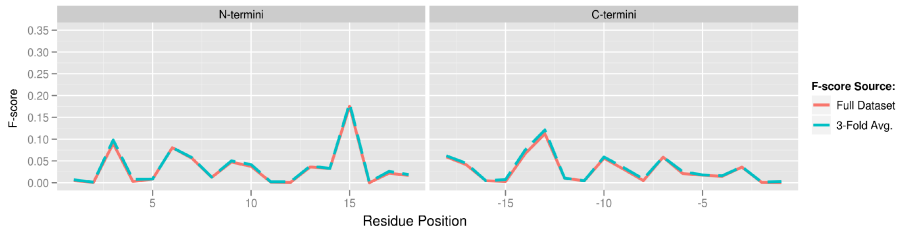


Figure D.167: Termini Profile for AAIndex ID: WILM950104: Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce et al. 1995)

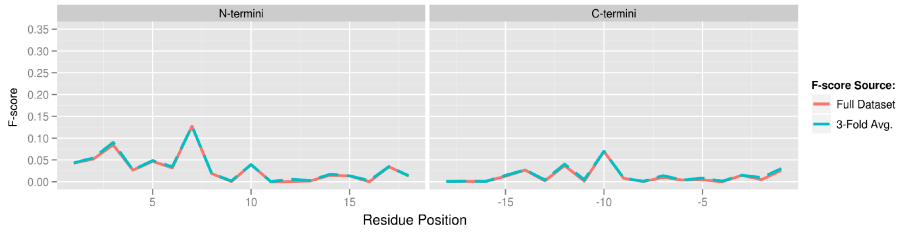


Figure D.168: Termini Profile for AAIndex ID: WIMW960101: Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996)

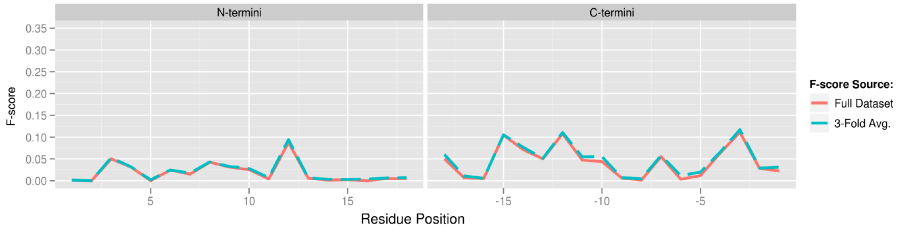


Figure D.169: Termini Profile for AAIndex ID: WOLS870102: Principal property value z2 (Wold et al., 1987)

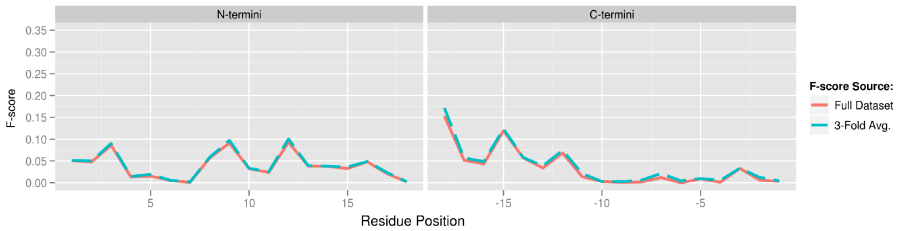


Figure D.170: Termini Profile for AAIndex ID: WOLS870103: Principal property value z3 (Wold et al., 1987)

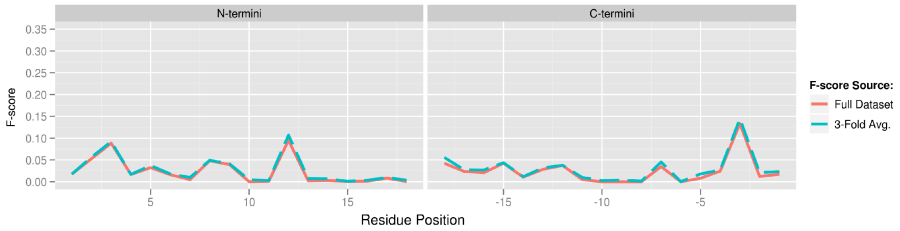


Figure D.171: Termini Profile for AAIndex ID: YUTK870103: Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)

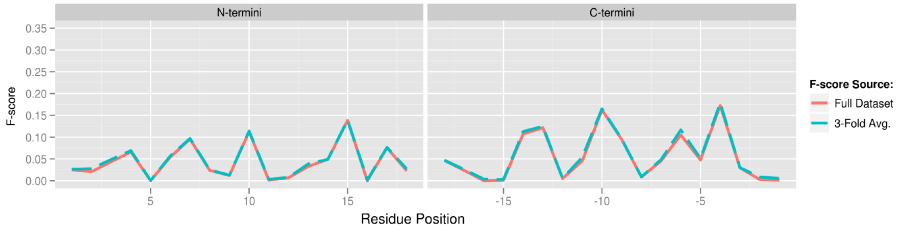


Figure D.172: Termini Profile for AAIndex ID: ZIMJ680101: Hydrophobicity (Zimmerman et al., 1968)

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 Staphylo-coccal 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] C. T. Bergstrom, M. Lo, and M. Lipsitch, “Ecological theory suggests that antimicrobial cycling will not reduce antimicrobial resistance in hospitals," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 36, pp. 13 285-13 290, 2004.

