APPLICATION OF A TOPOLOGICAL DESCRIPTOR FOR PROTEIN INTERFACE IDENTIFICATION AND PROTEIN BINDING PREDICTION

by

Olivia Peters

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

Committee:

_____	Dr. Iosif Vaisman, Dissertation Director
_____	Dr. Tim Born, Committee Member
_____	Dr. Saleet Jafri, Committee Member
_____	Dr. Iosif Vaisman, Department Chairperson
_____	Dr. Richard Diecchio, Associate Dean for Academic and Student Affairs, College of Science
_____	Dr. Vikas Chandhoke, Dean, College of Science

Date: _____ Spring Semester 2010
George Mason University
Fairfax, VA

Application of a Topological Descriptor for Protein Interface Identification and Protein Binding Prediction

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

By

Olivia Peters
Master of Science
University of Virginia, 2000

Director: Dr. Iosif Vaisman, Associate Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2010
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my family for their love and support in all my endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

APPLICATION OF A TOPOLOGICAL DESCRIPTOR FOR PROTEIN INTERFACE IDENTIFICATION AND PROTEIN BINDING PREDICTION

Olivia J. Peters, PhD

George Mason University, 2010

Dissertation Director: Dr. Iosif Vaisman

The identification of proteins which interact or form complexes is a critical step in advancing several aspects of computational biology, including intelligent protein design and functional prediction. Previous methods have focused primarily on sequence alignment or threading methods to accomplish this, requiring large libraries of sequences. This work is an attempt to advance the current field of protein prediction through the use of a structural geometry methodology proven successful for many other aspects of proteomic analyses. The method is extended in two ways; first, a classification approach is created to identify protein residues involved in the binding interface, with the intent of using this information to aid the prediction of protein complex formation. Results are promising, with better than eighty percent correct classification, comparable to the best techniques currently in use. Second, a methodology was created to score potential

docking conformations. Of the 54 proteins in the test data set, 43 had a near-native structure in the top 100 positions, and a median ratio of successfully identified residue contacts of 0.57. The structural geometry method has been successfully applied to these two problems to advance the state of the field of proteomics.

CHAPTER 1: INTRODUCTION

Many biological processes depend on the formation and separation of protein complexes; one of the key current proteomics challenges right now focuses on determining the correct three-dimensional structure of two proteins upon joining. Predicting the final form of binding molecules enables better understanding of the protein complex formation process, in addition to facilitating the design of new molecules, such as drugs. Frequently, biological information is used to constrain the docking search space in order to arrive more quickly at a viable solution. Knowledge of the binding interface of the two proteins greatly simplifies the search and improves search results.

In this work, two complementary studies are described. The first outlines a new method to identify binding residues on a protein, while the second algorithm attempts to identify which of several potential docking solutions is the correct one. The information from these studies together represents a significant advance of the field.

This report is laid out as follows. The next chapter provides an overview of the state of research in both interface residue and interacting protein predictions. Chapter 3 discusses the methods used to achieve these goals. The results are included in Chapters 4 and 5, and the conclusion follows in Chapter 6.

CHAPTER 2: BACKGROUND

Studies to identify potential interface residues began earlier than those to predict docking conformations, probably due to the computational complexity of docking prediction. However, almost all current docking procedures make use of any available biological information, including knowledge of the binding interface. This chapter will review the history of protein interface prediction in Sections 2.1 and 2.2, providing an overview of studies characterizing the protein interface, and algorithms developed to predict interface residues, respectively. Section 2.3 will go into the current understanding of protein binding energetics. The remainder of the chapter will cover the highlights of protein docking. Section 2.4 overviews the basics of the docking algorithms, while Section 2.5 specifically discusses the algorithms used to score the population of potential conformations. Finally, Section 2.6 describes data sets which have been developed to test docking and scoring methods.

2.1 Protein Interface Characterization

Analysis of protein interfaces to identify binding sites began in the mid 1970s [24, 25]. This initial work consisted of characterizing the binding areas both quantitatively and qualitatively to attempt to identify rules that would lead to identification of binding surfaces in novel structures. Early work was done with very small sample sizes, leading

to a detailed understanding of a few protein interfaces, but little ability to generalize. Later work recognized the importance of larger data sets and attempted to reveal more general characteristics; thus far, few generalizations can be made across all protein interfaces. The lack of common characteristics is perhaps a function of the different types of interface interactions; that is, that homodimers, enzyme inhibitors, heterocomplexes, and antibody-antigen interactions may have evolved their interfaces into an optimized form dependent on their varying functions. For example, transient proteins complexes have been observed to rely more on salt bridges and hydrogen bonds, while temporally stable complexes rely more on hydrophobic attractions [95]. It has more recently come to light that because of this, the data sets may need to be made up of a specific complex, e.g. including only homodimers for novel homodimer interface prediction. Table 2-1 lists protein interface characteristics and the conclusions made by different studies; this is followed by Table 2-2, which outlines the data sets used in the studies summarized in Table 2-1.

Table 2-1: Protein interface characteristics summary.

	Differentiates interface from surface	**Does not differentiate interface from surface**
Hydrophobicity	[12, 38, 66, 68, 133, 134], for homodimers [33], hydrophobic core surrounded by more hydrophilic rim [79, 134], for large interfaces [108], hydrophobicity decreases as interface size decreases [2]	[2, 61, 64, 67, 76, 91, 100], for small globular proteins [87]
Aromatic (His, Tyr, Phe)	[61, 66, 79, 87, 100]	Trp (possibly because of double aromatic) [67]
Charged (Asp, Glu, His, Lys, Arg)	[33, 66, 79]	[121], except Arg [87]
Polar (Asp, Glu, His, Lys, Asn, Gln, Arg, Ser, Thr, Tyr)	[12, 29, 33, 61, 66, 91, 121], polarity of interface increases with decreasing interface size [2]	[87], similar content, but clustered [100]
Specific Residues	[12], Ala [100], Arg [2, 66, 87], Cys [2, 100], Glu [100], His [2, 66, 100], Ile [134], Leu [134], Lys [100], Met [2, 91, 100, 134], Phe [2, 66, 91, 134], Pro [39, 100], Thr [100], Trp [2, 91], Tyr [2, 66, 100, 134], Val [134]	[64], not for protease, inhibitor, and antigen [56], Val [91]
Electrostatic	[95, 133]	[67], except enzymes [38]
Hydrogen-bonding	[133]	[38, 68]
Evolutionary Conservation	[3, 12, 19, 100] only in enzyme and enzyme protein-ligand interfaces [38], conservation of Trp, Phe and Met [91]	not in protein-protein interfaces [19]
Size	largest cavity on the enzyme surface [91]	[12, 67, 71], dimer indistinguishable [91]

Table 2-2: Data set summary (ordered from earliest to most recent). Data sets labeled as a "mix of" complexes indicate that protein types were not considered individually, but that the study attempted to identify characteristics across all protein interfaces.

Data Reference	**Data Set Description**
Chothia [24], 1975	59 protein-protein interfaces: 32 homodimers, 4 permanent complexes, 7 monomers, 10 enzyme-inhibitor complexes, 6 antibody-protein complexes
Janin [61], 1990	15 protease-inhibitors, 4 antibody-antigen complexes
Young [133], 1994	mix of 38 enzyme and protein complexes
Jones [67], 1995	32 protein dimer interfaces
McCoy [95], 1997	mix of 12 protein-protein interfaces
Tsai [121], 1997	mix of 362 protein-protein interfaces (subunit-subunit, receptor-ligand, and enzyme-inhibitor) and 57 symmetry-related oligomeric interfaces
Larsen [81], 1998	mix of 136 homodimeric proteins
lo Conte [89], 1999	mix of 75 protein-protein complexes (24 protease-inhibitor, 19 antibody-antigen and 32 other complexes)
Gallet [43], 2000	mix of 80,000 sequences
Hu [56], 2000	mix of 97 protein-protein interfaces
Jones [63], 2000	mix of 46 monomers and mix of 105 oligomers or protein-complexes
Glaser [45], 2001	mix of 621 protein-protein interfaces
Zhou [134], 2001	mix of 744 non-homologous protein-protein interfaces (hetero- and homo-dimers)
Fariselli [34], 2002	226 heterodimers
Ma [92], 2003	mix of 86 interfaces (obligate dimers, proteinase inhibitors, antigens, protease complexes, and hormones)
Caffrey [20], 2004	64 protein-protein interfaces: mix of homodimers, heterodimers, transients
Koike [73], 2004	324 heterocomplexes and 674 homocomplexes
Neuvirth [100], 2004	57 non-homologous heteromeric, transient protein-protein interfaces
Aytuna [3], 2005	6170 interfaces: mix of homodimers and heterodimers, monomerics and complexes
Keskin [70], 2005	mix of 292 protein oligomers
Burgoyne [19], 2006	97 pairwise non-obligate hetero-complexes (22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 56 other complexes) and 134 ligand-protein complexes (95 in enzymes, 39 not in enzymes)
de Vries [31], 2006	1494 protein-protein interfaces: 518 homodimers, 114 heterodimers, 862 multimers

The two previous tables illustrate the difficulty of identifying distinguishing characteristics. The inability to draw general conclusions may be due to either the small number of sample points in the earlier data sets, or the composition of the data sets throughout the studies [73]. One of the few generalizations that appears to be consistent is that protein interfaces more closely resemble protein surfaces than protein cores [115, 121], despite their becoming part of a core once the complex has been formed.

Some of the studies that have performed an in-depth analysis of single protein interfaces have observed that the structure of some protein interfaces seems to consist of a few critical residues, usually evolutionarily conserved [67, 81, 91], that contribute a large amount to the binding energies [19, 81, 91]. These key residues (and sometimes the structurally surrounding residues) are referred to as "hot spots," and may provide the interface scaffold [67, 70]. Observed characteristics of hot spots include:

- enriched in Trp, Tyr, and Arg [2, 38, 91];
- a number proportional to the interface size [67];
- enclosed in protein pockets [19, 67, 81];
- tighter packing than the rest of the interface, possibly to facilitate the removal of water molecules upon binding [67];
- not favored to form hydrogen bonds [67]; and
- no preference to be involved in charged electrostatic interactions [67].

If interfaces are characterized by a pattern of hot spots, surrounded by supporting residues, calculating properties by averaging across the entire interface would not be expected to provide accurate characteristics.

In summary, several studies have been performed to determine the characteristics of protein interfaces, with contradictory results. This may be due to the small amount of data used, the lack of focus on a specific interface type, or possibly because the structure within the interface is averaged out when the entire interface is considered. The investigated characteristics of protein-protein interfaces do not appear to give the ability to distinguish between the interface and the remainder of the protein surface, leading researchers to explore computational methods of interface prediction.

2.2 Prediction of Binding Interface Residues

A variety of algorithms have been developed in an attempt to accurately predict residues involved in the binding interface. The most popular algorithms include: a weighted combination of chemical and physical properties of the residues [36, 38, 39, 63, 65, 100], neural networks [33, 134], support vector machines [12, 13, 71, 131], and multiple sequence alignment [50, 91, 132]. These methods are summarized in Table 2-3, with a description of the data set used, the method of classification, and the reported results.

Table 2-3: Interface prediction result summary.

Study	Data	Method	Accuracy Results
Jones& Thornton [66], 1996	59 complexes: 32 homodimers, 10 enzyme-inhibitors, 6 antibody-proteins, 4 permanent complexes, 7 monomers	weighted combination of residue propensity, accessible surface area, protrusion index, planarity, and hydrophobicity	>70% (unclear how much overlap is required to declare success)
Jones & Thornton [64], 1997	59 complexes: 28 homodimers, 11 hetero-complexes, 14 homo-complexes, 6 antibody-antigens	weighted score of salvation potential, residue interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area	66% (groups considered separately)
Gallet *et al.* [43], 2000	818 and 136 non-redundant sequences	threshold: calculation of mean hydrophobic moment of residue and mean hydrophobicity of 11-residue window	59.1% and 80.1% (unclear if these are correct predictions or simply the amount of the sequence that is predicted to bind)
Zhou & Shan [134], 2001	615 (training) and 129 (testing) pairs of nonhomologous complex-forming homo- and hetero-dimers	Neural network, input sequence profiles and solvent exposure of target and surrounding residues	70% true positives, accounting for 65% of the true interface residues
Fariselli *et al.* [34], 2002	226 protein heterodimers	Neural network, input of 11 residue structural neighbors: identity and conservation	73% (considered only surface residues)
Ma *et al.* [92], 2003	86 obligate dimers, proteinase inhibitors, antigens, protease complexes, and hormones	multiple structure alignment to detect recurring substructural motifs	"higher correlation with experimental data"
Yan *et al.* [131], 2003	31 antibody-antigen and 19 protease-inhibitor	SVM, input of identity of target residue and 10 sequence neighbors	sensitivity: 82.3% and 78.5%; specificity: 81.0% and 77.6%
Yao *et al.* [112], 2003	79 proteins	for each family, determine importance of residues through sequence alignment; new protein is aligned to correct family and conserved residues are predicted on binding surface	"significant" overlap for 96% (no indication of what this overlap is)

Koike & Takagi [73], 2004	271 hetero-complexes, 292 homo-complex	svm, input sequence profiles of target and structural neighbors, relative accessible surface areas	63.2%-73.5%, classification performed on whole group and sub-groups
Neuvirth *et al*. [100], 2004	57 transient protein-protein heterocomplexes (required knowledge of both partners)	Weighted combination of non-regular secondary structures length, atom distribution, amino acid pairs, evolutionary conservation, chemical character, water binding, sequence distance, hydrophobic patch rank, and secondary structure	70% (50% overlap declared successful)
Yan *et al*. [130], 2004	77 (training) and 7 (testing) heterocomplexes	two stage: SVM followed by Bayesian classifier, input sequence	72%
Bordner & Abagyan [12], 2005	518 homodimers, 114 heterodimers, 862 multimers	SVM	97% (some overlap; 22% of the surface residues were included in an average predicted patch)
Bradford & Westhead [15], 2005	180 transient and obligate complexes (made sure all occurred *in vivo*)	SVM, inputs: surface shape, conservation, electrostatic potential, hydrophobicity, residue interface propensity, solvent accessible surface area	64% for enzyme-inhibitors, 85% for hetero-obligates, 82% for obligates, 63% for transients
Chen & Zhou [23], 2005	798 homodimers and 458 heterodimers	Consensus neural networks	80% with 51% coverage
Fernandez-Recio *et al*. [37], 2005	66 non-obligate, non-homologous hetero-complexes of known structure	threshold: favorable energy change when buried upon complex formation	80% (50% overlap declared successful)
Burgoyne & Jackson [19], 2006	134 protein-ligand complexes, 22 enzyme-inhibitors, 19 antibody-antigens, 56 other complexes	weighted combination of hydrophobicity, desolvation, electrostatics, and conservation	88% (25% overlap declared successful)
Porollo & Meller [108], 2007	262 heterocomplexes and 173 homocomplexes	Relative solvent accessibility	74%

Study comparison is difficult due to the various data sets used, inconsistent methods of data processing, and different definitions for both the interface itself and the success of the method are used. For example, many of the patch methods declare success if at least 50% of the predicted patch overlaps with the actual patch. Adding further complication, different studies use varying data selection and processing methods, which may include limiting the sequence identity [12, 33, 38] (the acceptable percent similarity also differs by study), including a resolution threshold at which the complex has been characterized [2, 71], and excluding chains annotated with specific words or phrases, including: membrane peptides [33], small proteins [33], coiled coils [33], glycoproteins [2], carbohydrates, [2], nucleic acids [2], etc.

Another discrepancy may result from dissimilar definitions of contacting residues. Residues are usually considered to be contacting if the difference between any two atoms of the residues is less than the sum of their van der Waals radii plus some small amount, usually 5 angstroms [3, 87], or the diameter of water, 2.8 angstroms [87].

2.3 Energetics of Protein-Protein Binding

Analogous to the studies attempting to identify binding sites, a similar analysis has been performed on interfaces known to interact in order to elucidate characteristics allowing differentiation of the most ideal conformation for two interacting proteins. Many of these characteristics have been investigated because of their contribution to the binding free energy of a complex. Alone, none of the characteristics appears able to

differentiate which components will come together to form a complex, but the combination determines how proteins form complexes.

Molecular docking is hypothesized to occur in two stages [21]. In the first stage, molecules diffuse in close proximity until interface patches are close enough for the second stage, binding, to begin, which results in modification of side chain and backbone conformations, and finishes with a high-affinity interaction. The driving force for the first stage, association, is the hydrophobic effect, with the electrostatic and/or desolvation contributions conferring specificity [58]. Those complexes composed of oppositely charged molecules form in regions with favorable electrostatic potential, while complexes with weak charge complementarity favor regions of low desolvation energy [26]. Finally, any conformational change of the protein upon binding involves burial of hydrophobic surfaces (desolvation), which enhances binding, but a change in entropy resulting from conformational changes, which discourages binding [1]. Many of these characteristics have been studied to further understand binding energetics.

Characteristics which favor protein interface binding include: regions of high surface complementarity interact [65, 81, 133, 135], charged residues pair with residues of complementary charge [2, 36, 54], and hydrophobic residues interact with each other [2, 43, 66, 121]. It has also been found that specific pairs of amino acids occur more frequently than others [29], including tryptophan and proline [2]; tryptophan and leucine [108]; phenylalanine and isoleucine [2]; and arginine and glutamic acid [2].

Unfavorable interactions include contacts between pairs of hydrophobic and polar residues [2, 108], contacts between pairs of hydrophobic and hydrophilic residues [2], and specifically contact between glycine and alanine [67].

Electrostatic complementarity has been found to differ in its impact with the type of complex, sometimes favoring binding [65, 115], sometimes indifferent [115], and sometimes opposing [115].

Prediction of the final complex structure may also be affected by conformational changes upon binding. It has been found that standard-size interfaces - 1600 $\pm$ 400 $Å^2$ - have small changes in conformation, such as shifts in surface loops, movement of short segments of polypeptide chain, or the rotation of side-chains, while large interfaces - 2000 to 4660 $Å^2$ - display large conformational changes [87], making their final conformation more difficult to predict.

2.4 Protein Docking Background

Utilizing knowledge of the factors that affect protein complex formation, docking algorithms attempt to use the structures of two sub-parts of a protein and predict how they join to form the final complex. This process is called docking, and offers the ability to predict the structure of both novel compounds and weak, transient complexes that are difficult to measure experimentally. Docking algorithms frequently have two phases, although they may be combined: candidate generation, and re-scoring of the docked candidates.

Initial generation of the potential complexes is done by keeping one protein stationary, and moving the other protein around it, generating a population of theoretical conformations. This is done using efficient mathematical algorithms, such as the fast Fourier transform, Monte Carlo simulations, or genetic algorithms. The process is usually done using the rigid-body assumption, where proteins are treated as solid objects, an accurate assumption if the molecules undergo little conformational change upon binding. However, if there is significant conformational change, the final structure of the complex can be more accurately predicted by incorporating side-chain or backbone flexibility into the docking algorithm.

To account for conformational change, different levels of flexibility can be introduced to the docking procedure. A minimal amount of flexibility results from the smoothing of protein surfaces or allowing some amount of overlap between the surfaces of the two proteins. Flexibility can also be incorporated explicitly by allowing sidechain and/or backbone flexibility either during docking or during the refinement step [11].

There are a number of different approaches to docking, and each method brings something unique to the field. Some of the more popular docking algorithms include:

- AutoDock: small molecule-receptor binding predictor using a genetic algorithm and empirical energy function [97]
- ClusPro: performs docking with PIPER, then clusters the top 1000 docked structures and selects the center as representative [27]
- DOCK: incremental construction docking method [98]

- EUDOC: generates ligand-receptor complexes for computational screening of chemical databases [105]
- FlexDock: predicts protein interactions with hinge motion in one of the docked molecules [114]
- FlexX: protein-ligand prediction by sampling conformation space with a discrete model and then performing a tree-search technique for placing the ligand in the active site [55]
- FTDOCK: rigid-body docking using Fourier correlation algorithm [42]
- GOLD: genetic algorithm [62]
- HADDOCK: allows both sidechain and backbone movements of the interface during the interface packing optimization stage [32]
- ICM-DISCO: Monte Carlo rigid-body search followed by ligand interface side-chain refinement [49]
- PIPER: FFT-based rigid body global search with pairwise potentials [77]
- RosettaDock: Monte Carlo simulation with explicit side chain flexibility [125]
- SOFTDOCK: coarse-grained docking method using Voronoi molecular surface [86]
- ZDOCK: FFT-based simulation [127]

These docking procedures generate a large number of candidate associations; in the simplest case, these candidates are ranked with various criteria, including geometric

fit or surface complementarity. The candidates may then be further refined using molecular dynamics or Monte Carlo simulations [114].

Biological information is almost always used to select the final docking candidates, including available interface data [51], biochemical data, information on sequence conservation in homologous proteins [84], interfacial statistics from known protein complexes, and binding free energy approximations [21].

2.5 Protein Docking Scoring Algorithms

Docking algorithms generate a large number of candidate conformations, and these potential solutions are ranked using a variety of methods, varying in complexity. Some of the most popular docking software programs (e.g. SOFTDOCK, FTDOCK, ZDOCK, PIPER) use only geometric or shape complementarity. However, many methods re-score the population of conformations after this initial laddering.

The majority of methods attempt to capture some or all of the features that comprise the complex physical chemistry that underlies the energetics of molecular binding. The most rigorous methods, including free energy perturbation, require molecular dynamics simulation and are extremely time-consuming [117]. Simplified scoring functions approximate the free energy of binding with terms including solvation energy, van der Waals forces, electrostatics (Poisson-Boltzman), interaction energy, buried surface area, desolvation energy, hydrophobicity, hydrogen bonds, or pair potential energies These empirical energy scoring functions attempt to calculate the binding energy and select the minimum as the correct solution. Coefficients of these

equations are either taken from experimental values, if known, or through linear regression of known interactions. This is problematic even if the minimum can be identified, because frequently the native solution is not the energy minimum.

Some methods model the molecular mechanics of the interaction of the two molecules and attempt to predict the correct conformation through these characteristics. Another class of algorithms select informative features, frequently representing some aspect of the energetics of the system, and attempt to combine them in a meaningful way, often through classification. A final group of algorithms, frequently called "knowledge-based" scoring methods, calculate characteristics or compile statistics on the preferences of atoms or amino acids from known structures [124].

A summary of scoring functions and the features or chemical properties they take into account is included in Table 2-4.

Table 2-4: Scoring functions.

Complementarity-based	
DOT-FADE [83]	Shape complementarity
Evolutionary Trace Method [69]	Complementarity of electrostatic potential, hydrophobicity and shape
Norel [102]	Geometric complementarity, simple hydrophobicity feature
Molecular Mechanics-like	
Chemscore [123]	Hydrogen bond score, acceptor-metal interaction score, lipophilic score, conformational entropy
Goldscore [123]	Hydrogen bond score, van der Waals score, intramolecular ligand strain
PROLEADS [7]	Lipophilic score, metal-binding score, hydrogen bond score, ligand internal energy, user defined active site

Empirical Energy Approximation	
ASP Method [126]	Desolvation energy
ATTRACT [94]	Pairwise interaction potentials
CAMLab [55]	Force field and solvation energies
ClusPro [26]	Desolvation, electrostatics
ComScore [46]	Atomic contact energy, van der Waals score, electrostatics
DOCK [98]	Electrostatics, van der Waals energy
DOT [93]	Electrostatic energy, van der Waals energy
Fitzjohn [40]	Electrostatics, van der Waals energy
HADDOCK [30]	Interaction restraint energy, buried surface area, desolvation energy
ICM-DISCO [35]	Van der Waals energy, electrostatics, hydrogen bonding energy, desolvation
IFACE [127]	Pair potentials
Jackson [59]	Electrostatic, van der Waals energy, hydrophobic
Moont [96]	Pair potential
pyDock [107]	Coulombic electrostatics, ASA-based desolvation, optional term for van der Waals energy
RDOCK [85]	Electrostatics and desolvation energies, shape complementarity
RosettaDock [125]	Van der Waals approximation, solvation energy, hydrogen bonding potential
SCore-RPScore [78]	Surface complementarity, pair potential score
ZRANK [106]	Van der Waals energy, electrostatics, desolvation
Feature Based	
BiGGER [104]	Neural network classifier: geometric complementarity, electrostatic interactions, desolvation energy, pairwise cross interface propensities
CIRCLE [119]	Regression: fraction of molecular surface area of the side-chain covered by polar atoms, fraction of side chain area buried by some other atom, secondary structure
FunHunt [90]	SVM classifier: docking environment score, energy decrease during Monte Carlo minimization, interface residue conservation, solvent accessible surface area, interface contact number, distance between two monomers centers of mass, number of unsatisfied hydrogen bond donors/acceptors

Feature Based (continued)	
Gottschalk [47]	Regression: tightness of fit, frequency of atoms, characteristics of atoms, chemical character, secondary structure, hydrophobic patches, distribution of water molecules, evolutionary conservation. Uses ProMate binding sites.
REGINA [53]	Multivariate analysis: correlation between molecular electric fields, number of different residue-residue interactions, interface conservation, surface shape complementarity, mean force potential, pair potentials, number of each type of residue in the interface along with their propensity to occur there
Voronoi Method [10]	Genetic algorithms: surface area, number of interface residues, fraction of interface residue type, interface residue Voronoi volume, fraction of pairs, centroid-to-centroid distance
Knowledge Based	
FastContact [18]	Electrostatic (classic distance dependent dielectric) and desolvation components (from PDB)
Kortemme [75]	Hydrogen bonding potential (derived from known structures)
Qin [109]	Interface prediction, experimental data
Muegge [99]	Protein-ligand atom pair interaction potentials calculated from known complexes

There are a few methods which don't fall neatly within any of the categories previously described. Kohlbacher *et al.* [72] score conformations by computing a theoretical ^{1}H-NMR spectrum for each structure and calculating the difference between the theoretical and calculated spectrum. The absolute areas of the difference spectra are used to rank the conformations. A consensus scoring method has been developed by Charifson *et al.* [22]; after comparing thirteen of the most popular scoring functions, the three found most effective on the dataset used (ChemScore, DOCK energy score, and Piecewise Linear Potential) were selected. The intersection of the top 300 from each of

the three methods was used to produce a final ranking. A final unique approach from Wang *et al.* [124] uses the information from all conformations generated by the docking method. They theorize that if the near-native conformations are sampled accurately, the center of the cluster is where the most near-native conformations should reside. Consequently, the score is based on the Cartesian distance between each conformation and its neighbors.

2.6 Protein Docking Data Sets

Several data sets for the entire docking process are available; most notably the CAPRI competition [60] that takes place two to four times per year since starting in 2001. This competition releases the unbound structures of new complexes and offers participants the opportunity to test their docking and scoring algorithms on novel structures. However, this competition (and many of the other data sets released) assume the use of a docking method to generate the population of conformations which are then rescored. Consequently, they are not suitable for testing scoring methods. A docking method can be selected and all scoring algorithms tested with that data, but different scoring algorithms perform differently with different docking data, as different docking methods utilize different energy minimization functions. Those scoring functions which complement the docking method appear to have improved performance over scoring methods which use the same information as the docking method [85].

In addition to the CAPRI competition dataset, additional data sets for testing docking algorithms in conjunction with scoring algorithms include:

- Protein-Protein Docking Benchmark, version 3 [57], contains 124 test cases: 88 rigid-body cases, 19 medium difficulty cases, and 17 difficult cases
- CCDC/Astex test set [101] contains 305 protein-ligand complexes
- Dockground [44] contains 99 unbound-unbound and 134 unbound-bound complexes

In recent years, decoy data sets have been developed to allow testing and comparison specifically of the scoring mechanisms. These decoy data sets include:

- Dockground [88]: 100 near-native decoys are calculated using Gramm-X for 99 unbound-unbound and 143 unbound-bound complexes
- CAPRI [60]: since round 10, putative solutions generated by participants using docking algorithms are available to other participants for re-ranking
- Gray Docking Decoys [48]: 1000 decoys for 54 targets

One of the biggest challenge in this area is the paucity of data, but recent efforts have made significant attempts to change this.

CHAPTER 3: METHODS

Related, but distinct methods were applied to the two studies addressed in this work. The data sets used for each study are described in Section 3.1, with those used for interface prediction included in Sections 3.1.1 and 3.1.2, and the set used for docking re-scoring given in Section 3.1.3. Section 3.2 presents an overview of the topological descriptor used in both studies. The final section, Section 3.3, focuses on the classification methods and metrics. Sections 3.3.1 and 3.3.3 describe the classification techniques used for interface prediction and scoring, respectively. The interface prediction classification metrics are included in Section 3.3.2, while measures of success for protein docking scoring are included in Section 3.3.4.

3.1 Data Sets

Most studies apply their method to a newly developed data set, resulting in an abundance of data sets, each with different characteristics. Many more data sets exist for the prediction of binding sites than for the prediction of interacting proteins. Table 2-2 summarizes the data sets for interface prediction; data sets for protein interaction are summarized in Section 2.6.

3.1.1 Interface Residue Prediction Proof of Concept Data Set

The data set used to select features and perform the initial classification was modified from the one developed by Halperin *et al.* [52]. This data set consists of 253 pairs of interacting interfaces; this included 439 proteins for which an interaction site could be predicted (some of the proteins had multiple partners in the data set). Of these, many of the chains (126 of the 439 possible) were unable to be tessellated, usually because one or more of the residues was missing the label indicating the C_α, and were removed from consideration. Chains with greater than 30% homology to another chain in the data set were removed (resulting in removal of an additional 193 chains). Nine proteins were removed because conservation scores were not calculated if less than five homologs could be identified in UniProt [129] for sequence alignment. The final data set was a mixed data set of 111 proteins; this data set consisted of 5,152 residues which interacted with another protein and served as positive examples, and 13,398 negative examples. During testing, the data set was adjusted to have an equal number of positive and negative examples by using all the positive examples, and 5,152 randomly selected negative examples. This resulted in a data set with a total of 10,304 samples, equally split between positive and negative. Interacting residues were determined by those labeled as interacting with a partner in PDBsum [82], a database that has an overview of all structures deposited in the Protein Data Bank [8], including how the proteins bind to each other.

In order to ensure that the training set was large enough, a learning curve experiment was conducted. The data were divided into 10 parts randomly; the size of the

data set was varied and classification accuracy was measured. This was repeated ten times; only the mean is reported here in Figure 3-1. Classification accuracy plateaus around 7000 residues, but as including additional samples resulted in slightly better classification accuracy, all possible data were included.

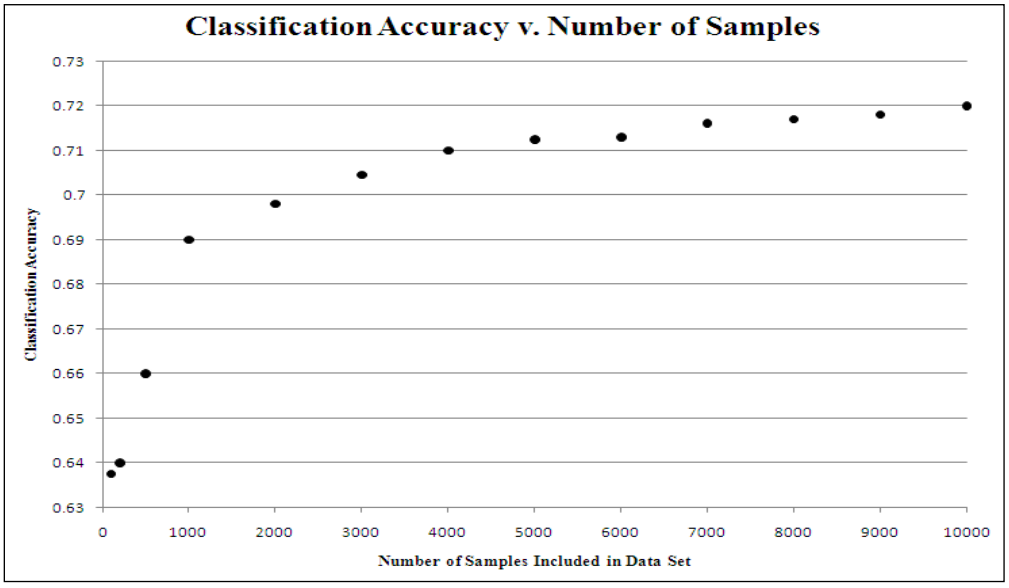

Figure 3-1: Classification accuracy as a function of data set size.

3.1.2 Interface Residue Prediction Comprehensive Data Set

After promising performance on the initial data set, a larger data set was required for more thorough testing. Despite an abundance of data sets developed through various studies, the decision was made to create a new data set that incorporated the largest possible number of samples. A search of the Protein Database [8] identified all multi-chain structures that contained proteins but not DNA or RNA molecules. After sequences with greater than 30% homology were removed, 4637 structures remained.

Following the same procedure outlined above, chains were removed if they could not be tessellated or if a conservation score could not be calculated. The final data set contained 1,476 chains, including 43,970 residues which interacted with another protein and served as positive examples, and 289,643 negative examples. Because the data was skewed and a high accuracy could be gained by assigning all residues as negative samples, a cost matrix of 5:1 was used to balance class distribution. PDBsum [82] was again used to identify interacting residues.

3.1.3 Scoring Function Data Set

The data set used to test the scoring function was developed by the Gray Lab [48] and includes 1000 decoys and the native structure for each of 54 proteins. Of these proteins, 22 are classified as enzyme/inhibitor complexes, 16 are antibody/antigen complexes, 6 are difficult complexes, and 10 are other complexes, distributed as shown in Table 3-1.

Table 3-1: Gray data set summary.

Enzyme / Inhibitor Complexes	1ACB, 1AVW, 1BRC, 1BRS, 1CGI, 1CHO, 1CSE, 1DFJ, 1FSS, 1MAH, 1PPE, 1STF, 1TAB, 1TGS, 1UDI, 1UGH, 2KAI, 2PTC, 2SIC, 2SNI, 2TEC, 4HTC
Antibody / Antigen Complexes	1AHW, 1BQL, 1BVK, 1DQJ, 1EO8, 1FBI, 1IAI, 1JHL, 1MEL, 1MLC, 1NCA, 1NMB, 1QFU, 1WEJ, 2JEL, 2VIR
Difficult Complexes	1BTH, 1EFU, 1FIN, 1FQ1, 1GOT, 3HHR
Other Complexes	1A0O, 1ATN, 1AVZ, 1GLA, 1IGC, 1MDA, 1SPB, 1WQ1, 2BTF, 2PCC

A detailed description of the dataset is included in the original article [48]. The first stage of docking was performed using a rigid-body Monte Carlo search, rotating one partner around the other with 500 Monte Carlo move attempts. Step sizes are continually adjusted to maintain a 50% move acceptance rate, with low-resolution, residue-scale interaction potentials based on a Bayesian expansion of the probability of the correctness of each decoy. Subsequently, explicit side chains were added to the protein backbone using a backbone-dependent rotamer packing algorithm, and the rigid body displacement is optimized. During this optimization, a full-atom scoring function is used with terms for van der Waals energy, solvation energy, hydrogen bonding energy, rotamer probabilities, residue-residue pair interactions, electrostatics, and surface area and atomic solvation.

3.2 Applied Topological Descriptor

Protein structures can be characterized using a computational geometry method based on three dimensional Delaunay tessellation. The use of statistical geometry to study the structure of disordered systems was introduced by Bernal [9], and further developed by Finney [38, 39] for Voronoi tessellation. Delaunay and Voronoi tessellations are duals of each other, as seen in Figure 3-2. To perform the tessellation, each amino acid is represented by its C_α (as opposed to the C_β or the center of mass of its side chain); it has been shown that this reduced representation allows accurate restoration to the full backbone structure [110]. The Delaunay tessellation divides the three dimensional space into convex polyhedra, with the four residues arranged at vertices of the tetrahedra. This

allows all sets of four nearest-neighbor points in space to be identified. The Delaunay simplex represents the ensemble of neighboring atoms, while the Voronoi polyhedron represents the environment of individual atoms.

Figure 3-2: Representation of the Delaunay (solid lines) and Voronoi (dashed lines) tessellations in two dimensional space.

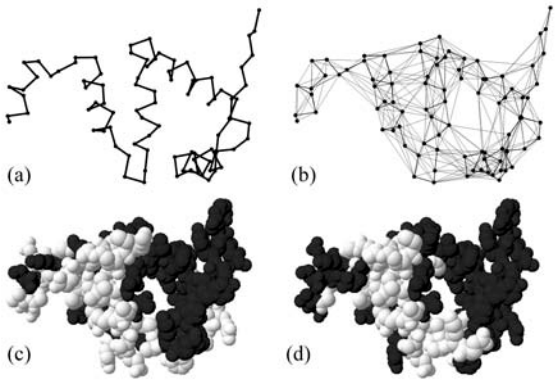

Figure 3-3: For a representative protein (2JD3-A), (a) the protein backbone, (b) the protein tessellated, (c) the protein with the correct interface residues indicated in black, and (d) the protein with the predicted interface residues indicated in black.

The Delaunay tessellation results in a series of non-overlapping, irregular tetrahedra, with the four residues at the vertices representing a set of four nearest-neighbor residues in structural space. Each of the four residues (*i, j, k, l*) form a four-body cluster in 3D space, but are separated by three distances (d_{ij}, d_{jk}, d_{kl}) in sequence space. The twenty naturally occurring amino acids are capable of yielding 8855 distinct quadruplets, but statistical analysis of the residue composition of these simplexes found nonrandom preferences for some amino acids to be clustered [116]. The simplexes can be classified into five nonredundant groups, shown in Figure 3-4, based on the relationship of the residues in the primary sequence: class **{4}**, where all four residues are consecutive in the primary sequence; class **{3,1}**, where three residues are consecutive, with the fourth removed; class **{2,2}**, in which two residues are consecutive, but separated from the other two, which are also consecutive; class **{2,1,1}**, with two consecutive residues and the other two distant from these two and each other; and class **{1,1,1,1}**, where none of the residues are consecutive. When multiple proteins are tessellated together, for example during docking, there is an additional class – class 5 – that includes all tetrahedra that have vertices on both proteins. The geometrical rules of tetrahedra, such as volume and tetrahedrality, can be used to characterize the simplexes using the equations

$$V = \frac{1}{3} A_0 h, \text{ and} \tag{1}$$

$$sT = \sum_{i>j} (l_i - l_j)^2 / 15\bar{l}^2, \tag{2}$$

where A_0 is the area of the tetrahedron base, h is the height from the base to the apex, l_i is the length of the i-th edge, and $\bar{l}$ is the mean length of the simplex edges. Tetrahedrality

provides a metric for the degree of difference between the simplex under consideration and the ideal simplex.

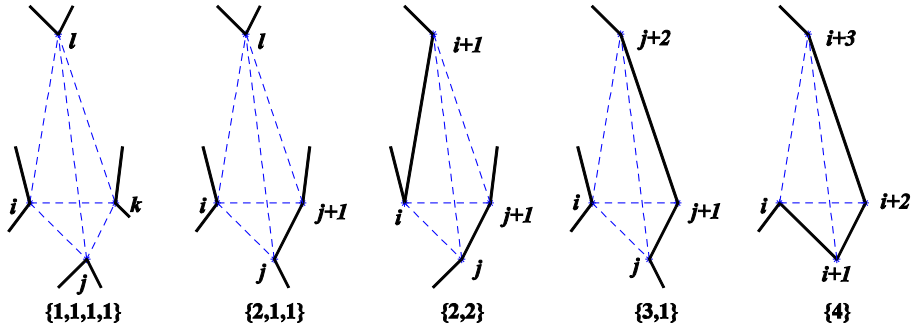

Figure 3-4: Examples of each of the simplex types.

For each quadruplet, a log-likelihood score is calculated, defined as

$$q_{ijkl} = \log \frac{f_{ijkl}}{p_{ijkl}}, \qquad (3)$$

where f_{ijkl} is the frequency of the quadruplet containing residues *i*, *j*, *k*, *l* in a non-redundant training set of high-resolution structures with low primary sequence identity obtained from the Protein Data Bank [8], and p_{ijkl} is the frequency of random occurrence of the quadruplet. The log-likelihood score can be interpreted as the non-random bias for four amino acid residues to be found in the same Delaunay simplex; this value is also known as the four-body statistical potential energy function, and frequently reflects important features of the protein. For example, the residues in local maxima values of the profile are frequently located in the hydrophobic core of the protein [12].

A suite of programs has been developed in Java and Perl to take the PDB files and perform the data extraction and formatting prior to tessellation. The Quickhull algorithm is then used to perform the protein tessellations. Originally developed for game theory, the Quickhull algorithm is commonly accepted as the most computationally efficient method of calculating the convex hull of a surface in two or more dimensions.

This method has been used successfully to: prioritize SNPs according to the degree of their functional effect on proteins [5], measure quantitative similarity between protein pairs [14], study protein structure-function correlations through computational mutagenesis [92], evaluate sequence-structure compatibility for inverted structure prediction [116], analyze the patterns of spatial proximity of residues in known protein structures [122], predict secondary structure [118], and evaluate the quantitative structural similarity between protein pairs [13].

3.3 Classification Methods and Metrics

3.3.1 Interface Prediction Classification Technique: Random Forests

All classification tests and evaluation were performed using Weka software [128]. Random Forest classification was chosen as the classification method that performed best for interface residue prediction after evaluation of several potential classifiers included within the Weka framework. The Random Forest algorithm was originally developed by Breiman [16], but built on the idea of random forests first proposed by Ho [54]. In this

method, multiple different decision trees - a forest - are created using random subsets of the training data and random elements of the feature vector. A novel residue is classified by presenting the feature vector to all of the trees, each of which votes, with the overall classification chosen as the one that has the most votes in all trees.

Random Forest classification has several advantages, which include estimating the importance of variables in determining classification, the ability to estimate missing data, and calculation of sample proximity.

3.3.2 Interface Prediction Classification Metrics

Because the interface residue prediction is a binary classification (either on the interface or not), several well known metrics can be used to assess classifier performance. Utilizing the number of True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*), these metrics were used to evaluate classifiers, as laid out by Baldi *et al.* [4], including accuracy (Acc), sensitivity (Sen), specificity (Spec), False Alarm Rate (FAR), Matthews Correlation Coefficient (MCC), and Bit Error Rate (BER):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Sen = 100\frac{TP}{TP + FN} \tag{5}$$

$$Spec = \frac{TP}{TP + FP} \tag{6}$$

$$FAR = \frac{FP}{FP + TN} \tag{7}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

$$BER = \frac{FN}{2(FN + TP)} + \frac{FP}{2(FP + TN)} \quad (9)$$

Each of these metrics brings additional information to assessing classifier performance. Accuracy indicates how close a measure is to the true value. Sensitivity, or recall, measures completeness; for these data, it represents the fraction of the interface correctly identified. The Specificity, also known as precision, can be considered a measure of the repeatability of the classifier. The False Alarm Rate indicates how frequently a residue is identified as on the interface when it is not. The Matthews Correlation Coefficient is a measure of the quality of binary classification, accurate even when the classes are different sizes, as in most protein data sets. This metric can be thought of as accuracy normalized to take into account different class sizes. Finally, the Bit Error Rate is the percentage of residue classifications that have errors. The function is a sum of the fraction of correct residues on the interface plus the fraction of correct residues everywhere but the interface. By evaluating each of these metrics, a true understanding of the classifier's strengths and weaknesses can be gained.

3.3.3 Scoring Classification Technique: Least Median Squared Linear Regression

Again with the Weka software, the method chosen for classification when ranking the docking data was the Least Median Squared (LMS) Linear Regression technique. This algorithm had the highest accuracy of all classifiers tested; the Support Vector

Machine achieved a similar accuracy, but with a much higher computational time. The Weka algorithm is based on the one described by Rousseeuw and Leroy [111], and generates least squared regression functions from random subsamples of the data. The Least Squared Regression with the lowest median squared error is chosen as the final model.

3.3.4 Protein Docking Measures of Success

Comparison of results from different docking and scoring algorithms is a challenge in itself. Rarely is the correct confirmation chosen, and metrics of success for the conformations chosen differ by study. The CAPRI competition has four classes into which predictions are placed: incorrect, acceptable (more than 10% of native residue-residue pairs in contact and within 4 angstroms RMS), medium (more than 30% of native residue-residue pairs in contact and within 2 angstroms RMS), and high (more than 50% of native residue-residue pairs in contact and within 1 angstrom RMS). Other investigations attempt to rank the correct solution in the top X conformations, where X varies from 50 to 2000 by study. Another metric for comparison is the mean rank and mean RMS. Some studies declare success if a near-native (instead of the native) conformation is found in the top X conformations. As with the interface prediction methods, different metrics of success make comparison between methods a challenge.

CHAPTER 4: RESULTS FOR INTERFACE PREDICTION

4.1 Interface Residue Prediction Feature Selection

A wide variety of features were considered for inclusion in the classifier. Several features were included from the topological descriptor, including: four-body statistical potential energy function (potential), the number of simplices of type [{1,1,1,1}, {2,1,1}, {2,2}, {3,1}, {4}] the residue participates in [T0, T1, T2, T3, T4], the sum of the volumes of the simplices the residue is part of (volume), and tetrahedrality (sT).

In addition to the features taken from the topological descriptor, additional potentially informational features were evaluated, including:

- Conservation (HSSP [113] and ConSurf [80]),
- Electrostatic Potential (Protein Continuum Electrostatics [6]),
- Secondary Structure (DSSP [68]),
- Residue Interface Propensity [68] (propensities for: interior, interface, and surface),
- Molecular Weight [17],
- Hydrophobic Potential [120] (values: positive, philic, phobic, negative),
- Side chain [74] (values: aliphatic, aromatic, neither),
- Hydrogen bonding ability [68] (values: yes, no), and

- Hydropathy [79], which is indicative of the hydrophobic or hydrophilic properties of the amino acid side chain.

In addition, the values of T0, T1, T2, T3, T4, Total, and Volume were normalized by protein to represent the percent instead of the absolute value, enabling comparison across proteins. These values were included in addition to the non-normalized values (not as a replacement). Those features which are consistent across amino acids are included in Table 4-1.

Table 4-1: Values for some of the features considered.