[7] B. M. Kuehn, “FDA targets antibiotic use in livestock," *JAMA: The Journal of the American Medical Association*, vol. 304, no. 4, p. 396, 2010.

[8] S. Kawashima and M. Kanehisa, “AAindex: amino acid index database," *Nucl. Acids Res.*, vol. 28, no. 1, p. 374, 2000.

[9] S. Lata, B. K. Sharma, and G. P. Raghava, “Analysis and prediction of antibacterial peptides," *BMC Bioinformatic*s, vol. 23, no. 8, pp. 263-272, 2007.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] D. Veltri and A. Shehu, “Physico-chemical features for recognition of antimicrobial peptides," in *Proceedings from the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*. IEEE, 2012, pp. 942-942.
[16] 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.
[17] H. G. Boman, “Antibacterial peptides: basic facts and emerging concepts," *Journal of Internal Medicine*, vol. 254, no. 3, pp. 197-215, 2003.
[18] R. E. Hancock, K. L. Brown, and N. Mookherjee, “Host defense peptides from invertebrates - emerging antimicrobial strategies," *Immunobiology*, vol. 211, no. 4, pp. 315-322, 2006.
[19] A. Tossi, L. Sandri, and A. Giangaspero, “Amphipathic, α-helical antimicrobial peptides," *Peptide Science*, vol. 55, no. 1, pp. 4-30, 2000.
[20] 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.
[21] K. G. Meade, S. Cahalane, F. Narciandi, P. Cormican, A. T. Lloyd, and C. OFarrelly, “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.
[22] G. Wang, *Antimicrobial Peptides: Discovery, Design and Novel Therapeutic Strategies*. Wallingford, England: CABI Bookshop, 2010.
[23] 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.
[24] Z. Wang and G. Wang, “APD: the antimicrobial peptide database," *Nucl. Acids Res*.,vol. 32, no. Sup.1, pp. D590-D592, 2004.
[25] Y. Shai, “Mode of action of membrane active antimicrobial peptides," *Peptide Science*, vol. 66, no. 4, pp. 236-248, 2002.
[26] Y. Pouny, D. Rapaport, A. Mor, P. Nicolas, and Y. Shai, “Interaction of antimicrobial dermaseptin and its uorescently labeled analogues with phospholipid membranes," *Biochemistry*, vol. 31, pp. 12 416-12 423, 1992.
[27] E. G and L. H., “Electrically gated ionic channels in lipid bilayers," *Q. Rev Biophys*, vol. 10, pp. 1-34, 1977.
[28] K. A. Brogden, “Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria?" *Nature Reviews Microbiology*, vol. 3, no. 3, pp. 238-250, 2005.
[29] 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.
[30] 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.
[31] 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.
[32] M. Seil, E. Kabr, C. Nagant, M. Vandenbranden, U. Fontanils, A. Marino, S. Pochet, and J.P. Dehaye, “Regulation by CRAMP of the responses of murine peritoneal macrophages to extracellular ATP," *Biochimica et Biophysica Acta Biomembranes*, vol. 1798, no. 3, pp. 569-578, 2010.
[33] K. Yamasaki, J. Schauber, A. Coda, H. Lin, R. A. Dorschner, N. M. Schechter, C. Bonnart, P. Descargues, A. Hovnanian, and R. L. Gallo, “Kallikrein-mediated proteolysis regulates the antimicrobial effects of cathelicidins in skin," *The FASEB Journal*, vol. 20, no. 12, pp. 2068-2080, 2006.
[34] K. Yamasaki, A. Di Nardo, A. Bardan, M. Murakami, T. Ohtake, A. Coda, R. A. Dorschner, C. Bonnart, P. Descargues, A. Hovnanian et al., “Increased serine protease activity and cathelicidin promotes skin inammation in rosacea*," Nature medicine*, vol. 13, no. 8, pp. 975-980, 2007.
[35] J. Schauber and R. L. Gallo, “Antimicrobial peptides and the skin immune defense system," *Journal of Allergy and Clinical Immunology*, vol. 122, no. 2, pp. 261-266, 2008.
[36] R. M. Dawson and C.-Q. Liu, “Cathelicidin peptide SMAP-29: comprehensive review of its properties and potential as a novel class of antibiotics," *Drug Development Research*, vol. 70, no. 7, pp. 481-498, 2009.
[37] Y. Xiao, Y. Cai, Y. R. Bommineni, S. C. Fernando, O. Prakash, S. E. Gilliland, and G. Zhang, “Identification and functional characterization of three chicken cathelicidins with potent antimicrobial activity," *Journal of Biological Chemistry*, vol. 281, no. 5, pp. 2858-2867, 2006.
[38] E. Bailey. (2008) The dermatology report: Advances in rosacea and acne vulgaris. Accessed: 03/09/2013. [Online]. Available: http://www.thedermatologyreport.com/derm/derm020104.html
[39] E. Alpaydin, *Introduction to machine learning*. MIT press, 2004, pp. 1-19.
[40] K. Jensen, M. Styczynski, and G. Stephanopoulos, “Machine learning approaches to modeling the physiochemical properties of small peptides," 2005.
[41] 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.
[42] 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.
[43] 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.
[44] J. R. Taylor, *An introduction to error analysis: the study of uncertainties in physical measurements*. Univ. Science Books, 1997.
[45] 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.
[46] 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.