	Molecular weight [17]	Hydro [120]	Side chain [74]	Hydrogen bonding [68]	residue propensity [68]			Hydropathy [79]
					interior	interface	surface	
ALA	89.09	phobic	aliphatic	No	12.56	7.15	8.00	1.8
ARG	174.20	pos	Neither	Yes	1.19	6.22	5.22	-4.5
ASN	132.12	philic	Neither	Yes	2.19	5.44	5.74	-3.5
ASP	133.10	neg	Neither	Yes	2.81	5.74	7.78	-3.5
CYS	121.16	philic	Neither	No	2.85	1.26	0.89	2.5
GLN	146.15	philic	Neither	Yes	1.30	3.44	4.85	-3.5
GLU	147.13	neg	Neither	Yes	1.56	5.11	7.11	-3.5
GLY	75.07	phobic	aliphatic	No	8.30	6.70	8.74	-0.4
HIS	155.16	pos	Neither	Yes	1.67	2.74	2.59	-3.2
ILE	131.17	phobic	aliphatic	No	10.59	5.26	3.44	4.5
LEU	131.17	phobic	aliphatic	No	14.33	8.67	4.96	3.8
LYS	146.19	pos	Neither	Yes	0.74	5.26	7.30	-3.9
MET	149.21	phobic	Neither	No	3.41	2.41	1.33	1.9
PHE	165.19	phobic	aromatic	No	7.30	4.52	2.89	2.8
PRO	115.13	phobic	Neither	No	1.89	5.00	5.89	-1.6
SER	105.09	philic	Neither	Yes	4.56	5.00	7.48	-0.8
THR	119.12	philic	Neither	Yes	4.59	5.96	6.22	-0.7
TRP	204.23	phobic	Aromatic	Yes	1.41	2.30	2.41	-0.9
TYR	181.19	philic	aromatic	Yes	3.81	5.37	3.59	-1.3
VAL	117.15	phobic	aliphatic	No	12.07	6.37	4.67	4.2

The features from the residue before and after the selected residue in the primary amino acid sequence were included to see if these features improved classification. Including the residue before or the residue after significantly improved classification; a further significant improvement was achieved if both the before and after residues were included, so features from all three residues were included in the final feature set.

Feature selection was done using RuleFit [41], an algorithm used with R [28]. RuleFit implements an ensemble learning methodology, identifying linear combinations of simple rules derived from the data to identify variables that are strongly indicative of the correct classification. The RuleFit algorithm provides a list of features in order of importance with a numerical value indicating the contribution to classification that each feature makes. Features were added in order of importance until classification achieved a plateau, at which time the feature set was finalized.

4.1.1 Proof of Concept Data Set

The feature selection procedure was applied to the Proof of Concept Data Set, and the results, a list of features with their corresponding importance in the data set, are summarized in Table 4-2.

Table 4-2: Importance from RuleFit for each feature.

Feature	RuleFit Importance
sT	100.0
Conservation	58.4
Molecular Weight	30.4
Volume	30.0
Trailing Residue	29.7
Trailing sT	27.8
Preceding Conservation	25.7
Preceding sT	23.6
Trailing Conservation	15.5
Potential	15.0
T1	12.7
Hydropathy	12.4
Trailing Volume	9.4
Trailing T3	8.7
Preceding T4	8.3
Preceding Hydropathy	7.9
Preceding Molecular Weight	7.6
Preceding T1	7.5
Preceding Volume	6.8
Trailing T4	6.3
T0	5.1
Trailing Hydropathy	5.0
Preceding Residue	4.8
Preceding Potential	3.7
Trailing Potential	3.7
T2	3.6
T4	3.5
Trailing T2	3.3
Preceding T3	3.1
Preceding T0	3.0
Trailing Side Chain	2.7
Trailing Molecular Weight	2.5
Trailing T0	2.4
Side Chain	2.0
Trailing T1	1.8
T3	1.8
Preceding T2	1.3
Preceding Side Chain	0.9
Residue	0.1

Features were added in order of importance until performance stabilized, resulting in a final feature set for the Proof of Concept Data Set of:

- From the residue being tested: potential, T0, T1, T2, sT, volume, conservation, molecular weight, hydropathy;
- From the residue preceding: residue, potential, T1, T4, sT, volume, conservation, molecular weight, hydropathy; and
- From the residue trailing: potential, T3, T4, sT, volume, residue, conservation, hydropathy.

4.1.2 Final Data Set

As described previously, feature selection was again performed using the RuleFit algorithm. The features selected as informative for the Final Data Set included:

- From the residue being tested: potential, total, volume, normalized volume, sT, conservation, molecular weight, hydropathy;
- From the residue preceding: volume, normalized T4, normalized total, normalized volume, sT, conservation; and
- From the residue trailing: volume, normalized total, normalized volume, sT, conservation, hydropathy.

Across both the Proof of Concept data set and the final data set, it is interesting to note that for both data sets, many of the same features are selected as most informative.

Tetrahedrality (sT) is consistently the most important feature for interface prediction. For a regular tetrahedron with four equilateral triangular faces, the

tetrahedrality is zero. The larger the tetrahedrality is, the further the tetrahedron is from regular. Internal residues demonstrate a lower value of the tetrahedrality, indicating more regularity, while those on the surface have higher values. Interface residues have the tetrahedrality values between the higher ones on the surface and the lower ones in the protein core. Since these residues begin on the protein surface, then become part of the core upon binding, this is reasonable.

Other consistently important features include the volume, the conservation, and the hydropathy. The average value of interface residue volumes is higher than the volume of those residues not on the interface (there is no significant difference between surface and core residue volumes). As expected, conservation is also a critical feature for interface residue classification. Surface residues show a high conservation, while those buried inside have a very low conservation. Interestingly, those surface residues not on the interface have, on average, a higher conservation than those on the interface. This may be a result of a few key interface residues which are highly conserved, while the remainder are not, so when the average conservation is taken across the interface, the values are lower. Hydropathy also adds significant information for residue classification. A higher values correlates to more hydrophobicity. Not surprisingly, core residues have a much higher hydropathy than those on the surface. As would be expected, those residues on the interface also have a higher hydropathy than those not on the interface, indicating their preference to be buried upon complex formation.

Some measure of the tetrahedral class (T0, T1, T2, T3, T4 or total) appears for every residue, but interestingly, not the same measure. Finally, the values for potential

show up in all of the residues for the proof of concept data set, but for only one of the residues for the comprehensive data set.

4.2 Classification

Classification tests and evaluation were performed using the Weka [128] software. Several of the classifiers included in the Weka framework were evaluated; performance of these classifiers on the Proof of Concept Data Set is included in Table 4-3.

Table 4-3: Performance of various classifiers on the proof of concept data set.

Classifier	Acc	Sen	Spec	FAR
Averaged One-Dependence Estimator	0.696	0.689	0.699	0.296
Bayes Network Classifier	0.663	0.652	0.666	0.327
Complement Class Naïve Bayes Classifier	0.604	0.437	0.656	0.229
Naïve Bayes Classifier using Estimator Classes	0.633	0.449	0.709	0.184
Naïve Bayes Multinomial	0.604	0.437	0.656	0.229
Simple Naïve Bayes Classifier	0.633	0.452	0.709	0.185
Updateable Naïve Bayes Classifier	0.633	0.449	0.709	0.184
Multinomial Logistic Regression Model	0.697	0.660	0.712	0.267
Multilayer Perceptron	0.677	0.670	0.680	0.316
Radial Basis Function Network	0.653	0.564	0.686	0.258
Logistic Regression Model with LogitBoost	0.693	0.668	0.704	0.281
Support Vector Machine	0.695	0.648	0.715	0.258
Voted Perceptron Algorithm	0.616	0.608	0.618	0.375
Winnow and Balanced Winnow Algorithms	0.579	0.586	0.578	0.427
IB1-type Classifier	0.604	0.602	0.605	0.394
K-nearest neighbors classifier	0.604	0.602	0.605	0.394
Instance-based Classifier	0.626	0.602	0.632	0.350
Lazy Bayesian Rules	0.708	0.729	0.699	0.314
Locally-weighted learning	0.643	0.827	0.604	0.542
HyperPipe Classifier	0.502	1.000	0.610	0.006
Voting Feature Interval Classifier	0.538	0.469	0.544	0.393
Alternating Decision Tree	0.681	0.682	0.680	0.321
Decision Stump	0.644	0.820	0.606	0.532
Ld3 Decision Tree Classifier	0.637	0.632	0.627	0.358
C4.5 Decision Tree	0.647	0.646	0.647	0.352
Logistic Model Tree	0.708	0.742	0.695	0.326
Naïve Bayes Tree	0.681	0.657	0.690	0.295
Random Forest	0.716	0.711	0.718	0.279
Random Tree	0.602	0.593	0.603	0.390
Fast Decision Tree Learner	0.670	0.690	0.664	0.349
User Defined Decision Tree	0.500	0.000	0.000	0.000
Single Conjunctive Rule Learner	0.643	0.857	0.600	0.572
Simple Decision Table Majority Classifier	0.665	0.711	0.651	0.381
Repeated Incremental Pruning to Produce Error Reduction	0.690	0.713	0.681	0.333
Separate & Conquer	0.678	0.685	0.676	0.329
Nearest-neighbor like using non-nested generalized exemplars	0.623	0.634	0.621	0.388
1R Classifier	0.515	0.529	0.514	0.500
Partial Decision Trees Decision List	0.649	0.698	0.636	0.400
Ripple-Down Rule Learner	0.647	0.668	0.640	0.375
0-R Classifier	0.500	0.000	0.000	0.000

The protein residue classification technique chosen, based both on the results presented here and previous work indicating that the algorithm worked well, was the Random Forest technique. All future training and testing performed on the interface classification data sets reported here were done using Random Forest classification.

4.2.1 Proof-of-Concept Data Set

As seen in Table 4.3, initial classification accuracy was 0.716, with a sensitivity of 0.711, a specificity of 0.718, and a false alarm rate of 0.279. For each of several criteria, the entire data set was split into subsets by binning the data according to the value for the criteria being tested. The performance on the data subset was compared to the performance on the data as a whole to understand if there was any benefit to performing training and testing on subsets of the data.

The first criteria used to separate the data was protein chain length. The data were split into smaller proteins, with a length of less than 46 residues, and larger proteins, with a length of 46 or more residues. Accuracy on the dataset of smaller proteins was 0.709 if the training and testing sets both only had smaller proteins, versus an accuracy of 0.646 when the test set contained the smaller proteins, but the training set included both small and large proteins. Similarly, accuracy on the dataset of larger proteins was 0.706 if the training set consisted of only larger proteins, and 0.704 if the training set contained proteins of all sizes (in both cases, the testing set contained only the larger proteins).

Next, the data were separated according to the amino acid of the sample. For this classification, molecular weight and hydropathy, which are specific to the amino acid,

were removed. Data for the classifiers designed for each amino acid are included in Table 4-4. It is important to note that when the data sets were limited to those of only a specific amino acid, some of the data sets were quite small.

Table 4-4: Results when a classifier was designed for each amino acid. For each classifier, the size of the data set used is listed, as well as the accuracy on a data set consisting of only the specific residue type (Specific Data Set), as opposed to the performance when the same size data set was taken from samples of the entire data set (Normal Data Set).

Amino Acid	Data Set Size	Specific Data Set	Normal Data Set
ALA	706	69.4	74.8
ARG	718	64.6	60.3
ASN	484	62.2	66.0
ASP	542	63.0	62.4
CYS	106	78.8	83.1
GLN	480	65.9	65.9
GLU	620	66.0	61.5
GLY	510	66.0	70.4
HIS	262	65.3	68.0
ILE	572	74.4	76.9
LEU	870	76.5	78.3
LYS	552	61.6	64.9
MET	304	73.3	73.3
PHE	502	74.6	74.7
PRO	428	64.3	63.9
SER	640	66.0	69.4
THR	604	67.5	67.3
TRP	122	64.2	72.1
TYR	568	68.0	67.8
VAL	638	78.2	79.7

Next, the data were split by their values for the four-body statistical potential. A threshold was set at 2.2, the value of the median plus one standard deviation. The data set containing samples with a potential energy above 2.2 displayed an accuracy of 0.761, while a data set of the same size containing random samples achieved an accuracy of

0.844. The data set containing samples with potential energy less than 2.2 had an accuracy of 0.707, while an equally sized normal data set had an accuracy of 0.676.

The data set was split into two subsets depending on their value for T0; those with a value less than the threshold of 13 were put in one data set, and those with values above the threshold were put in another. The threshold was chosen as the average plus two standard deviations. The data set with higher values of T0 achieved a classification accuracy of 0.725, versus 0.701 with a randomly selected normal data set, and the accuracy of the data set with the lower values was 0.712, while the randomly selected normal data set achieved an accuracy of 0.718.

Also considered was the volume of the tetrahedron the residue is involved in. Data were put into two different data sets depending on the value of the volume. A threshold of 56.7, the mean plus two standard deviations, was chosen. The data set with higher values had a classification accuracy of 0.643, while the normal data set demonstrated an accuracy of 0.568. The classification accuracy of the data set with lower values was 0.709, as opposed to 0.707 for the normal data set.

Because of the critical role conserved residues have been found to play in protein interfaces, the data were split into three groups by their conservation values. Residues with a conservation value less than the median were considered to have lower conservation values, and classification on this data set was 0.738, as compared to 0.718 on the normal data set. Those residues with a conservation value above the median plus one standard deviation are included in the group with higher conservation values, with a classification of 0.661 compared to 0.721 on the normal data set. The remaining data

were grouped in the middle class, and demonstrated an accuracy of 0.667 for both the experimental group and the normal group.

The data were then separated into three groups according to their molecular weights. The group with lower weights included Glycine (75.07), Alanine (89.09), and Serine (105.09); the middle weights included Proline (115.13), Valine (117.15), Threonine (119.12), Cysteine (121.16), Isoleucine (131.17), Leucine (131.17), Asparagine (132.12), and Aspartic Acid (133.1). The group with higher weights included Glutamine (146.15), Lysine (146.19), Glutamic Acid (147.13), Methionine (149.21), Histidine (155.16), Phenylanine (165.19), Arginine (174.2), Tyrosine (181.19), and Tryptophan (204.23). The low weight, middle weight, and high weight groups had accuracies of 0.680, 0.707, and 0.692, respectively, versus identical size normal groups that achieved accuracies of 0.719, 0.724, and 0.662, respectively.

Finally, the data were separated according to their values of hydropathy (which varied by amino acid). Data with lower values (less than -1.5) had an accuracy of 0.769 versus 0.637 for an equivalent normal data set; data with middle values, between -1.5 and 1.5, had accuracies of 0.693 versus 0.690 for the equivalent normal data set. The data with higher values (above 1.5) had an accuracy of 0.680, while the normal data set displayed an accuracy of 0.771.

Unfortunately, none of these data subsets significantly improved classification, and since they required an increase in the complexity of the classifier, the data were not split into additional classifiers based on this information.

Another test included implementing a final filter that classified isolated residues similarly to their neighbors. That is, if the residue was found to be on the interface when none of its surrounding members were on the interface, the residue was re-classified as not on the interface. Similarly, a residue that was found not to be on the interface when all those surrounding were on the interface was re-classified as part of the interface. Since most of the classification errors occurred at the interface boundaries, this filter did not improve the classification, and was not included in the final algorithm.

With the final feature list enumerated above, 10-fold cross validation was performed ten times. Overall average performance was 0.721 accuracy, 0.717 sensitivity, 0.723 specificity, and false alarm rate of 0.275. To gain further understanding of the limitations of the classification method, a leave-one-out training and testing technique was used to evaluate the data set. Average performance for the leave-one-out set was 0.697 accuracy, 0.792 sensitivity, 0.500 specificity, and a false alarm rate of 0.381 (the entire analysis of the data set is included in Appendix A).

Figure 4-1 shows examples of the best (a and b), middle (c and d), and worst (e and f) attempts at classification in the Halperin data set. In each sub-figure, there are four pictures. The top two are different views of the protein with the correct interface residues colored; the bottom two are the same views of the protein, but with the predicted interface residues colored. In sub-figures a and b, examples of the best performance, the predicted interface residues (the bottom two pictures) are very close to the actual interface residues (the top two pictures). Similarly, for sub-figures e and f, the predicted interface residues do not accurately reflect the true interfaces shown in the top pictures.

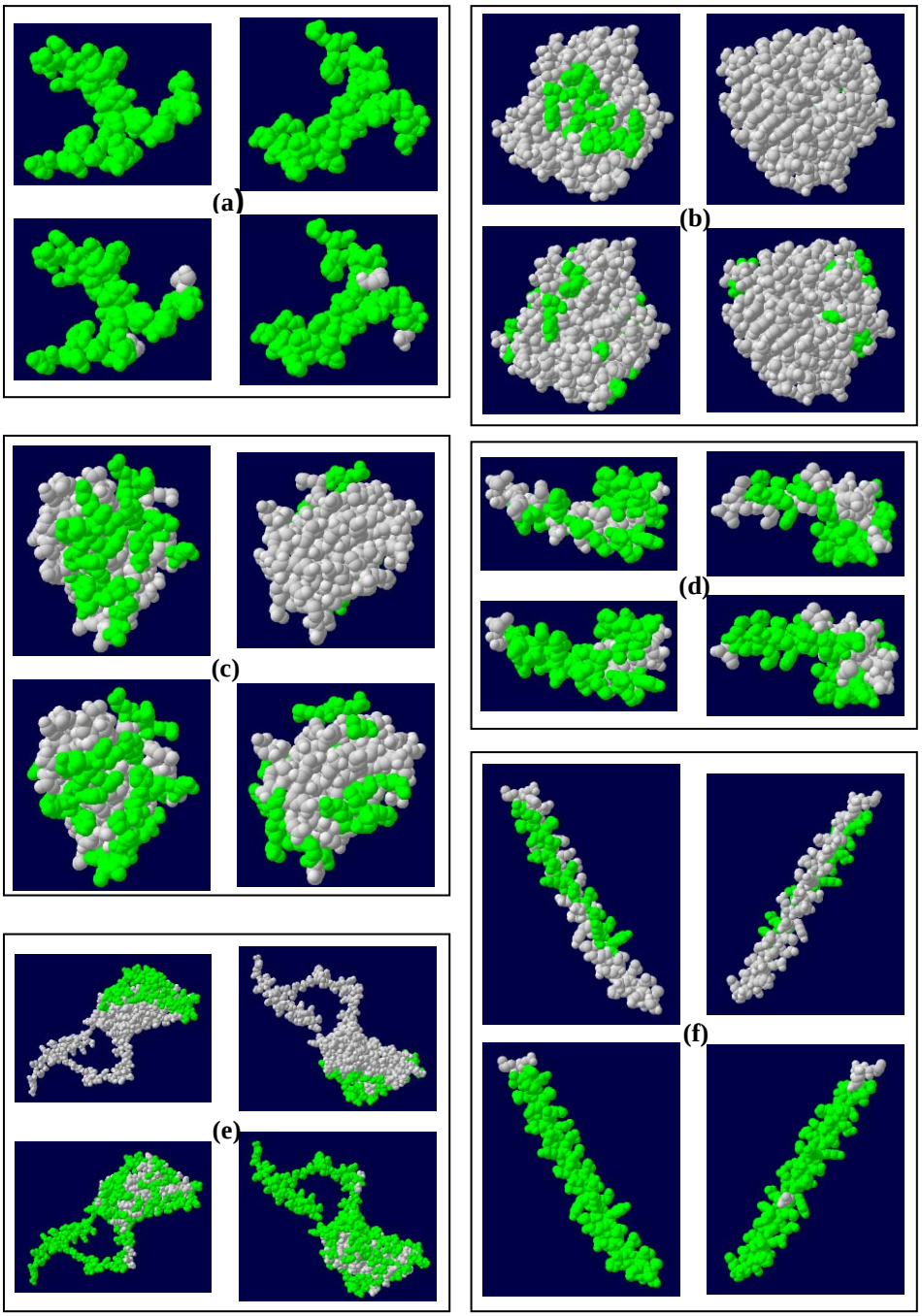

Figure 4-1: Examples of performance of the classifier on different protein chains. (a) 1C8O-B, (b) 2SNI-E, (c) 1AKJ-E, (d) 1BH8-A, (e) 1B35-B, and (f) 1KQL-B.

4.2.2 Final Data Set

A summary of the classification results (on the Final Data Set) is included in Table 4-5, and an example is provided in Figure 3-3, where the correct and predicted residues can be compared for protein 2JD3. Tests were done using 10-fold cross validation (CV), 66%-34% data split (DS), and leave-one-out (LOO) training and testing methods. The analysis of the entire LOO data set is included in the Appendix B; Figure 4-2 shows the range of distributions for the accuracies calculated for each of the 1476 proteins.

Table 4-5: Summary of classification results.

	Accuracy	**Sensitivity**	**Specificity**	**FAR**
Cross Validation	0.836	0.355	0.37	0.091
Data Split	0.862	0.198	0.43	0.039
Leave-One-Out	0.858	0.240	0.43	0.049

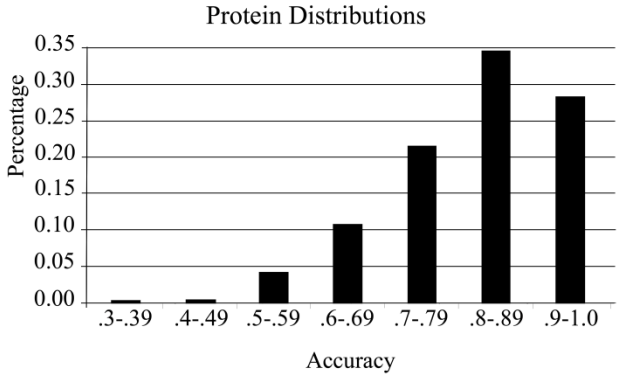

Figure 4-2: Range of classification accuracies for each protein in the data set using leave-one-out testing.

As seen in Figure 4-2, the range of accuracies is large, but the majority are in the 0.7 to 0.9 range. There are a small number of proteins (11 of the 1476) who have accuracies less than 0.5, indicating that performance on these proteins is worse than random. The size of either the protein or the binding interface may possibly have an impact on the classifier performance; many of the poorly classified chains were smaller proteins or had larger binding interfaces. To further investigate this, the data were binned by either the length of the chain (Table 4-6a, top of Figure 4-3), the number of residues on the interface (Table 4-6b, middle of Figure 4-3), or the ratio of interface to total residues (Table 4-6c, bottom of Figure 4-3). The size of the bins was selected to include approximately the same number of proteins per bin, while still maintaining a reasonable range for the bins to span. By binning the data in this way, overall trends of the data could be investigated without the noise included when each data point is considered. In Figure 4-3, the accuracy is plotted against the length of the protein, the number of residues in the interface, and the percent of residues involved in the interface.

Table 4-6a: Binning of data by length of protein.

Bin (L = Chain Length)	Median Length	Mean Accuracy	Accuracy Standard Deviation	Number of Proteins
1: $L < 120$	89.5	0.702	0.11	348
2: $120 \leq L < 185$	149.0	0.812	0.09	377
3: $185 \leq L \leq 300$	239.0	0.859	0.09	380
4: $L > 300$	402.0	0.895	0.07	371

Table 4-6b: Binning of data by number of interface residues.

Bin (N = Number of Interface Residues)	Median Number	Mean Accuracy	Accuracy Standard Deviation	Number of Proteins
1: N < 12	6.0	0.891	0.11	358
2: 12 ≤ N < 24	17.0	0.841	0.10	372
3: 24 ≤ N ≤ 39	30.0	0.793	0.11	365
4: N > 39	55.0	0.755	0.10	381

Table 4-6c: Binning of data by the ratio of residues on the interface to the total number of residues.

Bin (R = Ratio of Interface to Total Residues)	Median Ratio	Mean Accuracy	Accuracy Standard Deviation	Number of Proteins
1: R < 4.25	1.9	0.925	0.07	297
2: 4.25 ≤ R < 10	7.3	0.891	0.06	300
3: 10 ≤ R < 17	13.1	0.836	0.06	307
4: 17 ≤ R < 29	22.1	0.768	0.07	306
5: R ≥ 29	40.4	0.659	0.08	266

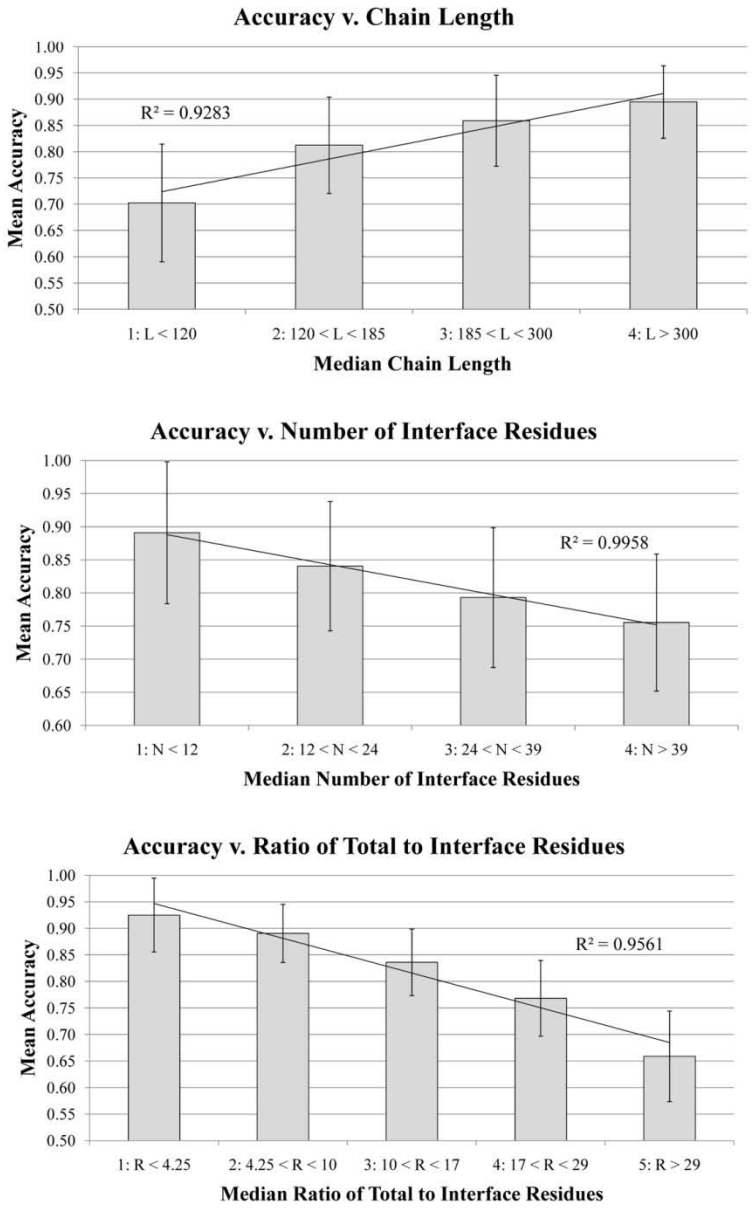

Figure 4-3: Classification accuracy plotted against: length of protein (top), number of residues on the interface (middle), and percent of residues involved in the interface (bottom).

It appears that the classifier may not perform as well on smaller proteins. To test this, the data were split by amino acid length into three groups: large proteins (L-492),

medium length proteins (M-494), and small proteins (S-490). Because the number of proteins was split relatively evenly between the groups, the total number of residues varied. To compensate for this, decoy data sets (D-492, D-494, and D-490) were created by randomly selecting the same number of positive and negative residues as found in the corresponding data set. The results are reported in Table 4-7 below.

Table 4-7: Summary results if data sets are created with only small, medium, or large proteins.

Data Set	Accuracy	Sensitivity	Specificity	FAR
L-492	0.894	0.134	0.398	0.022
D-492	0.895	0.149	0.420	0.023
M-494	0.871	0.075	0.451	0.013
D-494	0.861	0.119	0.514	0.018
S-490	0.724	0.479	0.448	0.195
D-490	0.740	0.537	0.479	0.193

While the size of the protein seems to correlate with lower accuracy values, it doesn't appear that the smaller size causes them. The smaller proteins were further investigated to understand why classification performance was lower. It was observed that many of the proteins fell into certain classes, and these classes were more highly represented in the proteins with lower classification accuracies than across the entire data set, as seen in Table 4-8. It can be seen that the general classifier does not perform as well on some specific types of proteins.

Table 4-8: Representation of classes of protein.

Category	% of proteins below 50% accuracy in category	% of proteins below 60% accuracy in category	% of category in entire data set
DNA/RNA binding	9.1% (1/11)	21.9% (16/73)	9.5% (140/1476)
Viral	18.2% (2/11)	15.1% (11/73)	6.4% (94/1476)
Cell Cycle	9.1% (1/11)	2.7% (2/73)	1.2% (17/1476)
Coiled Coil	27.3% (3/11)	6.8% (5/73)	0.3% (5/1476)
Mitochondrial	18.2% (2/11)	13.7% (10/73)	2.0% (29/1476)

4.3 Comparison with Other Methods

In order to compare the newly developed method to similar methods, a subset of the data set was randomly selected and used to test four other interface prediction methods: cons-PPISP, SPPIDER, PPI-Pred, and ProMate; these methods were chosen because they are also structural-based methods. The data subset consisted of 55 proteins, with a total of 11,794 residues. The mean number of residues per protein is 216.3, with a mean of 27.0 residues on the interface. All results are included in Appendix C, and are summarized below in Table 4-9.

Table 4-9: Comparison of several methods of interface identification.

	Acc	Sen	Spec	FAR	MCC	BER
Topological	0.867	0.284	0.442	0.051	0.28	0.38
PPI-PRED	0.769	0.302	0.294	0.135	0.15	0.42
cons-PPISP	0.803	0.288	0.347	0.106	0.18	0.41
ProMate	0.810	0.075	0.269	0.045	0.06	0.48
SPPIDER	0.768	0.526	0.315	0.229	0.25	0.35

As shown in Figure 4-3, the Topological Description Classifier demonstrates better performance on proteins which are larger and have fewer residues on the interface. In order to investigate if there is a specific class of proteins for which the Topological Description Classifier performs better than other classifiers, the data subset was further broken down. First, a subset was created on which the Topological Descriptor Classifier would be expected to perform well: proteins larger than 300 residues and with an interface smaller than 24 residues. This data set had 13 proteins with 4,660 residues, and a mean number of 359.8 residues, of which 11.2 were interface residues. Results on this data set are included in Table 4-10.

Table 4-10: Method comparison on a data set with larger proteins and smaller interfaces.

	Acc	Sen	Spec	FAR	MCC	BER
Topological	0.944	0.069	0.074	0.028	0.20	0.48
PPI-PRED	0.859	0.400	0.092	0.126	0.14	0.36
cons-PPISP	0.936	0.193	0.132	0.041	0.13	0.42
ProMate	0.958	0.048	0.109	0.013	0.05	0.48
SPPIDER	0.918	0.400	0.164	0.065	0.22	0.33

Alternatively, a data set was considered with smaller proteins, with 200 or fewer residues, and interfaces with 24 or more residues. These are the characteristics of the proteins on which the Topological Descriptor Classifier has the worst performance. This data set contained 18 proteins, with a total of 2,283 residues. There was a mean of 126.8 residues per protein, and 37.4 of those residues on the protein. Analysis of this data set is included below in Table 4-11.

Table 4-11: Method comparison on a data set with smaller proteins and larger interfaces.

	Acc	**Sen**	**Spec**	**FAR**	**MCC**	**BER**
Topological	0.727	0.325	0.564	0.105	0.27	0.39
PPI-PRED	0.710	0.289	0.532	0.109	0.22	0.41
cons-PPISP	0.726	0.357	0.570	0.115	0.28	0.38
ProMate	0.703	0.125	0.520	0.050	0.13	0.46
SPPIDER	0.706	0.704	0.507	0.293	0.38	0.29

Finally, a data set was developed with the remainder of the data subset, consisting of those proteins that are either small with small interfaces or large with large interfaces. This set consisted of 23 proteins with a total of 4,743 residues, and a mean length of 206.2 and a mean interface size of 28.8 residues. Table 4-12 has a summary of the results on this data set.

Table 4-12: Method comparison on a data set of proteins that are either small with small interfaces or large with large interfaces.

	Acc	**Sen**	**Spec**	**FAR**	**MCC**	**BER**
Topological	0.854	0.291	0.458	0.056	0.29	0.38
PPI-PRED	0.803	0.290	0.292	0.114	0.18	0.41
cons-PPISP	0.834	0.238	0.358	0.069	0.20	0.42
ProMate	0.839	0.042	0.178	0.032	0.02	0.49
SPPIDER	0.809	0.599	0.382	0.157	0.37	0.28

From these data, it can be seen that the new method performs comparably to other methods on all types of proteins. There does not appear to be a specific subset of protein types that the Topological Description Classifier performs significantly better or

significantly worse on when compared to other classifiers; rather it appears to consistently perform slightly better on all protein types.

For the Topological Description Classifier, both the training and the testing set could be controlled, but for the other classifiers, only the testing set could be determined. It is possible that the training sets specified by the other classifiers had homologous, or even identical, proteins included. To test this, a data set was developed with additional homologous proteins, for a total of 6,091 proteins (1,283,795 residues). On this data set, the Topological Description Classifier achieved 0.902 accuracy, with sensitivity of 0.753, specificity of 0.688, and FAR of 0.068. The addition of homologous data significantly improves performance of the new classifier.

In order to investigate if the data classification would improve by further subdividing the types of proteins, the data set from Guharoy and Chakrabarti [50] was analyzed; these data included 15 obligate homodimers, and 114 nonobligate heterodimers. The nonobligate data was further split into 31 chains with only a single partner, and 83 chains with more than one partner. The classifier accuracy results are included below in Table 4-13; ten trials were averaged for the results displayed.

Table 4-13: Accuracy results on data split into obligate and non-obligate subsets.

All Data	**Obligate Data**	**All Non-obligate Data**	**Non-obligate nmers (n=2)**	**Non-obligate nmers (n>2)**
0.737	0.757	0.739	0.680	0.759

The results of this test indicate that performance may be improved by splitting the data into separate subsets including only obligate or non-obligate proteins.

A further test was performed using a data set of proteins that bind with DNA to see if the same classifier could be trained to predict those residues which bound to DNA. The data set [103] used had 693 proteins. Of these proteins, 91 were missing a C_α, and couldn't be tessellated, and one of the listed proteins (1AN2-C) didn't have the chain listed, only A and B. 54 proteins did not have enough homologous proteins in UniProt or SwissProt to achieve results from ConSurf. This left 547 proteins, with 10,512 residues that bound to DNA, and 79,957 that did not. All the residues that bound to DNA were included in the data set, and 10,512 of the residues that were not on the interface were randomly selected to be included in the data set. 10-fold cross validation was repeated 10 times using Random Forest classification, resulting in an average accuracy of 0.902. The unusually high accuracy of the classifier is probably due to a high degree of redundancy in the data set used, but these results at least suggest that the method described here could be used to predict those residues which bind to DNA as well as those residues which bind to other proteins.

CHAPTER 5: RESULTS FOR DOCKING RE-SCORING

While the underlying methodology for the docking scoring and interface prediction classifiers is the same, there are differences in both the features selected as informative and the classifiers themselves. This chapter discusses the features selected (Section 5.1), the performance of the classifier (Section 5.2), and a comparison of the new docking re-scoring method with others (Section 5.3).

5.1 Dock Scoring Feature Selection

As with the interface prediction classifier feature selection process, several features were considered for incorporation into the final scoring classifier. In this instance, features were not calculated for individual residues (as in the interface prediction feature selection), but for the entire protein conformation. From the topological descriptor, features considered included:

- Mean volume for each of the six simplex types (as described in Section 3.2) – 6 features
- Total four-body statistical potential energy function for the whole complex – 1 feature
- Mean four-body statistical potential energy function over all residues – 1 feature

- Mean four-body statistical potential energy function for interface residues – 1 feature
- Ratio of mean four-body statistical potential energy function for interface residues to mean four-body statistical potential energy function for all residues – 1 feature
- Mean value of each simplex type (T0, T1, T2, T3, T4, T5) for interface residues – 6 features
- Ratio of mean value of each simplex type (T0, T1, T2, T3, T4, T5) for interface residues to mean value of each simplex type (T0, T1, T2, T3, T4, T5) for all residues – 6 features
- Mean total number of simplices interface residues participate in – 1 feature
- Ratio of mean total number of simplices interface residues participate in to mean total number of simplices all residues participate in – 1 feature
- Mean volume of interface simplices – 1 feature
- Ratio of mean volume of interface simplices to mean volume of all simplices – 1 feature
- Mean tetrahedrality of interface residues – 1 feature
- Ratio of mean tetrahedrality of interface simplices to mean tetrahedrality of all simplices – 1 feature
- Mean volume of simplices which cross the interface – 1 feature
- Ratio of mean volume of simplices which cross the interface to mean volume of all simplices – 1 feature

A number of additional features were considered for inclusion in the classifier that were thought to potentially be informative:

- Number of residues on the interface for each protein – 2 features
- Fraction of residues on the interface for each protein – 2 features
- Total number of interface residues – 1 feature
- Ratio of total interface residues to total residues in the complex – 1 feature
- For each of the 20 amino acids, ratio of number of amino acids on the interface to total count of that amino acid in the protein – 20 features
- For each of 6 categories (hydrophobic, aromatic, positively charged, negatively charged, polar, and small), the ratio of each pair of interface interactions to the total number of interface interactions – 15 features
- Mean conservation of interface residues – 1 feature
- Ratio of mean conservation of interface residues to mean conservation of all residues – 1 feature

5.1.1 Initial Feature Selection

As with the feature selection for the interface selection method, the features were selected using the RuleFit algorithm. Table 5-1 has a list of the top 25 features (all features are included in Appendix D) and the mean importance over the ten trials.

Table 5-1:RuleFit importance of scoring features on original data set.

Feature	RuleFit Importance
Mean interface residue tetrahedrality	100.0
Mean interface residue tetrahedrality / mean residue tetrahedrality	46.7
Ratio of interactions of class aromatic-small	44.2
Mean interface residue T5	42.1
Number of interface residues for protein A	41.8
Total number of interface residues	38.7
Ratio of interface / total residues for protein B	34.8
Ratio of interactions of class hydrophobic-aromatic	33.9
Mean interface residue volume / mean residue volume	33.2
Mean interface residue potential	30.6
Mean conservation of interface residues	26.7
Ratio of interface to total number of CYS residues	24.3
Mean interface residue potential / mean residue potential	21.9
Raton of interactions of class positively charged-negatively charged	21.7
Mean volume for T0 simplices	20.4
Number of interface residues for protein B	20.2
Total volume of simplices that cross interface / total volume of both chains	18.8
Ratio of interface to total number of SER residues	17.4
Ratio of interface to total number of TRP residues	17.4
Ratio of interface to total number of PRO residues	17.2
Ratio of interface to total number of VAL residues	15.1
Mean interface residue conservation / mean conservation of all residues	14.5
Mean interface residue T4	14.4
Ratio of interface to total number of LEU residues	13.6
Ratio of interface to total number of GLN residues	13.4

Features were added one at a time until classification accuracy reached a plateau (trials were repeated three times). The ideal number of features for this data set is 18, and includes:

- Mean interface residue tetrahedrality
- Mean interface residue tetrahedrality / mean residue tetrahedrality
- Ratio of interactions for classes aromatic-small, hydrophobic-aromatic, and positively charged-negatively charged (3 features)

- Mean interface residue T5
- Number of interface residues for each protein (2 features)
- Total number of interface residues
- Ratio of interface / total residues for protein B
- Mean interface residue volume / mean residue volume
- Mean volume for T0 simplices
- Total volume of simplices that cross interface / total volume of both chains
- Mean interface residue potential
- Mean interface residue potential / mean residue potential
- Mean conservation of interface residues
- Ratio of interface to total number for cysteine and serine residues (2 features)

5.1.2 Feature Selection after Addition of Data

After the addition of more randomly selected data, the feature selection algorithm was run again, and another list of features were selected for this data set. The same features were used after the addition of homologous data. Again, the top 25 features are included in Table 5-2, with the entire results included in Appendix D.

Table 5-2:RuleFit importance of scoring features on data set with additional data.

Feature	RuleFit Importance
Mean interface residue tetrahedrality	100.0
Mean interface residue tetrahedrality / mean residue tetrahedrality	46.9
Mean interface residue T5	42.7
Number of interface residues for protein A	40.8
Ratio of interactions of class aromatic-small	38.3
Ratio of interactions of class hydrophobic-aromatic	34.4
Total number of interface residues	33.0
Mean interface residue potential	31.8
Ratio of interface / total residues for protein B	31.3
Mean conservation of interface residues	26.5
Mean interface residue volume / mean residue volume	25.8
Number of interface residues for protein B	25.5
Ratio of interface to total number of CYS residues	24.9
Mean interface residue potential / mean residue potential	19.3
Ratio of interface to total number of PRO residues	18.6
Mean volume for T0 simplices	18.0
Ratio of interactions of class positively charged-negatively charged	17.5
Ratio of interface to total number of TRP residues	17.3
Ratio of interface to total number of SER residues	16.7
Mean interface residue T4	15.2
Ratio of interface to total number of GLN residues	13.0
Ratio of interface to total number of VAL residues	12.8
Total volume of simplices that cross interface / total volume of both chains	12.8
Ratio of interface to total number of GLY residues	11.8
Volume of simplices that cross interface	11.8

On this dataset, the ideal number of features is 14, with the final feature list for the data set supplemented with additional data including:

- Mean interface residue tetrahedrality
- Mean interface residue tetrahedrality / mean residue tetrahedrality
- Mean interface residue T5
- Number of interface residues for each protein (2 features)
- Total number of interface residues

- Ratio of interface / total residues for protein B
- Ratio of interactions for classes aromatic-small and hydrophobic-aromatic (2 features)
- Mean interface residue potential
- Mean interface residue potential / mean residue potential
- Mean conservation of interface residues
- Mean interface residue volume / mean residue volume
- Ratio of interface to total number of cysteine residues

5.1.3 Feature Selection for Antibody-Antigen Data Subset

A subset of the overall data set was created for just the antibody-antigen complexes. The results of the RuleFit algorithm on these data are included below in Table 5-3 for the top 25 features, and in Appendix D for all features.

Table 5-3:RuleFit importance of scoring features on antibody-antigen data subset.

Feature	**RuleFit Importance**
Ratio of interactions of class aromatic-small	98.9
Ratio of interface to total number of ASN residues	87.8
Mean interface residue tetrahedrality	71.4
Mean conservation of interface residues	69.3
Mean interface residue tetrahedrality / mean residue tetrahedrality	62.9
Mean interface residue conservation / mean conservation of all residues	61.0
Ratio of interface to total number of ILE residues	60.8
Ratio of interface to total number of ASP residues	51.8
Ratio of interface to total number of GLU residues	51.2
Ratio of interface to total number of GLN residues	47.4
Mean volume for T1 simplices	47.1
Mean volume for T2 simplices	44.7
Ratio of interface to total number of PRO residues	42.1
Ratio of interface to total number of THR residues	36.8
Ratio of interface to total number of VAL residues	32.6
Ratio of interaction of class aromatic-negatively charged	31.7
Ratio of interface to total number of TRP residues	30.8
Total volume of simplices that cross interface / total volume of both chains	27.7
Total number of interface residues	25.8
Ratio of interactions of class aromatic-positively charged	25.7
Ratio of interface to total number of PHE residues	25.4
Ratio of interface to total number of TYR residues	24.7
Ratio of interface to total number of SER residues	23.1
Mean interface residue T5	22.6
Number of interface residues for protein B	20.6

The ideal number of features, 24, on the antibody-antigen data subset included:

- Ratio of interactions for classes aromatic-small, aromatic-negatively charged, and aromatic-positively charged (3 features)
- Ratio of interface to total number of residues for asparagine, aspartic acid, glutamic acid, glutamine, isoleucine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine (12 features)
- Mean interface residue tetrahedrality

- Mean interface residue tetrahedrality / mean residue tetrahedrality
- Mean conservation of interface residues
- Mean interface residue conservation / mean conservation of all residues
- Mean volume for T1 and T2 simplices (2 features)
- Total volume of simplices that cross interface / total volume of both chains
- Total number of interface residues
- Mean interface residue T5

5.1.4 Feature Selection for Enzyme-Inhibitor Data Subset

Another data subset was developed for just the enzyme-inhibitor data. The top 25 features that were found using RuleFit for this data set are included in Table 5-4; the entire table can be found in Appendix D.

Table 5-4:RuleFit importance of scoring features on enzyme-inhibitor data subset.

Feature	**RuleFit Importance**
Ratio of interface to total number of SER residues	96.8
Ratio of interface / total residues for protein B	85.0
Mean interface residue tetrahedrality	85.0
Ratio of interface to total number of CYS residues	80.4
Mean interface residue T4	71.1
Mean conservation of interface residues	69.0
Total number of interface residues	68.2
Ratio of interactions of class hydrophobic-aromatic	60.7
Mean interface residue T5	53.8
Mean interface residue T4 / mean residue T4	51.6
Ratio of interface to total number of PHE residues	47.8
Mean interface residue T5 / mean residue T5	43.2
Mean interface residue tetrahedrality / mean residue tetrahedrality	38.7
Ratio of interface to total number of GLN residues	32.2
Ratio of interface to total number of TRP residues	28.8
Mean interface residue T3	26.0
Ratio of interface to total number of ARG residues	25.2
Mean volume for T0 simplices	24.6
Number of interface residues for protein A	23.8
Ratio of interface to total number of HIS residues	21.6
Mean interface residue volume / mean residue volume	19.5
Mean interface residue conservation / mean conservation of all residues	18.5
Volume of simplices that cross interface	18.1
Mean interface residue potential	15.9
Mean volume for T5 simplices	15.6

The ideal number of features for this data set is 15, and includes:

- Ratio of interface to total number of residues for cysteine, glutamine, phenylalanine, serine, and tryptophan (5 features)
- Ratio of interface / total residues for protein B
- Total number of interface residues
- Mean interface residue tetrahedrality
- Mean interface residue tetrahedrality / mean residue tetrahedrality

- Mean interface residue T4 and T5 (2 features)
- Mean interface residue / mean residue for T4 and T5 (2 features)
- Mean conservation of interface residues
- Ratio of interactions of class hydrophobic-aromatic

5.1.5 Selected Feature Comparison

When comparing the features selected for the different subsets of data, the original data set and the data set supplemented with additional data identified many of the same features as important, although in a slightly different order. In fact, the original feature set is a superset of the supplemented data feature set, with the addition of four more features: the impact of ionic interactions, the mean volume of T0 simplices, the percent of the total complex volume that the interface takes up, and the percent of serine residues on the interface.

Those features which are represented in both data sets reveal characteristics of the protein complex formations. As with the prediction of interface residues, the tetrahedrality plays the most critical role in definition of the correct docking conformations. Also important are the interactions across the interface of two classes: aromatic and small residues, and aromatic and hydrophobic residues. This is not surprising as aromatic residues have been found to facilitate differentiation of the binding interface from the remainder of the protein interface [61, 66, 79, 87, 100]. Another critical feature is the mean number of the T5 residues for the complex. The T5 residues are those which cross the interface, and can be considered a measure of the size of the

protein interface. There are several other features which also indicate the size of the protein interface, including the number of interface residues for each chain of the protein, the total number of interface residues, the percent of volume for a protein that the interface takes, and the mean size of the interface residue volumes as compared to the mean size of all residues. Two features focus on the four-body statistical potential: the mean interface residue potential, and the ratio of this value to the mean potential for all residues. On average, the potential and the ratio are lower for the native conformations than for non-native ones. This indicates there is a higher bias for four residues to occur together in a simplex away from the interface than on the interface. Another informative feature is the conservation; this was seen in the interface prediction feature selection as well and is not surprising. The final feature is the ratio of interface to total number of cysteine residues; as cysteine normally has a very low representation on the protein interface, this is expected.