[47] B. Boser, “A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, Pennsylvania, United States, 1992, pp. 144-152.
[48] C. Cortes and V. Vapnik, “Support-vector networks." Springer Netherlands, 1995, vol. 20, pp. 273-297.
[49] T. Furey, “Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906-914, 2000.
[50] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.
[51] C. Chen, Y.-X. Tian, X.-Y. Zou, P.-X. Cai, and J.-Y. Mo, “Using pseudo-amino acid composition and support vector machine to predict protein structural class," *Journal of Theoretical Biology*, vol. 243, no. 3, pp. 444-448, 2006.
[52] A. Zien, G. Ratsch, S. Mika, B. Scholkopf, and T. Lengauer, “Engineering support vector machine kernels that recognize translation initiation sites," *Bioinformatics*, vol. 16, pp. 799-807, September 2000.
[53] C.W. Hsu, C.C. Chang, and C.-J. Lin, “A practical guide to support vector classification," National Taiwan University, Tech. Rep., 2003.
[54] W. Noble and S.William, “What is a support vector machine?" *Nature Biotech*., vol. 24, no. 12, pp. 1565-1567, 2004.
[55] E. Alpaydin, *Introduction to machine learning*. MIT press, 2004, pp. 309-321.
[56] Y. Chen and C. Lin, “Combining SVMs with various feature selection strategies," in *Feature Extraction*, ser. Studies in Fuzziness and Soft Computing, I. Guyon,
[57] 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.
[58] 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.
[59] A. Cherkasov and B. Jankovic, “Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides," *Molecules*, vol. 9, no. 12, pp. 1034-1052, 2004.
[60] 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.
[61] 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.
[62] 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
[63] M. Magrane and the UniProt consortium, “UniProt knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, no. bar009, pp. 1-13, 2011.

[64] A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, “Large-scale computational analysis of protein arrangement in the lipid bilayer," *Biophys. J.*, vol. 100, no. S1, p. 492a, 2010.
[65] Z. Lu, Y. Wang, L. Zhai, Q. Che, H. Wang, S. Du, D. Wang, F. Feng, J. Liu, R. Lai et al., “Novel cathelicidin-derived antimicrobial peptides from Equus asinus," *FEBS Journal*, vol. 277, no. 10, pp. 2329-2339, 2010.
[66] 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
[67] Y. Igarashi, A. Eroshkin, S. Gramatikova, G. Gramatikoff, Y. Zhang, J. W. Smith, A.L. Osterman, and G. A., “CutDB: a proteolytic event database," *Nucl. Acids Res*., vol. 35, pp. D546-D549, 2007. [Online]. Available: http://cutdb.burnham.org
[68] K. Pearson, “On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559-572, 1901.
[69] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.
[70] 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.
[71] C. Staelin, “*Parameter selection for support vector machines*," Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1, 2003.
[72] M. B. Wilk and R. Gnanadesikan, “Probability plotting methods for the analysis for the analysis of data," *Biometrika*, vol. 55, no. 1, pp. 1-17, 1968.
[73] R Development Core Team, R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010.
[74] W. Hui, Y. R. Gel, and J. L. Gastwirth, “lawstat: An R package for law, public policy and biostatistics," *J Stat Software*, vol. 28, no. 3, pp. 1-26, 2005.
[75] T. Hothorn and K. Hornik, “The exactRankTests package," pp. 1-25, 2006.
[76] M. Zvelebil and B. Jeremy, *Understanding bioinformatics*. Garland Science, 2007.
[77] M. Wilce, M. Aguilar, and M. T. Hearn, “Physicochemical basis of amino acid hydrophobicity scales: Evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides," *Analytical Chemistry*, vol. 67, no. 7, pp. 1210-1219, 1995.
[78] C. Jameson, “Understanding NMR chemical shifts," *Annual Review of Phys Cchem*, vol. 47, no. 1, pp. 135-169, 1996.
[79] R. Tibshirani, “Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, 1996.
[80] Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J. W. Smith, L. Osterman, and A. Godzik, “Cutdb: a proteolytic event database," *Nucl. Acids Res*., vol. 35, no. suppl 1, pp. D546-D549, 2006.

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.