The antibody-antigen data subset feature list has some interesting differences from the feature list developed when considering all of the data. Most of the features are identical or related to those from the full feature set except there is no inclusion of the four-body statistical potential, indicating that there is no pattern to the occurrence of amino acids in the simplices. However, two related metrics, the mean volume for the T1 simplices and for the T2 simplices appear in this feature list but not the other, so the primary sequence relationship does have some impact on the correct conformation. This happens because the different conformations result in different tessellations of the interface residues, impacting the tetrahedra surrounding the interface residues. So those

surrounding tetrahedral volumes indicate whether the protein is in the correct conformation or not. Finally, while the percent of cysteine residues in the interface is not included, the percent of several other amino acids is, including asparagine, aspartic acid, glutamic acid, glutamine, isoleucine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine. The specific amino acid composition of the interface for antibody-antigen complexes has more of an impact than when the entire data set is considered.

The final feature list developed was for the enzyme-inhibitor data subset. The differences from both the total data set and the antibody-antigen data set are interesting. As with the antibody-antigen set, all of the features from the total data set are included except for the four-body statistical potential. And, similar to the antibody-antigen data set, a type of simplex has shown as informative; in this case, the mean interface residue and ratio of mean interface residue to mean residue for T4 type simplices. Finally, several of the amino acid residue type were important in classification: cysteine, glutamine, phenylalanine, serine, and tryptophan. It appears that if the data are separated by the type of interaction, the type of amino acid plays a far more crucial role in determining the correct conformation.

While the features were not selected to approximate the terms of the energetic of binding, several of the terms relate to them. The interface area, which was selected in some form for all data sets, is linearly related to the free enthalpy contribution of the hydrophobic effect [10]. The residue volumes describe the atomic packing, which is directly related to the van der Waals energy [10]. The amino acid percents, selected as

informative for both the antibody-antigen and enzyme-inhibitor data subsets, distinguish the interface from the rest of the protein surface and correlate with the desolvation free energy [10].

5.2 Docking Conformation Scoring Classifier

While the same Weka software was used for classification, the classifiers for interface prediction and docking scoring were quite different. The most significant difference is that a binary classifier is used for interface prediction, and a continuous classifier is used for the docking. As described in the Chapter 3, all classification for the docking re-scoring work was done using the Least Median Square Linear Regression classifier. The correct output of this classifier can be set as either the Root Mean Square (RMS) from the native conformation or the correct rank of the conformations. The first classification test involved determining if the classifier should be trained to output the RMS value or the rank value. Performing a simple 3-fold cross validation test (the proteins were randomly split into 3 groups of 18 proteins each), the mean position of the native conformation when RMS was used was 142.8, while when rank was used, the mean position was 161.1, so all future tests used RMS.

5.2.1 Classification Results on Original Data Set

With the exception of the test selecting RMS or rank, all testing was done using leave-one-out classification, where the data from all proteins except the one being tested were used to train the classifier (all results are included in Appendix E, and the best

results for all proteins are included in Section 5.2.4). On the original data set, the median position of the native conformation is 110.5 (out of approximately 1000), while the median RMS of the predicted top conformation is 13.97 angstroms. The median fraction of correct interface residues identified on the larger molecule was 0.611, with a median fraction of 0.620 for the smaller molecule. The median highest ranked near-native conformation (less than 5 angstroms from the native) is in position 21.5.

These data were then re-ranked to see if performance could be improved: the top 200 predicted conformations from each protein were used to create a new training and testing dataset. The complete results are included in Appendix E, but there was no improvement in performance. Fifteen of the 54 proteins did not have the native conformation in the top 200, so re-ranking did not offer the opportunity for the native conformation rank to be improved. Considering only those proteins which had the native conformation in the top 200, the median rank of the native conformation after re-ranking was 50, while before re-ranking it was 49. If the top ranked conformation of the predicted top 200 conformations is used (for both the original results and the re-ranked results), the median rank of the top conformation is 95 before re-scoring, and 96.5 after rescoring. Since re-ranking the proteins did not improve the performance, this step was not used.

5.2.2 Classification Results after Addition of Data

Next, additional correct conformations were selected to supplement the data. The supplemental data were put in two groups. First, 207 proteins were randomly selected

and included as samples with an RMS of 0 to give the classifier more correct instances. In a second test, 44 proteins which were homologous to one of the proteins in the dataset were included. Table 5-5 has the results of these tests, and includes the position of the native conformation with the addition of the first set of non-homologous data and then the addition of the second set of homologous data (in addition to the non-homologous data). The number of homologs for each protein that was added with the second data set is also listed.

Table 5-5:Scoring Results After Inclusion of Additional Data.

Protein	**Position of Native Conformation with Non-Homologous Data**	**Position of Native Conformation with Homologous and Non-Homologus Data**	**Number of Homologs Added**
1A0O	550	571	4
1ACB	137	141	
1AHW	46	43	3
1ATN	350	346	
1AVW	47	37	2
1AVZ	260	259	1
1BQL	293	285	
1BRC	329	344	
1BRS	292	292	
1BTH	46	50	1
1BVK	219	215	
1CGI	63	61	
1CHO	192	197	1
1CSE	54	40	
1DFJ	1	1	4
1DQJ	2	1	
1EFU	1	1	2
1EO8	208	227	
1FBI	94	88	

1FIN	7	9	
1FQ1	329	336	
1FSS	118	115	4
1GLA	601	573	4
1GOT	113	117	2
1IAI	7	11	
1IGC	219	223	2
1JHL	165	149	
1MAH	50	53	4
1MDA	421	373	
1MEL	1	1	
1MLC	178	180	
1NCA	3	3	
1NMB	144	153	
1PPE	7	7	
1QFU	64	67	
1SPB	2	2	
1STF	137	136	2
1TAB	146	139	5
1TGS	7	9	
1UDI	116	117	
1UGH	15	15	
1WEJ	53	56	
1WQ1	1	1	
2BTF	199	181	1
2JEL	17	16	
2KAI	254	249	
2PCC	518	522	
2PTC	170	178	
2SIC	156	167	2
2SNI	67	63	
2TEC	155	163	
2VIR	48	47	
3HHR	1	1	
4HTC	34	35	

The mean position of the native conformation after addition of the non-homologous data is 142.7 (versus 143.9 without this data). The further addition of the homologous data improved performance again to a mean rank of 142.0. For the proteins which had homologs that were added, mean performance went from 169.5 to 168.4; the proteins which did not have homologs added had a mean performance improvement of 0.6 (rank of the native conformation went from 130.4 to 129.8). Further study of the data set found that of the 131 chains that make up the data set, only 32 of those chains did not have homologs (BLAST E value < 0.0001) already existing in the data. The proteins that did not have homologs, but had homologs added through the second data increment included: 1A0O, 1AHW, 1AVW, 1AVZ, 1DFJ, 1EFU, 1GLA, 1GOT, 1IGC, 1STF, 1TAB, and 2BTF. Performance on these proteins went from a median conformation of 141.5 without the homologous data to 137.5 after inclusion of the homologous data. However, performance on the remainder of the data (without new homologous data) went from 80.5 to 77.5. So it appears that the improvement in performance is from the addition of data, not the specific addition of homologous data.

5.2.3 Classification Results after Splitting out Enzyme-Inhibitor and Antibody-Antigen Data

Results may potentially be improved by splitting out the data into subsets. This was checked by splitting out both the enzyme-inhibitor and the antibody-antigen classes from the main data set (Table 3-1 shows which proteins are included in each class).

Training and testing were performed on just these subsets of data using the features selected specifically for these data sets (described in Sections 5.1.4 and 5.1.5).

For the enzyme-inhibitor data subset, which included 22 proteins, the median rank of the native conformation is 81.5, and the median RMS of the top ranked conformation from the true native conformation is 11.6 angstroms. This is an improvement on the performance on these same proteins using all of the data for training, which results in a median rank for the top conformation of 106.0 and a median RMS of 12.4 angstroms.

Finally, the 16 proteins of the antibody-antigen data set had a median rank for the native conformation of 95.5, and an RMS of 15.2. This contrasts with a median rank of 61.5 and median RMS of 18.3 if all of the data are used for training. Interestingly, separating out the antibody-antigen proteins actually decreases performance on the native conformation rank, but improves performance for the RMS.

5.2.4 Final Scoring Results

Included in Table 5-6 is a list of the best performance on each protein in the dataset. For all proteins except the enzyme-inhibitor complexes, the data used for training was the original data set with the addition of both the non-homologous and the homologous data. The enzyme-inhibitor complexes (the proteins which are included in Table 5-6 in boldface) used a classifier trained using only the data from other enzyme-inhibitor complexes. The median position of the native conformation for all proteins is 96; however, the median position of the highest ranked near-native conformation (less than 5 angstroms RMS) is 10. The median number of correctly identified residues on the

larger and smaller sub-proteins is 0.595 and 0.534, respectively. The median RMS of the top prediction is 13.635 angstroms.

Table 5-6:Final Results.

Protein	Position of Native Conformation	RMS of Predicted Top Confirmation	Fraction of Correct Receptor Residues	Fraction of Correct Ligand Residues	Highest Ranked Confirmation with RMS $\leq$ 5 Å
1A0O	554	12.59	0.63	1.00	149
1ACB	68	11.03	0.73	0.82	2
1AHW	44	15.7	0.70	0.50	17
1ATN	380	18.08	0.56	0.14	217
1AVW	95	20.98	0.38	0.36	7
1AVZ	203	16.43	0.42	0.50	153
1BQL	306	23.19	0.40	0.29	210
1BRC	192	15.51	0.69	0.90	112
1BRS	172	12.22	0.40	0.31	5
1BTH	27	18.3	0.75	0.60	27
1BVK	265	19.77	0.40	0.31	11
1CGI	58	1.89	0.81	0.70	1
1CHO	52	16.33	0.75	0.46	18
1CSE	97	14.42	0.64	1.00	22
1DFJ	49	2.34	0.90	0.86	1
1DQJ	7	7.31	0.53	0.22	7
1EFU	1	0	1.00	1.00	1
1EO8	237	16.89	0.63	0.45	237
1FBI	71	16.82	0.40	0.22	3
1FIN	9	12.02	0.57	0.53	5
1FQ1	305	18.43	0.68	0.67	38
1FSS	177	9.76	0.57	0.75	25
1GLA	496	29.41	0.80	0.40	346
1GOT	100	13.68	0.47	0.31	88
1IAI	13	4.97	0.88	0.84	1
1IGC	251	23.64	0.19	0.25	69

1JHL	201	20.49	0.08	0.77	153
1MAH	122	12.7	0.45	0.47	4
1MDA	410	13.06	0.57	0.53	55
1MEL	2	7.93	0.17	0.10	2
1MLC	175	16.73	0.38	0.25	113
1NCA	2	22.26	0.47	0.47	2
1NMB	142	14.91	0.71	0.13	56
1PPE	40	0.74	0.91	0.92	1
1QFU	64	10.89	0.30	0.50	23
1SPB	18	9.92	0.68	0.31	18
1STF	105	1.12	1.00	0.77	1
1TAB	127	6.69	0.76	0.90	3
1TGS	10	16.38	0.33	0.60	9
1UDI	221	6.06	0.82	0.53	2
1UGH	43	7.77	0.60	0.64	3
1WEJ	111	14.43	0.64	0.45	111
1WQ1	1	0	1.00	1.00	1
2BTF	178	9.29	0.59	0.76	5
2JEL	11	23.21	0.07	0.18	3
2KAI	36	12.13	0.40	0.42	8
2PCC	487	13.8	0.89	0.90	208
2PTC	177	13.59	0.36	0.69	83
2SIC	29	22.07	0.29	0.17	4
2SNI	31	5.38	0.76	0.80	8
2TEC	170	15.64	0.35	0.58	33
2VIR	49	25.93	0.44	0.57	14
3HHR	1	0	1.00	1.00	1
4HTC	42	5.87	0.70	0.88	3

Figures 5-1 and 5-2 include plots of the predicted rank as a function of the RMS for each of the proteins. Ideally these scores would fall in a straight diagonal line, with increasing RMS values assigned an increasing score. Performance on the proteins with data along a straight line is better than performance on the proteins with non-linear plots.

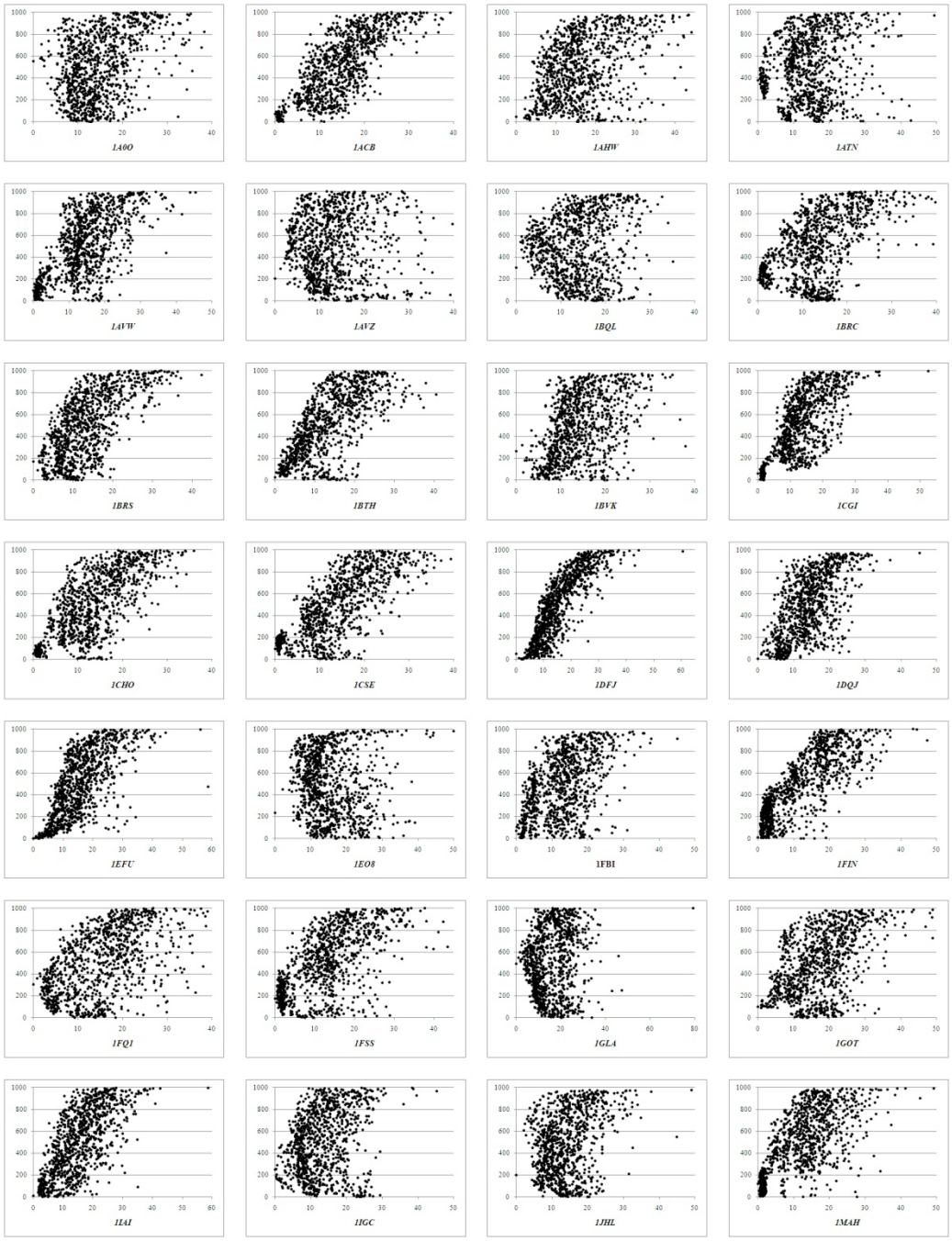

Figure 5-1: Sub plots for the first half of the proteins. Each plot is the predicted score as a function of the RMS, where each conformation is represented by a single point.

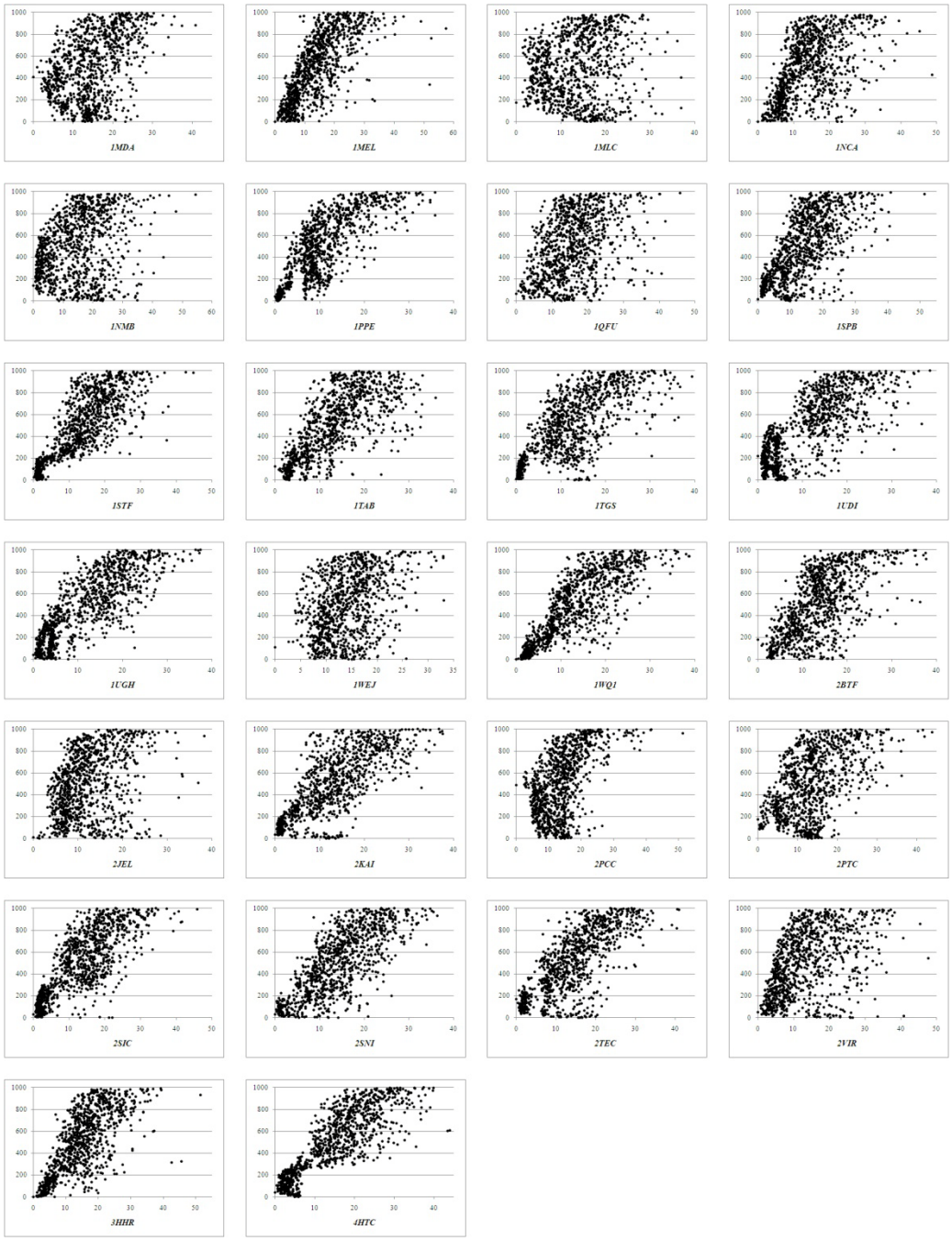

Figure 5-2: Sub plots for the remainder of the proteins. Each plot is the predicted score as a function of the RMS, where each conformation is represented by a single point.

5.2.5 Analysis of Scoring Results

There is a wide variability in the performance of the classifier on different proteins, from selection of the correct native structure for three of the proteins, to ranking the native confirmation at position 554 for protein 1A0O. The remainder of this section discusses factors that may have impacted the classifier performance for specific proteins.

Different protein classes, such as enzyme/inhibitor and antibody antigen complexes, may impact classifier performance. Enzymes and their inhibitors have co-evolved to form an interface with a high degree of surface complementarity, while the immune system produces many different antibodies in response to an antigen with varying degrees of success, so that some antibodies bind strongly to their antigen, while others do not. So a specific antibody/antigen complex does not necessarily have the best potential binding interface [42]. This may explain why separating the enzyme/inhibitor complexes improves their performance, while separation of antibody/antigen complexes does not.

Table 5-7 breaks the entire data set into subsets according to the type of protein: enzyme/inhibitor, antibody/antigen, other, or difficult to compare the performance on different classes. Despite the class labels, the classifier developed here demonstrated the best performance on the “difficult” class, while maintaining acceptable performance for both the antibody/antigen and enzyme/inhibitor classes. Interestingly, the class which provided the most challenge was the “other” data subset. Because these complexes are so different from the other proteins in the training data set, performance suffered.

Table 5-7:Results on when Data are Split by Type.

Data Subset	Median Position of Native Conformation	Median RMS of Predicted Top Confirmation	Median Fraction of Correct Receptor Residues	Median Fraction of Correct Ligand Residues	Median Highest Ranked Near Native (RMS $\leq$ 5 Å) Conformation
Antibody / Antigen	67.5	16.8	0.42	0.38	15.5
Enzyme / Inhibitor	91.5	11.6	0.66	0.70	4.5
Other	315.5	13.4	0.61	0.50	109.0
Difficult	18.0	12.9	0.72	0.63	16.0

Most proteins in the data set had homologs within the data set even before inclusion of the additional homologous data. Of the 54 proteins in the data set, only 8 (1A0O, 1ATN, 1BRS, 1GLA, 1GOT, 1MDA, 1STF, and 3HHR) did not have homologs for at least one molecule of the complex within the original data set. Another category of interest are those proteins which had a homolog for one molecule of the complex, but not the other. These complexes included: 1AHW, 1AVW, 1AVZ, 1DFJ, 1EFU, 1FIN, 1FQ1, 1IGC, 1PPE, 1SPB, 1TAB, 1WQ1, 2BTF, and 2JEL. Finally, there were some complexes that had one or more proteins for which the entire protein (both molecules) was homologous: 1BQL, 1BRC, 1BVK, 1CGI, 1CHO, 1CSE, 1DQJ, 1FBI, 1FSS, 1JHL, 1MAH, 1MEL, 1MLC, 1NCA, 1NMB, 1QFU, 1TGS, 1UDI, 1UGH, 2PCC, 2PTC, 2SNI, 2TEC, and 2VIR. Performance on each of these data subsets in included in Table 5-8.

Table 5-8:Results on Data Subsets with Varying Degrees of Homology.

Data Subset	Median Position of Native Conformation	Median RMS of Predicted Top Confirmation	Median Fraction of Correct Receptor Residues	Median Fraction of Correct Ligand Residues	Median Highest Ranked Near Native (RMS $\leq$ 5 Å) Conformation
No Homologs	276.0	12.8	0.63	0.38	71.5
Homolog for One Molecule	46.5	11.0	0.68	0.60	5.0
Homolog for Both Molecules	109.5	14.7	0.46	0.55	16.0

The inclusion of homologous data may have been advantageous for some proteins, but disruptive for others. Frequently, close homologs interact in similar orientations, but there are also examples of homologous proteins associating in different orientations [51]. For these data, the inclusion of homologous data was found to improve the overall performance of the classifier.

5.3 Comparison with Other Methods

As discussed in the background, comparison between specific methods is something of a challenge, especially when only the scoring methods are being compared. Results of alternative methods are included in Table 5-9, and it can be seen that the method developed and reported here performs admirably.

Table 5-9:Results of Other Scoring Methods.

Study	**Results**
Li 2007	Correct solutions ranked within top 2000 for 66 out of 83 complexes. Average rank of near-native solutions for 83 complexes was 1018 and the average rmsd 11.003 angstroms.
Mandell 2001	Geometry very close to crystallographic orientation within the best 266 minimum energies for all systems
Gray 2003	25 of 54 in the top 20 with 50% of contacts and 31 of 54 with 25% contacts
Comeau 2004	Successful if a certain number of the top clusters include at least one conformation with less than 10 angstroms RMSD from the native
Baster 1998	PRO_LEADS accurately predicted the binding mode of 86% of the complexes
Palma 2000	Near-native docked geometries were found with RMS $\leq$ 4 angstroms in 22 out of 25 complexes, and 14 of those were in the top 20
London 2007	FunHunt developed to distinguish between energy funnels – able to choose near-native funnel from the set of all 10 funnels with accuracy 72%
Qin 2007	Near-native poses were found for 23 of the 24 targets, but the poses with the lowest RMSD were ranked among the top 100 only for seven of the targets
Moont 1999	For all the systems, a correct docking was placed within the top 12% of the pair potential score ranked complexes

A direct comparison can be made between the method described here and the study by Gray *et al.* [48], as both methods used the same data set. Gray *et al*. were able to predict conformations for 32 of the 54 proteins, with 7 demonstrating at least 75% of the correct interface contacts; 23 predicting at least 50% of the contacts, and 28 predicting 25% or more. Comparatively, the method described here was able to rank the conformations for all 54 proteins; of these, 16 had 75% or more of the correct interface contacts, 30 proteins demonstrated at least 50% of the contacts, and 49 predicted 25% or more of the protein contacts.

CHAPTER 6: CONCLUSION

Two new procedures are presented in this work, both based on a topological descriptor. The method is applied first to identification of binding interface residues, and then to score different docking conformations. The two studies presented in this work have demonstrated two additional areas in which the topological descriptor can advance the current state of the field, and the success is viewed as promising.

In the first process, a new method to identify residues involved in protein-protein interactions is described. This method uses structural information to classify whether a specific residue is on the interaction interface or not. The random forest classifier was used to achieve a classification accuracy of 0.836 on a data set of 1476 non-homologous proteins, results which are comparable to other popular methods for protein-protein interface prediction. The classification algorithm could be further improved through: (1) inclusion of additional data, even if homologous, to give more redundant examples, and (2) identification and inclusion of additional features.

The topological descriptor was then used to develop a method of ranking docking conformations, and the method performed exceptionally well, placing the native structure within the top 100 for 29 of the 54 proteins, with an overall median position of 96. Additionally, 43 of the 54 proteins had a near-native structure (less than 5 angstroms from the native) in the top 100 positions. The median ratio of correctly

identified residue contacts is 0.57. Improvement for this work will result from inclusion of additional data, splitting the data into further appropriate subsets, and development of additional informative features.

APPENDIX

Appendix A: Leave-one-out analysis for proof-of-concept data set (described in Section 4.2.1)

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1A8M	B	0.671	0.591	0.448	0.296
1AE1	B	0.647	0.607	0.175	0.348
1AKJ	E	0.711	0.680	0.405	0.281
1AL2	3	0.651	0.757	0.617	0.450
1AOH	B	0.728	0.571	0.190	0.256
1AOI	A	0.673	0.934	0.671	0.757
1AOI	B	0.747	0.887	0.758	0.500
1AOI	C	0.600	0.965	0.556	0.759
1AOI	D	0.657	0.949	0.644	0.775
1AR8	1	0.643	0.908	0.548	0.552
1AVP	A	0.721	0.500	0.193	0.253
1AZD	A	0.747	0.610	0.362	0.224
1B35	B	0.400	0.592	0.179	0.646
1B35	C	0.603	0.873	0.473	0.550
1B48	A	0.751	0.739	0.258	0.247
1B67	A	0.603	0.786	0.512	0.525
1BH8	A	0.711	0.840	0.700	0.450
1BH8	B	0.472	0.929	0.366	0.738
1BQP	A	0.641	0.565	0.772	0.247
1BZX	I	0.707	0.500	0.353	0.239
1C14	A	0.750	0.526	0.303	0.211
1C2Y	A	0.684	0.630	0.475	0.294
1C72	A	0.714	0.708	0.236	0.285
1C8O	A	0.720	0.720	0.340	0.280
1C8O	B	0.938	0.938	1.000	0.000
1CDO	A	0.818	0.639	0.295	0.163
1CJQ	B	0.663	0.750	0.341	0.358
1CYD	A	0.715	0.555	0.805	0.130
1D3B	B	0.617	0.643	0.462	0.396
1D3B	C	0.704	6.745	0.667	0.250
1D5S	A	0.701	0.673	0.311	0.294
1D5S	B	0.780	0.970	0.800	1.000
1DCI	A	0.767	0.829	0.553	0.256
1DEE	D	0.659	0.557	0.410	0.302
1DIR	A	0.742	0.568	0.318	0.226
1DPS	A	0.730	0.574	0.673	0.173

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1EJB	A	0.696	0.604	0.516	0.261
1EOB	B	0.584	0.627	0.398	0.436
1F2D	D	0.777	0.593	0.372	0.188
1F2E	A	0.761	0.732	0.448	0.231
1F4M	B	0.679	0.793	0.657	0.444
1FNT	A	0.790	0.814	0.606	0.220
1FNT	C	0.817	0.855	0.663	0.200
1FZA	A	0.647	0.894	0.627	0.658
1FZA	B	0.645	0.766	0.388	0.394
1G1K	A	0.804	0.813	0.342	0.197
1G3I	G	0.809	0.805	0.569	0.189
1G8Q	A	0.778	0.895	0.486	0.254
1GCQ	A	0.643	0.467	0.778	0.154
1GCQ	C	0.652	0.481	0.565	0.238
1GEG	A	0.788	0.627	0.592	0.154
1GK4	A	0.532	0.953	0.539	0.972
1GL2	C	0.683	0.976	0.690	0.947
1GNW	A	0.681	0.333	0.098	0.286
1GO4	H	0.516	0.977	0.494	0.898
1GWC	C	0.734	0.905	0.250	0.284
1H59	B	0.711	0.600	0.400	0.257
1H5Q	A	0.746	0.701	0.557	0.235
1HEZ	A	0.603	0.560	0.303	0.378
1HEZ	E	0.721	0.760	0.633	0.306
1HFO	A	0.708	0.761	0.614	0.328
1HG3	A	0.817	0.442	0.528	0.094
1HRI	2	0.643	0.595	0.419	0.337
1HZD	B	0.687	0.536	0.947	0.052
1I8F	B	0.653	0.647	0.629	0.342
1IC2	B	0.545	1.000	0.541	0.897
1IJD	A	0.556	1.000	0.500	0.800
1IRJ	B	0.643	0.600	0.568	0.327
1IRU	F	0.744	0.729	0.548	0.250
1IRU	G	0.784	0.838	0.602	0.240
1IRU	H	0.752	0.710	0.579	0.229
1IRU	I	0.809	0.855	0.703	0.219
1IRU	J	0.760	0.513	0.765	0.094
1IRU	K	0.789	0.746	0.644	0.191
1IRU	L	0.776	0.657	0.667	0.164
1IRU	M	0.798	0.671	0.721	0.136
1IRU	N	0.811	0.822	0.682	0.194
1JFI	B	0.422	0.977	0.353	0.837
1JH5	A	0.667	0.735	0.391	0.355
1JK8	B	0.521	0.782	0.352	0.585

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1JLV	A	0.845	0.893	0.463	0.162
1JXZ	B	0.725	0.747	0.522	0.284
1KIL	A	0.769	0.918	0.804	0.688
1KIL	B	0.831	0.958	0.852	0.727
1KIL	D	0.682	0.952	0.678	0.792
1KKQ	A	0.625	0.889	0.139	0.394
1KQL	B	0.370	1.000	0.320	0.895
1LJR	A	0.738	0.833	0.298	0.276
1LLD	A	0.738	0.588	0.118	0.253
1MR8	A	0.567	0.889	0.302	0.514
1OTG	A	0.760	0.907	0.708	0.279
1PD2	1	0.578	0.857	0.228	0.343
1PMA	B	0.818	0.845	0.636	0.193
1PPF	I	0.411	0.364	0.133	0.578
1PSR	B	0.730	0.846	0.489	0.311
1QD9	A	0.806	0.667	0.703	0.129
1QGH	A	0.773	0.618	0.723	0.137
1RVF	1	0.652	0.905	0.556	0.535
1RVF	4	0.750	1.000	0.750	1.000
1SCJ	B	0.789	0.706	0.545	0.185
1TAF	A	0.647	0.933	0.560	0.579
1TAF	B	0.486	0.719	0.460	0.711
1TME	2	0.643	0.565	0.390	0.328
1YDV	A	0.793	0.690	0.323	0.194
2AAI	A	0.685	0.235	0.121	0.249
2AAI	B	0.817	0.793	0.354	0.180
2SIC	I	0.710	0.667	0.229	0.284
2SIV	A	0.889	0.939	0.939	0.667
2SIV	B	0.588	0.900	0.600	0.857
2SNI	E	0.891	0.190	0.235	0.051
2SNI	I	0.734	0.909	0.385	0.302

Appendix B: Leave-one-out Results for Final Data Set (described in Section 4.2.2)

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
12AS	A	0.899	0.525	0.600	0.049
1A0C	A	0.817	0.452	0.838	0.035
1A12	A	0.898	0.000	0.000	0.045
1A3A	A	0.869	0.333	0.462	0.055
1A4I	B	0.851	0.267	0.267	0.083
1A4M	A	0.971	0.000	0.000	0.006
1A6J	A	0.833	0.417	0.476	0.087
1AD3	A	0.843	0.338	0.533	0.057
1AHS	A	0.722	0.424	0.467	0.172
1AIH	A	0.694	0.686	0.369	0.304
1AJO	A	0.854	0.333	0.069	0.131
1AJY	A	0.493	0.950	0.352	0.686
1AOC	A	0.720	0.077	0.026	0.228
1ASH	A	0.849	0.000	0.000	0.139
1ASO	A	0.931	0.100	0.042	0.054
1ATZ	A	0.918	0.143	0.100	0.051
1AUU	A	0.673	0.846	0.407	0.381
1AUY	A	0.730	0.000	0.000	0.264
1AVO	A	0.600	0.821	0.653	0.810
1AVO	B	0.614	0.444	0.696	0.206
1AVQ	A	0.859	0.381	0.296	0.092
1B35	A	0.692	0.500	0.538	0.213
1B35	B	0.514	0.204	0.105	0.413
1B35	C	0.667	0.618	0.534	0.306
1B3U	A	0.838	0.364	0.054	0.164
1B5E	A	0.793	0.304	0.438	0.092
1B5Q	A	0.858	0.344	0.208	0.104
1B67	A	0.588	0.786	0.500	0.550
1B77	A	0.829	0.053	0.045	0.100
1B9L	A	0.790	0.750	0.702	0.187
1BEB	A	0.846	0.300	0.150	0.116
1BG8	A	0.684	0.706	0.387	0.322
1BGF	A	0.642	0.000	0.000	0.347
1BGV	A	0.902	0.333	0.111	0.075
1BH9	B	0.652	0.967	0.492	0.508
1BJA	A	0.611	0.533	0.211	0.375

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1BO4	A	0.752	0.656	0.477	0.219
1BOU	A	0.659	0.500	0.511	0.256
1BOU	B	0.829	0.294	0.870	0.013
1BRT	A	0.964	0.000	0.000	0.029
1BSL	A	0.879	0.412	0.424	0.066
1BYF	A	0.756	0.300	0.273	0.155
1BYK	A	0.910	0.480	0.545	0.043
1BYR	A	0.908	0.000	0.000	0.080
1C4Q	A	0.623	0.700	0.412	0.408
1C5E	A	0.684	0.536	0.469	0.254
1C7N	A	0.886	0.190	0.421	0.031
1C8N	A	0.778	0.324	0.367	0.123
1C8U	A	0.786	0.319	0.341	0.122
1C9K	B	0.800	0.381	0.258	0.145
1CBY	A	0.806	0.667	0.087	0.190
1CCW	A	0.825	0.143	0.333	0.052
1CCW	B	0.886	0.220	0.733	0.010
1CFZ	A	0.802	0.333	0.333	0.116
1CG2	A	0.907	0.474	0.529	0.046
1CHK	A	0.916	0.000	0.000	0.056
1CHM	A	0.908	0.672	0.707	0.050
1CI9	A	0.966	0.167	0.111	0.022
1CJX	A	0.892	0.310	0.333	0.056
1CKM	A	0.836	0.313	0.250	0.105
1CMC	A	0.750	0.862	0.532	0.293
1COL	A	0.898	0.000	0.000	0.043
1COZ	A	0.738	0.235	0.167	0.183
1CP2	A	0.907	0.400	0.381	0.052
1CQ3	A	0.884	0.421	0.348	0.073
1CQX	A	0.849	0.000	0.000	0.109
1CRU	B	0.920	0.048	0.059	0.038
1CSH	A	0.855	0.333	0.033	0.138
1CTF	A	0.824	0.500	0.083	0.167
1CTT	A	0.918	1.000	0.077	0.082
1D0C	A	0.834	0.491	0.397	0.114
1D0Q	A	0.725	0.357	0.208	0.216
1D1G	A	0.750	0.323	0.333	0.150

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1D2N	A	0.866	0.000	0.000	0.116
1D2O	A	0.888	0.400	0.211	0.085
1D2Z	B	0.753	0.500	0.378	0.189
1D3B	A	0.653	0.677	0.583	0.366
1D3B	B	0.531	0.643	0.391	0.528
1D9C	A	0.620	0.667	0.627	0.431
1DAA	A	0.892	0.436	0.680	0.034
1DAB	A	0.852	0.667	0.030	0.147
1DCE	B	0.830	0.238	0.652	0.030
1DCI	A	0.764	0.447	0.596	0.116
1DJ0	A	0.902	0.346	0.500	0.038
1DK0	A	0.850	0.200	0.455	0.041
1DL5	A	0.868	0.091	0.030	0.105
1DM9	A	0.558	0.267	0.103	0.393
1DMH	A	0.816	0.607	0.529	0.133
1DP4	A	0.913	0.118	0.083	0.054
1DPG	A	0.880	0.349	0.349	0.071
1DQE	A	0.854	0.200	0.143	0.094
1DQN	A	0.817	0.333	0.355	0.102
1DQZ	A	0.893	0.214	0.136	0.071
1DRW	A	0.801	0.667	0.036	0.197
1DZK	A	0.892	0.214	0.375	0.037
1E0B	A	0.443	0.692	0.231	0.625
1E19	A	0.827	0.255	0.387	0.071
1E6U	A	0.892	0.500	0.029	0.105
1EAJ	A	0.815	0.643	0.333	0.164
1EBF	A	0.891	0.429	0.343	0.070
1ECE	A	0.961	0.267	0.571	0.009
1ECM	A	0.670	0.921	0.565	0.509
1ECS	A	0.758	0.667	0.474	0.215
1EDZ	A	0.896	0.000	0.000	0.098
1EE8	A	0.906	0.000	0.000	0.069
1EER	A	0.783	0.515	0.459	0.150
1EER	B	0.737	0.286	0.128	0.214
1EEX	A	0.746	0.299	0.809	0.029
1EEX	B	0.809	0.346	0.346	0.112
1EEX	G	0.672	0.509	0.587	0.226

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1EFD	N	0.916	0.000	0.000	0.077
1EH9	A	0.961	0.500	0.048	0.046
1EI7	A	0.703	0.167	0.023	0.276
1EI9	A	0.943	0.000	0.000	0.051
1EJ2	A	0.862	0.000	0.000	0.117
1EJ6	D	0.854	0.150	0.474	0.028
1EJE	A	0.781	0.500	0.048	0.213
1EJF	A	0.655	0.364	0.114	0.313
1EKJ	A	0.714	0.597	0.548	0.231
1EL6	A	0.688	0.671	0.606	0.301
1ELK	A	0.856	0.500	0.045	0.139
1ELU	A	0.861	0.288	0.607	0.034
1EM8	A	0.844	0.154	0.143	0.090
1EM8	B	0.736	0.462	0.214	0.227
1EM9	A	0.762	0.667	0.108	0.234
1EPA	A	0.750	0.214	0.094	0.199
1ES9	A	0.858	0.000	0.000	0.133
1ETE	A	0.828	0.188	0.231	0.085
1EV0	A	0.517	0.875	0.457	0.735
1EX2	A	0.822	0.364	0.133	0.149
1EXT	A	0.675	0.565	0.236	0.307
1EYQ	A	0.807	0.118	0.071	0.133
1EYV	A	0.809	0.588	0.357	0.158
1EZ0	A	0.883	0.412	0.389	0.055
1EZG	A	0.695	0.400	0.174	0.264
1F06	A	0.828	0.438	0.429	0.103
1F0K	A	0.900	0.111	0.036	0.079
1F15	A	0.694	0.636	0.259	0.296
1F1M	A	0.735	0.500	0.488	0.183
1F2N	A	0.746	0.265	0.281	0.148
1F2T	B	0.685	0.477	0.488	0.222
1F2V	A	0.823	0.750	0.077	0.176
1F3U	A	0.644	0.717	0.585	0.415
1F46	A	0.705	0.357	0.135	0.256
1F7D	A	0.542	0.571	0.073	0.459
1F86	A	0.791	0.737	0.424	0.198
1F8M	A	0.796	0.503	0.864	0.043

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1F8Y	A	0.827	0.333	0.500	0.070
1FC3	A	0.706	0.476	0.294	0.245
1FCD	A	0.903	0.167	0.636	0.011
1FCD	C	0.839	0.161	0.714	0.014
1FE0	A	0.652	0.500	0.217	0.321
1FGJ	A	0.868	0.071	0.026	0.090
1FGU	A	0.764	0.111	0.020	0.211
1FIP	A	0.589	0.893	0.481	0.600
1FJR	A	0.814	0.538	0.194	0.166
1FLC	A	0.803	0.527	0.443	0.139
1FLC	B	0.512	0.442	0.551	0.408
1FLM	A	0.738	0.318	0.292	0.170
1FN9	A	0.893	0.333	0.300	0.062
1FP2	A	0.788	0.500	0.014	0.210
1FPO	A	0.661	0.429	0.107	0.318
1FPZ	A	0.858	0.000	0.000	0.137
1FSE	A	0.522	0.579	0.314	0.500
1FTR	A	0.733	0.291	0.581	0.086
1FUI	A	0.831	0.218	0.607	0.031
1FVK	A	0.846	0.063	0.067	0.081
1G0S	B	0.757	0.605	0.708	0.151
1G2C	B	0.750	0.964	0.750	0.750
1G31	A	0.617	0.548	0.386	0.355
1G3K	A	0.855	0.556	0.370	0.110
1G5B	A	0.918	0.429	0.375	0.049
1G5Q	A	0.845	0.500	0.259	0.125
1G5T	A	0.752	0.500	0.026	0.245
1G61	A	0.951	0.125	0.200	0.018
1G6G	A	0.827	0.111	0.067	0.119
1G8E	A	0.531	0.795	0.486	0.685
1G8K	A	0.887	0.125	0.600	0.010
1G8K	B	0.684	0.311	0.560	0.125
1G8Q	A	0.789	0.737	0.500	0.197
1GD8	A	0.610	0.571	0.273	0.381
1GL4	A	0.901	0.412	0.292	0.066
1GL4	B	0.708	0.722	0.382	0.296
1GME	A	0.647	0.730	0.386	0.381

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1GMU	A	0.681	0.389	0.175	0.275
1GO3	F	0.766	0.771	0.614	0.236
1GPE	A	0.947	0.130	0.600	0.005
1GQ6	B	0.864	0.238	0.526	0.035
1GSA	A	0.901	0.333	0.069	0.088
1GU2	A	0.750	0.667	0.229	0.241
1GU7	A	0.901	0.238	0.200	0.058
1GUQ	A	0.818	0.506	0.609	0.093
1GUT	A	0.791	0.897	0.778	0.357
1GVJ	A	0.660	0.560	0.275	0.319
1GVN	A	0.920	0.533	1.000	0.000
1GVN	B	0.868	0.100	0.036	0.103
1GWY	A	0.897	0.250	0.143	0.072
1GXC	A	0.784	0.000	0.000	0.188
1GXJ	A	0.783	0.370	0.357	0.134
1GXM	A	0.886	0.000	0.000	0.068
1GXY	A	0.937	0.000	0.000	0.041
1GY7	A	0.760	0.435	0.385	0.163
1GYG	A	0.905	0.375	0.091	0.083
1GYT	A	0.837	0.387	0.518	0.075
1H2I	A	0.613	0.545	0.732	0.289
1H3L	A	0.520	0.600	0.231	0.500
1H3O	A	0.729	0.897	0.722	0.526
1H4R	A	0.840	0.125	0.024	0.140
1H6D	A	0.723	0.292	0.559	0.096
1H7E	A	0.849	0.308	0.296	0.087
1H8U	A	0.661	0.333	0.026	0.330
1H97	A	0.952	0.200	0.250	0.021
1HBN	B	0.719	0.228	0.733	0.041
1HBN	C	0.789	0.516	0.873	0.045
1HCN	A	0.576	0.725	0.537	0.556
1HCN	B	0.500	0.676	0.368	0.589
1HF2	B	0.777	0.294	0.313	0.128
1HF8	A	0.932	0.333	0.059	0.062
1HI9	A	0.876	0.242	0.471	0.037
1HJR	A	0.734	0.370	0.286	0.191
1HKQ	A	0.632	0.278	0.132	0.308

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1HO1	C	0.880	0.273	0.125	0.091
1HQZ	1	0.835	0.333	0.150	0.131
1HST	A	0.500	0.857	0.143	0.537
1HTW	A	0.785	0.310	0.391	0.109
1HUL	A	0.694	0.833	0.615	0.417
1HUX	A	0.923	0.556	0.238	0.064
1HW1	A	0.819	0.667	0.392	0.158
1HW5	A	0.731	0.333	0.191	0.210
1HWX	A	0.824	0.423	0.550	0.075
1HXR	B	0.817	0.333	0.050	0.170
1HYN	P	0.829	0.431	0.512	0.087
1HYO	A	0.877	0.423	0.512	0.058
1I0R	A	0.801	0.558	0.649	0.110
1I4U	A	0.818	0.429	0.414	0.111
1I52	A	0.822	0.200	0.027	0.164
1I58	A	0.762	0.071	0.030	0.183
1I6A	A	0.863	0.000	0.000	0.129
1I6P	A	0.706	0.500	0.032	0.290
1IA9	B	0.771	0.588	0.411	0.188
1IBY	A	0.750	0.750	0.587	0.250
1IDP	A	0.653	0.316	0.324	0.229
1IG0	A	0.899	0.367	0.458	0.045
1IG3	A	0.846	0.370	0.313	0.097
1IGQ	A	0.481	0.778	0.368	0.667
1II7	A	0.862	0.188	0.083	0.104
1IIE	A	0.600	0.833	0.603	0.697
1IJY	A	0.754	0.364	0.148	0.207
1IK9	C	0.607	0.882	0.625	0.818
1ILK	A	0.437	0.833	0.056	0.579
1IN0	A	0.759	0.545	0.150	0.225
1INL	C	0.775	0.224	0.517	0.062
1IO0	A	0.849	0.500	0.040	0.146
1IQ4	A	0.732	0.308	0.093	0.235
1IQ8	A	0.913	0.333	0.556	0.031
1IR6	A	0.927	0.000	0.000	0.068
1ITH	A	0.851	0.364	0.222	0.108
1ITU	A	0.886	0.195	0.471	0.027

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1IU4	A	0.843	0.316	0.133	0.125
1IU8	A	0.893	0.500	0.364	0.074
1IUJ	A	0.706	0.679	0.475	0.284
1IV2	A	0.720	0.237	0.409	0.116
1IWM	A	0.887	0.000	0.000	0.082
1IYK	A	0.872	0.622	0.390	0.101
1IZC	A	0.776	0.500	0.045	0.218
1IZN	A	0.822	0.768	0.616	0.160
1IZN	B	0.585	0.474	0.421	0.354
1IZO	A	0.959	0.000	0.000	0.037
1J1J	A	0.774	0.438	0.311	0.168
1J1N	A	0.911	0.333	0.063	0.070
1J24	A	0.827	0.000	0.000	0.160
1J2G	A	0.765	0.523	0.727	0.106
1J2R	A	0.856	0.489	0.846	0.028
1J3W	A	0.866	0.692	0.643	0.093
1J5S	A	0.829	0.182	0.500	0.039
1J6R	A	0.782	0.308	0.105	0.185
1J9I	A	0.382	0.500	0.119	0.638
1JB3	A	0.772	0.500	0.034	0.224
1JCL	A	0.924	0.111	0.083	0.045
1JDW	A	0.939	0.000	0.000	0.056
1JEK	A	0.400	0.789	0.429	0.952
1JFL	A	0.838	0.269	0.280	0.089
1JFM	A	0.747	0.154	0.057	0.205
1JFR	A	0.977	0.333	0.200	0.016
1JFU	A	0.909	0.000	0.000	0.059
1JG5	A	0.711	0.767	0.575	0.321
1JH6	A	0.840	0.200	0.235	0.081
1JHF	A	0.792	0.375	0.162	0.171
1JHG	A	0.406	0.667	0.033	0.602
1JI1	A	0.959	0.000	0.000	0.035
1JIH	A	0.835	0.294	0.079	0.138
1JKE	A	0.793	0.612	0.732	0.115
1JKM	A	0.902	0.103	0.250	0.027
1JKX	A	0.818	0.000	0.000	0.123
1JLY	A	0.732	0.224	0.208	0.168

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1JMV	A	0.779	0.625	0.405	0.190
1JNR	B	0.624	0.637	0.716	0.397
1JO0	A	0.711	0.846	0.297	0.310
1JOC	A	0.528	0.694	0.347	0.540
1JOY	A	0.313	0.750	0.120	0.746
1JPY	A	0.669	0.730	0.667	0.397
1JQE	A	0.932	0.200	0.063	0.055
1JR2	A	0.812	0.273	0.068	0.165
1JR8	A	0.695	0.609	0.378	0.280
1JT6	A	0.731	0.059	0.100	0.118
1JU2	A	0.969	0.143	0.125	0.016
1JYO	A	0.669	0.429	0.686	0.149
1JYO	E	0.578	0.774	0.569	0.633
1K04	A	0.479	1.000	0.051	0.536
1K12	A	0.854	0.500	0.043	0.141
1K1E	A	0.785	0.314	0.842	0.024
1K2E	A	0.678	0.640	0.286	0.315
1K3R	A	0.821	0.313	0.286	0.109
1K3Y	A	0.887	0.536	0.556	0.062
1K4Z	A	0.739	0.667	0.136	0.257
1K8Q	A	0.926	0.000	0.000	0.054
1K9X	A	0.871	0.140	0.194	0.063
1KA8	A	0.630	0.467	0.194	0.341
1KAF	A	0.713	0.182	0.083	0.227
1KBP	A	0.887	0.200	0.077	0.088
1KDG	A	0.917	0.118	0.074	0.060
1KGN	A	0.838	0.537	0.558	0.095
1KHI	A	0.653	0.500	0.020	0.345
1KHV	B	0.883	0.333	0.085	0.101
1KLO	A	0.753	0.500	0.025	0.244
1KMT	A	0.783	0.636	0.212	0.205
1KNC	A	0.707	0.362	0.600	0.121
1KNQ	A	0.836	0.333	0.217	0.115
1KNY	A	0.818	0.536	0.600	0.102
1KO6	A	0.743	0.636	0.167	0.248
1KOL	A	0.909	0.133	0.080	0.060
1KQ1	A	0.700	0.710	0.710	0.310

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1KQF	B	0.785	0.549	0.643	0.121
1KQF	C	0.718	0.563	0.277	0.255
1KQP	A	0.838	0.472	0.610	0.073
1KTG	A	0.847	0.000	0.000	0.147
1KWG	A	0.921	0.300	0.075	0.087
1KXU	A	0.870	0.500	0.028	0.128
1KZQ	A	0.830	0.300	0.171	0.124
1L0W	A	0.802	0.414	0.387	0.125
1L3A	A	0.602	0.580	0.392	0.388
1L5A	A	0.892	0.571	0.085	0.103
1L5J	A	0.945	0.000	0.000	0.039
1L6W	A	0.745	0.384	0.718	0.075
1L7A	A	0.925	0.200	0.333	0.027
1L8D	A	0.495	0.429	0.120	0.494
1LB6	A	0.794	1.000	0.030	0.208
1LC5	A	0.927	0.333	0.040	0.068
1LDD	A	0.635	0.500	0.259	0.333
1LF6	A	0.967	0.000	0.000	0.019
1LGP	A	0.434	1.000	0.045	0.582
1LI4	A	0.874	0.500	0.111	0.115
1LJ2	A	0.594	0.741	0.580	0.558
1LJ9	A	0.688	0.622	0.500	0.283
1LKT	A	0.683	0.659	0.587	0.302
1LLF	A	0.949	0.000	0.000	0.023
1LO7	A	0.707	0.333	0.025	0.285
1LQ9	A	0.750	0.781	0.543	0.263
1LR5	A	0.769	0.481	0.361	0.173
1LTL	A	0.799	0.769	0.182	0.199
1LVF	A	0.651	0.571	0.205	0.337
1LVM	B	0.822	0.200	0.286	0.079
1LZL	A	0.950	0.500	0.063	0.048
1M0D	A	0.620	0.621	0.571	0.380
1M1C	A	0.896	0.167	0.098	0.090
1M1L	A	0.852	0.292	0.280	0.085
1M1N	A	0.816	0.429	0.639	0.064
1M1N	B	0.791	0.418	0.676	0.067
1M2D	A	0.842	0.667	0.400	0.135

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1M3Y	A	0.855	0.417	0.086	0.132
1M4I	A	0.746	0.304	0.189	0.190
1M4R	A	0.768	0.190	0.200	0.132
1M55	A	0.829	0.188	0.130	0.113
1M70	A	0.905	0.000	0.000	0.080
1M98	A	0.883	0.261	0.231	0.068
1MBM	A	0.864	0.190	0.286	0.056
1MBY	A	0.493	0.840	0.382	0.680
1MK4	A	0.782	0.381	0.276	0.156
1MKA	A	0.871	0.694	0.694	0.081
1MKK	A	0.495	0.667	0.321	0.576
1MN8	A	0.632	0.522	0.333	0.333
1MO9	A	0.835	0.333	0.422	0.068
1MP9	A	0.808	0.500	0.324	0.148
1MPG	A	0.911	0.000	0.000	0.069
1MPY	A	0.837	0.308	0.533	0.055
1MSC	A	0.473	1.000	0.029	0.535
1MT5	A	0.916	0.162	0.400	0.023
1MTY	B	0.729	0.405	0.827	0.056
1MTY	D	0.752	0.276	0.654	0.057
1MTY	G	0.815	0.579	0.846	0.057
1MV8	A	0.814	0.534	0.708	0.082
1MWW	A	0.717	0.596	0.705	0.191
1MXR	A	0.844	0.390	0.575	0.061
1MY7	A	0.598	0.615	0.174	0.404
1N0E	A	0.688	0.592	0.547	0.261
1N2Z	A	0.833	0.250	0.121	0.127
1N62	A	0.702	0.306	0.792	0.051
1N62	B	0.866	0.371	0.433	0.042
1N62	F	0.878	0.378	0.538	0.048
1N69	A	0.603	0.864	0.404	0.500
1N71	A	0.722	0.300	0.353	0.157
1N81	A	0.769	0.500	0.023	0.228
1N8V	A	0.683	0.308	0.148	0.261
1N97	A	0.917	0.000	0.000	0.078
1NBA	A	0.711	0.429	0.712	0.110
1NBC	A	0.897	0.375	0.214	0.075

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1NC5	A	0.956	0.000	0.000	0.039
1NCN	A	0.661	0.300	0.091	0.303
1ND4	A	0.925	0.167	0.067	0.056
1ND6	A	0.851	0.356	0.421	0.074
1NEI	A	0.433	0.870	0.392	0.838
1NF2	A	0.895	0.077	0.059	0.063
1NF9	A	0.903	1.000	0.091	0.098
1NG7	A	0.283	0.909	0.192	0.857
1NJF	A	0.820	0.414	0.316	0.124
1NKS	A	0.768	0.143	0.250	0.094
1NLN	A	0.837	0.130	0.188	0.072
1NLT	A	0.719	0.000	0.000	0.278
1NLX	A	0.596	0.423	0.289	0.346
1NNW	A	0.888	0.056	0.083	0.047
1NO4	A	0.610	0.914	0.542	0.643
1NO5	A	0.620	0.125	0.031	0.337
1NOX	A	0.705	0.400	0.034	0.287
1NP6	B	0.775	0.630	0.580	0.171
1NQJ	B	0.737	0.615	0.242	0.248
1NRZ	A	0.779	0.200	0.067	0.183
1NTH	A	0.932	0.200	0.037	0.060
1NTV	A	0.836	1.000	0.038	0.166
1NUY	A	0.881	0.600	0.075	0.115
1NVM	A	0.824	0.238	0.556	0.043
1NVM	B	0.875	0.516	0.400	0.085
1NYC	A	0.636	0.909	0.204	0.394
1O0W	A	0.797	0.238	0.135	0.148
1O26	A	0.803	0.591	0.709	0.105
1O5L	A	0.636	0.333	0.022	0.357
1O7D	B	0.907	0.957	0.944	0.800
1O7I	A	0.765	0.375	0.120	0.206
1O7Q	A	0.909	0.105	0.182	0.034
1O91	A	0.809	0.590	0.719	0.098
1O9I	A	0.654	0.503	0.892	0.099
1O9Y	A	0.690	0.913	0.700	0.720
1OA8	A	0.836	0.594	0.704	0.083
1OCY	A	0.485	0.583	0.067	0.522

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1OGD	A	0.748	0.400	0.200	0.207
1OH0	A	0.744	0.407	0.407	0.163
1OHF	A	0.768	0.308	0.343	0.128
1OHV	A	0.859	0.434	0.827	0.027
1OJL	B	0.793	0.485	0.314	0.161
1OK7	B	0.792	0.400	0.141	0.179
1OMO	A	0.878	0.286	0.571	0.032
1ON2	A	0.644	0.407	0.256	0.296
1ON3	A	0.760	0.356	0.578	0.081
1ONR	A	0.949	0.000	0.000	0.048
1OOE	A	0.787	0.147	0.192	0.104
1OOH	A	0.872	0.308	0.364	0.063
1OPO	A	0.861	0.400	0.242	0.101
1OPO	C	0.847	0.286	0.188	0.105
1OQJ	A	0.600	0.643	0.225	0.408
1OR4	A	0.799	0.324	0.500	0.081
1OR7	C	0.773	0.809	0.864	0.316
1ORJ	A	0.778	0.417	0.417	0.137
1ORR	A	0.870	0.209	0.474	0.034
1OSD	A	0.653	0.000	0.000	0.338
1OTG	A	0.720	0.596	0.739	0.176
1OTK	A	0.881	0.308	0.167	0.087
1OTV	A	0.811	0.400	0.400	0.112
1OU8	A	0.679	0.500	0.265	0.284
1OV9	A	0.578	0.917	0.564	0.810
1OYJ	C	0.863	0.556	0.441	0.095
1P0Y	B	0.800	0.444	0.241	0.161
1P1J	A	0.825	0.434	0.569	0.069
1P1M	A	0.918	0.250	0.032	0.075
1P35	A	0.863	0.385	0.132	0.115
1P6O	A	0.808	0.258	0.533	0.056
1P94	A	0.408	0.800	0.381	0.848
1P9E	A	0.881	0.581	0.595	0.068
1P9Y	A	0.590	0.500	0.021	0.409
1PB6	A	0.773	0.436	0.425	0.145
1PBE	A	0.921	0.750	0.091	0.078
1PBW	A	0.837	0.091	0.048	0.116

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1PC6	A	0.546	0.686	0.422	0.533
1PCF	A	0.470	0.724	0.438	0.730
1PD3	A	0.611	0.864	0.514	0.563
1PEA	A	0.957	0.000	0.000	0.041
1PF5	A	0.731	1.000	0.054	0.273
1PFO	A	0.877	0.000	0.000	0.112
1PGW	1	0.730	0.513	0.392	0.212
1PGW	2	0.812	0.239	0.262	0.102
1PIX	A	0.816	0.371	0.523	0.056
1PKH	A	0.742	0.688	0.208	0.253
1POI	B	0.831	0.309	0.739	0.029
1PPR	M	0.862	0.176	0.094	0.098
1PUC	A	0.426	1.000	0.033	0.586
1PXZ	A	0.962	0.000	0.000	0.023
1PYA	A	0.716	0.746	0.914	0.500
1Q08	A	0.649	0.780	0.639	0.500
1Q0Q	A	0.905	0.212	0.368	0.033
1Q2H	A	0.619	1.000	0.579	0.800
1Q4U	A	0.807	0.588	0.606	0.123
1Q5Y	A	0.738	0.625	0.667	0.192
1Q6O	A	0.873	0.533	0.552	0.071
1Q7F	A	0.882	0.000	0.000	0.061
1Q7L	A	0.750	0.582	0.754	0.133
1Q7L	B	0.807	0.820	0.893	0.222
1Q88	B	0.763	0.125	0.024	0.211
1QBE	B	0.462	0.722	0.165	0.579
1QC7	A	0.604	0.091	0.032	0.333
1QD6	C	0.767	0.234	0.355	0.104
1QFT	A	0.846	0.500	0.148	0.138
1QGT	A	0.585	0.486	0.293	0.383
1QHD	A	0.798	0.333	0.053	0.187
1QHX	A	0.837	0.000	0.000	0.153
1QKS	A	0.912	0.133	0.182	0.044
1QL0	A	0.900	0.063	0.100	0.040
1QLM	A	0.946	0.000	0.000	0.048
1QLW	A	0.833	0.414	0.558	0.073
1QMG	A	0.885	0.178	0.400	0.031

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1QMH	A	0.919	0.267	0.200	0.050
1QMY	A	0.846	0.091	0.067	0.097
1QQP	3	0.686	0.615	0.308	0.298
1QQR	A	0.688	0.690	0.370	0.312
1QRE	A	0.819	0.500	0.053	0.175
1QS1	A	0.920	0.375	0.100	0.069
1QSD	A	0.676	0.588	0.278	0.306
1QSM	A	0.793	0.608	0.738	0.111
1QU7	A	0.612	0.474	0.440	0.315
1QW9	A	0.928	0.300	0.214	0.053
1QWD	A	0.856	0.250	0.045	0.129
1QWG	A	0.936	0.500	0.125	0.057
1QWJ	A	0.811	0.553	0.542	0.122
1QWT	A	0.820	0.615	0.174	0.168
1QXN	A	0.693	0.606	0.408	0.279
1QYN	A	0.649	0.344	0.297	0.255
1QYR	A	0.861	0.000	0.000	0.114
1QZ9	A	0.906	1.000	0.050	0.095
1R0V	A	0.758	0.294	0.395	0.119
1R1T	A	0.582	0.757	0.467	0.525
1R30	A	0.913	0.227	0.333	0.034
1R31	A	0.755	0.374	0.493	0.123
1R44	A	0.901	0.000	0.000	0.057
1R45	A	0.851	0.111	0.125	0.077
1R46	A	0.895	0.231	0.222	0.058
1R6R	A	0.800	0.958	0.605	0.268
1R6R	B	0.675	1.000	0.480	0.464
1R77	A	0.687	0.800	0.216	0.326
1R7J	A	0.567	0.500	0.026	0.432
1R89	A	0.831	0.500	0.027	0.166
1R9D	A	0.976	0.632	0.706	0.012
1R9D	B	0.976	0.706	0.706	0.012
1RA0	A	0.849	0.667	0.061	0.149
1REG	X	0.697	0.182	0.067	0.252
1RFY	A	0.528	0.462	0.146	0.461
1RGX	A	0.573	0.735	0.463	0.527
1RH5	B	0.536	0.929	0.520	0.857

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1RH5	C	0.594	0.900	0.621	0.917
1RJD	A	0.875	0.161	0.250	0.051
1RJJ	A	0.649	0.679	0.388	0.361
1RKI	A	0.535	0.000	0.000	0.419
1RKQ	A	0.919	0.000	0.000	0.064
1RKU	A	0.810	0.200	0.208	0.106
1RP0	A	0.734	0.552	0.208	0.245
1RSO	A	0.650	0.806	0.674	0.583
1RTQ	A	0.969	0.000	0.000	0.024
1RVE	A	0.795	0.476	0.204	0.175
1RW6	A	0.708	0.250	0.019	0.282
1RWZ	A	0.889	0.000	0.000	0.107
1RY9	A	0.707	0.536	0.366	0.248
1S0P	A	0.864	0.250	0.100	0.107
1S1D	A	0.915	0.067	0.071	0.043
1S3E	A	0.872	0.227	0.435	0.033
1S3M	A	0.842	0.182	0.105	0.110
1S3Z	A	0.753	0.676	0.510	0.220
1S4C	B	0.781	0.348	0.296	0.144
1S5U	A	0.744	0.657	0.523	0.223
1S7I	A	0.581	0.500	0.019	0.418
1S7M	A	0.658	0.663	0.711	0.348
1S98	A	0.742	0.870	0.476	0.297
1S9R	A	0.917	0.240	0.286	0.039
1SAC	A	0.853	0.321	0.450	0.063
1SC3	A	0.763	0.500	0.659	0.118
1SC3	B	0.693	0.717	0.760	0.343
1SEF	A	0.808	0.500	0.021	0.190
1SEI	A	0.769	0.000	0.000	0.200
1SFK	A	0.603	0.719	0.535	0.488
1SG4	C	0.876	0.219	0.538	0.028
1SGM	A	0.815	0.419	0.448	0.105
1SH0	A	0.882	0.000	0.000	0.093
1SHS	A	0.635	0.500	0.714	0.218
1SJW	A	0.824	0.500	0.120	0.162
1SMO	A	0.611	0.769	0.196	0.410
1SQU	A	0.849	0.333	0.533	0.055

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1SSE	B	0.349	0.500	0.071	0.667
1SSQ	A	0.842	0.571	0.200	0.141
1STM	A	0.645	0.174	0.114	0.263
1STZ	A	0.814	0.419	0.340	0.125
1SU1	A	0.793	0.324	0.480	0.088
1SU8	A	0.946	0.000	0.000	0.053
1SUM	B	0.871	0.000	0.000	0.109
1SUR	A	0.860	0.500	0.033	0.136
1SVB	A	0.737	0.125	0.010	0.251
1SVM	A	0.787	0.312	0.500	0.084
1SVP	A	0.750	0.125	0.029	0.217
1SW5	A	0.893	0.111	0.133	0.052
1SWV	A	0.903	0.077	0.071	0.053
1SZ9	A	0.797	0.150	0.200	0.098
1SZH	A	0.728	0.167	0.063	0.222
1T06	A	0.877	0.476	0.357	0.084
1T0B	A	0.817	0.433	0.722	0.056
1T0F	A	0.765	0.111	0.185	0.100
1T0I	A	0.832	0.417	0.370	0.106
1T0T	V	0.770	0.368	0.658	0.074
1T15	A	0.853	0.000	0.000	0.139
1T16	A	0.913	0.100	0.034	0.067
1T1D	A	0.720	0.700	0.219	0.278
1T1V	A	0.677	0.500	0.200	0.296
1T2B	A	0.899	0.118	0.074	0.066
1T2W	A	0.772	0.071	0.048	0.153
1T33	A	0.814	0.444	0.556	0.091
1T4B	A	0.801	0.414	0.475	0.108
1T4O	A	0.506	0.875	0.152	0.534
1T56	A	0.860	0.250	0.040	0.127
1T6S	A	0.679	0.793	0.333	0.346
1T71	A	0.890	0.500	0.032	0.108
1T77	A	0.889	0.333	0.045	0.103
1T7R	A	0.876	0.000	0.000	0.120
1T92	A	0.685	0.654	0.405	0.305
1TAF	A	0.603	0.867	0.531	0.605
1TAF	B	0.514	0.750	0.480	0.684

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1TE5	A	0.846	0.280	0.250	0.092
1TH0	A	0.912	0.250	0.056	0.077
1TH7	A	0.564	0.410	0.593	0.282
1TH8	B	0.817	0.313	0.333	0.101
1THF	D	0.933	0.500	0.059	0.064
1TJL	A	0.524	0.414	0.188	0.448
1TJV	A	0.768	0.423	0.663	0.091
1TKV	A	0.652	0.563	0.273	0.329
1TO6	A	0.881	0.320	0.229	0.078
1TOA	A	0.845	0.250	0.194	0.099
1TR0	A	0.613	0.473	0.684	0.235
1TTW	A	0.669	0.846	0.229	0.352
1TU1	A	0.755	0.667	0.511	0.215
1TUW	A	0.604	1.000	0.045	0.404
1TVF	A	0.894	0.000	0.000	0.093
1TVX	A	0.578	0.655	0.528	0.486
1TX9	A	0.759	0.458	0.344	0.179
1TXG	A	0.893	0.286	0.476	0.037
1TY9	A	0.797	0.564	0.646	0.116
1TZJ	C	0.837	0.160	0.381	0.045
1U07	A	0.622	0.727	0.364	0.412
1U19	A	0.851	0.154	0.047	0.122
1U1I	A	0.865	0.348	0.697	0.031
1U1S	A	0.712	0.867	0.634	0.417
1U20	A	0.750	0.355	0.275	0.176
1U2W	B	0.654	0.765	0.609	0.446
1U55	A	0.830	0.500	0.219	0.144
1U6M	A	0.820	0.250	0.208	0.112
1U6Z	A	0.861	0.444	0.129	0.129
1U7P	A	0.805	0.227	0.250	0.106
1U8V	A	0.776	0.320	0.635	0.057
1U9L	A	0.544	0.200	0.036	0.429
1UB9	A	0.590	1.000	0.047	0.418
1UC2	A	0.933	0.167	0.167	0.036
1UCR	A	0.419	0.471	0.190	0.596
1UF2	E	0.736	0.300	0.764	0.046
1UFH	A	0.748	0.125	0.030	0.218

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1UFI	A	0.563	0.960	0.545	0.870
1UJ2	A	0.840	0.111	0.100	0.092
1USR	A	0.866	0.108	0.133	0.065
1UTC	A	0.855	0.176	0.075	0.110
1UTY	A	0.654	0.550	0.386	0.310
1UUN	A	0.658	0.550	0.328	0.313
1UW4	A	0.670	0.583	0.412	0.299
1UW4	B	0.794	0.359	0.350	0.124
1UWC	A	0.969	0.000	0.000	0.023
1UWK	A	0.906	0.333	0.500	0.027
1UXZ	A	0.802	0.143	0.048	0.161
1UYP	A	0.889	0.143	0.091	0.073
1UZ3	A	0.588	0.773	0.315	0.463
1V37	A	0.813	0.000	0.000	0.120
1V3E	A	0.893	0.250	0.267	0.055
1V4P	A	0.821	0.273	0.136	0.136
1V70	A	0.686	0.500	0.030	0.311
1V74	A	0.720	0.333	0.250	0.202
1V7L	A	0.784	0.000	0.000	0.170
1V7Z	A	0.728	0.195	0.652	0.044
1V8H	A	0.764	0.571	0.296	0.207
1V8Q	A	0.515	0.688	0.289	0.540
1V96	B	0.799	0.238	0.278	0.106
1VBK	A	0.801	0.200	0.058	0.168
1VC1	A	0.818	0.583	0.318	0.153
1VC4	A	0.957	0.000	0.000	0.028
1VDM	G	0.684	0.432	0.452	0.213
1VDR	A	0.828	0.182	0.100	0.123
1VE9	A	0.912	0.286	0.444	0.032
1VF7	A	0.650	0.500	0.217	0.323
1VH5	A	0.774	0.556	0.441	0.173
1VHM	A	0.818	0.333	0.381	0.096
1VHW	A	0.797	0.367	0.688	0.056
1VJ0	A	0.858	0.290	0.692	0.026
1VJ2	A	0.675	0.600	0.477	0.291
1VKC	A	0.651	0.417	0.100	0.328
1VKI	A	0.873	0.667	0.444	0.102

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1VLG	C	0.829	0.523	0.767	0.058
1VLG	H	0.823	0.690	0.500	0.148
1VMO	A	0.804	0.125	0.038	0.161
1VPS	A	0.642	0.374	0.430	0.232
1VSG	A	0.751	0.317	0.433	0.121
1VYB	A	0.928	0.500	0.235	0.057
1VZ0	A	0.661	0.551	0.606	0.259
1VZI	A	0.728	0.500	0.441	0.200
1VZY	A	0.814	0.250	0.143	0.135
1W18	A	0.955	0.200	0.056	0.039
1W1H	A	0.748	0.500	0.135	0.234
1W23	A	0.894	0.289	0.684	0.019
1W33	A	0.751	0.615	0.444	0.211
1W61	A	0.887	0.300	0.321	0.059
1W6S	A	0.931	0.661	0.911	0.011
1W6S	B	0.639	1.000	0.527	0.605
1W6S	C	0.919	0.649	0.740	0.034
1W6S	D	0.639	0.964	0.519	0.568
1W79	A	0.918	0.500	0.028	0.080
1W8S	A	0.752	0.063	0.667	0.011
1W91	A	0.838	0.397	0.500	0.067
1W9C	A	0.829	0.235	0.087	0.138
1W9Z	A	0.739	0.423	0.536	0.140
1WA8	A	0.455	0.639	0.359	0.651
1WA8	B	0.589	0.800	0.467	0.533
1WAP	A	0.647	0.650	0.722	0.357
1WCV	1	0.909	0.000	0.000	0.083
1WDJ	A	0.715	0.085	0.286	0.072
1WHI	A	0.762	1.000	0.065	0.242
1WKO	A	0.842	0.000	0.000	0.120
1WKQ	B	0.600	0.413	0.688	0.200
1WLE	B	0.827	0.283	0.278	0.102
1WLG	A	0.826	0.143	0.022	0.157
1WLZ	A	0.612	0.400	0.129	0.360
1WMH	A	0.627	0.385	0.179	0.329
1WMH	B	0.622	0.600	0.182	0.375
1WMI	A	0.614	0.489	0.697	0.244

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1WMW	A	0.866	0.200	0.529	0.028
1WMX	A	0.798	0.500	0.143	0.184
1WO8	A	0.738	0.343	0.545	0.110
1WOL	A	0.779	0.000	0.000	0.208
1WOM	A	0.922	0.167	0.154	0.043
1WOQ	A	0.877	0.240	0.333	0.053
1WP1	A	0.783	1.000	0.048	0.229
1WPV	A	0.721	0.429	0.154	0.248
1WR8	A	0.883	0.176	0.188	0.061
1WS8	A	0.721	0.071	0.059	0.178
1WSP	A	0.410	0.714	0.096	0.618
1WTJ	A	0.831	0.414	0.522	0.080
1WU9	A	0.661	0.967	0.604	0.655
1WUI	S	0.760	0.260	0.741	0.037
1WUR	A	0.692	0.500	0.421	0.241
1WW7	A	0.794	0.286	0.250	0.129
1WWH	A	0.568	0.357	0.161	0.388
1WWJ	A	0.586	0.471	0.410	0.354
1WWL	A	0.896	0.462	0.194	0.085
1WWZ	A	0.873	0.444	0.211	0.101
1WXC	A	0.894	0.040	0.167	0.020
1WY5	A	0.791	0.297	0.220	0.142
1WYU	B	0.734	0.377	0.836	0.044
1WZ3	A	0.655	0.745	0.732	0.517
1WZC	B	0.918	0.000	0.000	0.067
1WZD	A	0.900	0.286	0.111	0.079
1X1N	A	0.868	0.500	0.016	0.140
1X2I	A	0.544	0.500	0.452	0.425
1X6M	A	0.806	0.412	0.200	0.156
1X89	A	0.810	0.125	0.037	0.157
1X8D	A	0.788	0.722	0.684	0.176
1X8L	A	0.953	0.000	0.000	0.044
1X9V	A	0.244	1.000	0.190	0.919
1X9X	A	0.548	0.636	0.226	0.471
1XCR	A	0.930	0.000	0.000	0.027
1XEQ	A	0.551	0.735	0.446	0.564
1XEY	A	0.808	0.545	0.516	0.129

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1XFK	A	0.935	0.714	0.208	0.060
1XG7	B	0.874	0.053	0.077	0.055
1XHX	A	0.890	0.375	0.083	0.077
1XM5	A	0.763	0.682	0.341	0.223
1XOD	B	0.686	0.500	0.081	0.304
1XPJ	A	0.774	0.596	0.757	0.117
1XRK	A	0.717	0.656	0.477	0.261
1XRS	A	0.946	0.500	0.280	0.043
1XT5	A	0.748	0.000	0.000	0.241
1XTE	A	0.629	0.400	0.048	0.360
1XTT	A	0.800	0.446	0.676	0.075
1XU1	A	0.693	0.354	0.607	0.124
1XVA	A	0.795	0.512	0.361	0.157
1XWR	A	0.646	0.889	0.490	0.481
1XX1	A	0.909	0.238	0.333	0.038
1XZO	A	0.872	0.083	0.083	0.069
1Y0H	A	0.733	0.739	0.447	0.269
1Y1L	A	0.806	0.273	0.429	0.078
1Y1P	A	0.921	0.133	0.125	0.043
1Y23	A	0.633	0.571	0.600	0.316
1Y37	A	0.918	0.167	0.500	0.015
1Y56	B	0.904	0.207	0.316	0.038
1Y60	B	0.631	0.301	0.667	0.116
1Y6V	A	0.849	0.229	0.864	0.008
1Y96	A	0.581	0.579	0.524	0.417
1Y96	B	0.659	1.000	0.453	0.475
1Y9W	A	0.679	0.741	0.563	0.360
1YAC	A	0.765	0.130	0.097	0.155
1YAV	A	0.746	0.704	0.422	0.243
1YB0	A	0.898	0.462	0.400	0.063
1YBI	A	0.820	0.500	0.118	0.165
1YCD	A	0.911	0.250	0.053	0.077
1YCO	A	0.888	0.520	0.406	0.076
1YEW	A	0.767	0.437	0.592	0.111
1YEW	B	0.647	0.522	0.778	0.192
1YF2	A	0.776	0.438	0.075	0.210
1YFU	A	0.776	0.500	0.026	0.221

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1YGT	A	0.596	1.000	0.045	0.412
1YK3	A	0.798	0.240	0.222	0.121
1YKI	A	0.796	0.603	0.667	0.124
1YKU	A	0.841	0.500	0.429	0.105
1YL7	A	0.763	0.585	0.551	0.172
1YLK	A	0.736	0.636	0.766	0.174
1YN9	B	0.905	0.333	0.143	0.074
1YNB	A	0.677	0.426	0.500	0.204
1YOC	A	0.841	0.593	0.571	0.102
1YPY	A	0.857	0.100	0.056	0.099
1YQF	A	0.725	0.560	0.500	0.212
1YQZ	A	0.876	0.380	0.452	0.059
1YRL	A	0.901	0.222	0.174	0.045
1YT5	A	0.832	0.429	0.308	0.118
1YTL	B	0.842	0.071	0.077	0.083
1YUM	A	0.915	0.583	0.350	0.065
1YXY	A	0.887	0.143	0.048	0.090
1YY7	A	0.835	0.417	0.333	0.110
1YZY	A	0.908	0.125	0.031	0.077
1Z0S	A	0.843	0.422	0.594	0.064
1Z2L	A	0.866	0.475	0.358	0.092
1Z2W	A	0.841	0.400	0.417	0.089
1Z3E	A	0.678	0.222	0.143	0.240
1Z3E	B	0.537	0.231	0.125	0.389
1Z4E	A	0.713	0.460	0.590	0.160
1Z56	A	0.545	0.939	0.484	0.750
1Z6N	A	0.831	0.000	0.000	0.159
1Z9H	A	0.836	0.323	0.294	0.099
1Z9M	A	0.654	0.133	0.080	0.258
1ZA7	A	0.715	0.476	0.238	0.246
1ZB1	A	0.907	0.000	0.000	0.075
1ZCD	A	0.830	0.000	0.000	0.106
1ZCZ	A	0.765	0.400	0.458	0.132
1ZH1	A	0.632	0.412	0.123	0.342
1ZHS	A	0.735	0.640	0.727	0.190
1ZHV	A	0.657	0.000	0.000	0.328
1ZHX	A	0.894	0.500	0.022	0.104

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
1ZJ9	A	0.908	0.182	0.059	0.075
1ZKE	A	0.481	0.586	0.362	0.577
1ZKK	A	0.781	0.526	0.278	0.184
1ZL8	B	0.463	0.778	0.359	0.694
1ZLP	A	0.803	0.450	0.167	0.170
1ZMT	A	0.766	0.372	0.744	0.057
1ZOS	A	0.857	0.415	0.654	0.048
1ZPS	A	0.718	0.704	0.413	0.278
1ZRN	A	0.886	0.500	0.040	0.110
1ZS3	A	0.673	0.280	0.412	0.165
1ZT2	A	0.820	0.417	0.182	0.149
1ZT2	B	0.841	0.400	0.200	0.125
1ZTD	A	0.720	0.375	0.194	0.229
1ZV1	A	0.593	0.913	0.488	0.611
1ZVP	D	0.771	0.622	0.590	0.170
1ZVT	B	0.862	0.500	0.412	0.092
1ZX0	A	0.934	0.000	0.000	0.045
1ZXA	A	0.278	1.000	0.133	0.813
1ZXX	A	0.893	0.000	0.000	0.095
1ZZ1	A	0.924	0.378	0.737	0.015
1ZZW	A	0.844	0.200	0.118	0.109
2A01	A	0.761	0.000	0.000	0.236
2A10	A	0.740	0.545	0.774	0.117
2A15	A	0.805	0.500	0.077	0.186
2A1K	A	0.647	0.462	0.080	0.342
2A2L	A	0.724	0.447	0.472	0.178
2A6P	A	0.881	0.238	0.417	0.041
2A6S	A	0.530	0.571	0.195	0.478
2A7K	B	0.835	0.162	0.462	0.036
2A7U	B	0.762	0.500	0.240	0.204
2ADL	A	0.444	0.897	0.413	0.860
2AEB	A	0.917	0.375	0.125	0.069
2AFF	A	0.653	0.379	0.407	0.232
2AG4	A	0.665	0.000	0.000	0.314
2AGH	B	0.632	0.792	0.413	0.429
2AHF	A	0.947	0.100	0.500	0.006
2AHM	A	0.623	0.667	0.585	0.415

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2AHM	E	0.632	0.816	0.383	0.427
2AJT	A	0.849	0.472	0.446	0.081
2AN7	A	0.349	0.900	0.257	0.825
2ANE	A	0.727	0.545	0.375	0.227
2AO3	A	0.869	0.000	0.000	0.096
2ASK	A	0.614	0.743	0.464	0.455
2ASW	A	0.357	0.500	0.250	0.711
2AUK	A	0.624	0.260	0.277	0.245
2AZ4	A	0.899	0.000	0.000	0.071
2AZE	A	0.738	0.846	0.655	0.345
2AZE	B	0.673	0.833	0.714	0.629
2AZK	A	0.877	0.226	0.412	0.041
2B0J	A	0.794	0.250	0.014	0.200
2B30	A	0.884	0.091	0.133	0.050
2B3F	A	0.888	0.170	1.000	0.000
2B5A	A	0.610	0.647	0.314	0.400
2B7F	A	0.741	0.634	0.634	0.200
2B82	A	0.768	0.333	0.225	0.168
2B98	A	0.695	0.410	0.444	0.196
2B9D	A	0.365	0.625	0.143	0.682
2BA2	A	0.654	0.915	0.642	0.706
2BAY	A	0.661	0.667	0.348	0.341
2BB6	A	0.911	0.455	0.139	0.077
2BEM	A	0.900	0.200	0.182	0.056
2BF5	A	0.772	0.583	0.304	0.200
2BGR	A	0.911	0.286	0.250	0.043
2BGX	A	0.787	0.250	0.019	0.204
2BH1	A	0.866	0.444	0.414	0.081
2BH1	X	0.603	0.733	0.324	0.434
2BH8	A	0.706	0.889	0.667	0.500
2BHW	A	0.583	0.565	0.135	0.415
2BIW	A	0.935	0.000	0.000	0.036
2BJI	A	0.923	0.227	0.556	0.016
2BKM	A	0.727	0.500	0.229	0.241
2BKX	A	0.934	0.222	0.182	0.039
2BL2	A	0.667	0.383	0.605	0.156
2BL8	B	0.741	0.444	0.421	0.175

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2BM5	A	0.858	0.529	0.333	0.108
2BNM	A	0.716	0.606	0.323	0.261
2BO4	A	0.816	0.155	0.300	0.065
2BPS	A	0.700	0.846	0.333	0.328
2BSF	A	0.676	1.000	0.034	0.328
2BSK	A	0.575	0.909	0.517	0.700
2BVF	A	0.934	0.143	0.043	0.051
2BVU	A	0.642	0.500	0.159	0.339
2BWR	A	0.945	0.211	0.364	0.018
2BYC	A	0.757	0.231	0.115	0.187
2C0A	A	0.887	0.136	0.375	0.026
2C0G	B	0.747	0.267	0.082	0.218
2C12	A	0.830	0.357	0.612	0.055
2C1V	A	0.854	0.250	0.242	0.083
2C2U	A	0.848	0.364	0.167	0.120
2C81	A	0.944	0.500	0.087	0.052
2C92	A	0.721	0.409	0.545	0.146
2CB8	A	0.605	0.375	0.094	0.372
2CBI	A	0.923	0.800	0.114	0.072
2CC3	A	0.681	0.688	0.212	0.320
2CC6	A	0.438	0.625	0.132	0.589
2CCM	A	0.942	0.000	0.000	0.048
2CHC	A	0.631	0.333	0.319	0.260
2CHG	A	0.857	0.231	0.120	0.105
2CJG	A	0.901	0.500	0.023	0.097
2CJP	A	0.919	0.000	0.000	0.049
2CLY	A	0.476	0.709	0.500	0.780
2CLY	B	0.625	0.877	0.568	0.603
2CLY	C	0.530	0.935	0.500	0.829
2CMG	A	0.874	0.321	0.391	0.060
2CMZ	A	0.719	0.400	0.277	0.215
2CN3	A	0.966	0.000	0.000	0.023
2CNT	A	0.768	0.476	0.294	0.185
2CS7	A	0.611	0.783	0.529	0.516
2CU2	A	0.910	0.000	0.000	0.084
2CUA	A	0.795	0.500	0.080	0.195
2CW6	A	0.885	0.241	0.368	0.045

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2CXN	A	0.844	0.449	0.727	0.043
2CZQ	A	0.956	0.000	0.000	0.039
2CZS	A	0.300	1.000	0.020	0.710
2D00	A	0.633	0.741	0.606	0.473
2D0O	A	0.861	0.469	0.326	0.077
2D0O	B	0.806	0.636	0.519	0.151
2D3Q	A	0.945	0.000	0.000	0.031
2D73	A	0.927	0.161	0.294	0.030
2D7E	A	0.590	0.620	0.564	0.436
2D8D	A	0.750	0.881	0.712	0.395
2DBB	A	0.685	0.597	0.638	0.250
2DC4	A	0.841	0.438	0.292	0.115
2DDR	C	0.946	0.000	0.000	0.041
2DDZ	A	0.819	0.556	0.750	0.075
2DE3	A	0.953	0.000	0.000	0.032
2DF7	A	0.726	0.318	0.337	0.166
2DFJ	A	0.936	0.200	0.182	0.035
2DG1	A	0.894	0.200	0.032	0.095
2DI3	A	0.810	0.444	0.400	0.123
2DJ6	B	0.765	0.606	0.588	0.171
2DLA	A	0.847	0.750	0.341	0.144
2DPF	A	0.649	0.731	0.373	0.376
2DR3	D	0.851	0.385	0.556	0.059
2DRW	A	0.928	0.087	0.286	0.015
2DS2	B	0.627	0.852	0.523	0.525
2DS5	A	0.442	0.933	0.378	0.821
2DSC	A	0.744	0.525	0.585	0.162
2DSJ	A	0.939	0.438	0.292	0.042
2DSK	A	0.930	0.222	0.125	0.048
2DSN	A	0.951	0.250	0.133	0.034
2DT5	A	0.776	0.582	0.571	0.155
2DT7	B	0.424	0.700	0.132	0.613
2DTJ	A	0.787	0.674	0.580	0.174
2DUR	B	0.864	0.000	0.000	0.077
2DVM	A	0.760	0.252	0.491	0.081
2DVT	A	0.778	0.169	0.619	0.032
2DVY	A	0.852	0.804	0.607	0.136

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2DXN	A	0.841	0.538	0.596	0.087
2DY0	A	0.786	0.320	0.267	0.140
2DYJ	A	0.648	0.429	0.200	0.312
2DYR	C	0.595	0.322	0.578	0.188
2DYR	D	0.632	0.670	0.726	0.434
2DYR	E	0.619	0.489	0.564	0.283
2DYR	F	0.673	0.691	0.717	0.349
2DYR	H	0.544	0.750	0.462	0.596
2DYR	J	0.534	0.844	0.551	0.846
2DYR	L	0.630	0.853	0.707	1.000
2DYU	A	0.795	0.383	0.660	0.066
2E0Z	A	0.750	0.355	0.220	0.190
2E11	A	0.887	0.532	0.758	0.037
2E12	A	0.548	0.600	0.070	0.455
2E1M	B	0.722	0.838	0.803	0.636
2E1M	C	0.563	0.543	0.829	0.371
2E1N	A	0.664	0.636	0.326	0.330
2E2A	A	0.721	0.556	0.606	0.191
2E2X	A	0.852	0.400	0.114	0.129
2E5F	A	0.886	0.509	0.711	0.040
2E5Y	A	0.774	0.579	0.333	0.193
2E67	A	0.905	0.130	0.375	0.021
2E6F	A	0.875	0.488	0.553	0.063
2E79	A	0.704	0.231	0.120	0.232
2E7D	A	0.767	0.615	0.267	0.214
2E8G	A	0.888	0.000	0.000	0.097
2E8Y	A	0.954	0.667	0.100	0.041
2E9X	A	0.556	0.468	0.627	0.338
2E9X	B	0.663	0.323	0.541	0.150
2EAB	A	0.974	0.000	0.000	0.009
2EBY	A	0.657	0.783	0.375	0.380
2ECU	A	0.846	0.560	0.966	0.010
2ED6	A	0.771	0.467	0.583	0.120
2EGJ	A	0.762	0.789	0.366	0.243
2EGV	A	0.856	0.310	0.409	0.065
2EIX	A	0.872	0.444	0.133	0.111
2EIY	B	0.845	0.290	0.265	0.092

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2EJN	A	0.800	0.652	0.417	0.172
2EJQ	A	0.561	0.417	0.111	0.421
2EJW	A	0.858	0.190	0.381	0.045
2EK0	A	0.600	0.529	0.243	0.384
2ELC	A	0.918	0.095	0.200	0.026
2EQ5	A	0.842	0.269	0.318	0.079
2ERB	A	0.789	0.154	0.118	0.136
2ERV	A	0.760	0.800	0.103	0.241
2ET1	A	0.701	0.500	0.017	0.296
2ETX	B	0.841	0.091	0.045	0.114
2EX0	A	0.946	0.000	0.000	0.036
2EX2	A	0.895	0.250	0.022	0.102
2EZ9	A	0.880	0.069	0.067	0.069
2F01	B	0.633	0.421	0.421	0.268
2F07	A	0.763	0.500	0.304	0.193
2F23	A	0.773	0.579	0.289	0.200
2F2H	A	0.922	0.508	0.780	0.024
2F48	A	0.917	0.344	0.423	0.037
2F5G	A	0.738	0.638	0.638	0.205
2F5K	A	0.494	0.765	0.255	0.576
2F5V	A	0.877	0.500	0.016	0.143
2F6M	A	0.554	0.706	0.558	0.613
2F6M	B	0.645	0.767	0.426	0.403
2F8B	A	0.518	0.944	0.395	0.684
2F8J	B	0.829	0.339	0.568	0.059
2F9D	A	0.649	0.571	0.444	0.316
2FAO	A	0.908	0.350	0.333	0.051
2FB2	A	0.859	0.172	0.179	0.077
2FB5	A	0.770	0.333	0.068	0.210
2FD5	A	0.811	0.500	0.029	0.185
2FDV	A	0.908	0.238	0.179	0.055
2FE1	A	0.762	0.000	0.000	0.227
2FE8	A	0.816	0.341	0.311	0.113
2FF4	A	0.860	0.286	0.136	0.106
2FGQ	X	0.833	0.333	0.019	0.162
2FHZ	A	0.792	0.500	0.636	0.103
2FHZ	B	0.710	0.636	0.424	0.268

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2FI2	A	0.596	0.771	0.474	0.508
2FIA	A	0.822	0.250	0.136	0.131
2FIP	A	0.757	0.579	0.355	0.208
2FJR	A	0.804	0.182	0.174	0.114
2FK5	A	0.856	0.000	0.000	0.112
2FL4	A	0.664	0.750	0.059	0.338
2FLH	B	0.739	0.280	0.241	0.172
2FMY	A	0.784	0.414	0.286	0.159
2FN9	A	0.900	0.500	0.036	0.097
2FNU	A	0.887	0.431	0.629	0.040
2FP8	A	0.917	0.000	0.000	0.048
2FQL	A	0.607	0.000	0.000	0.382
2FQM	A	0.677	0.932	0.695	0.857
2FR5	A	0.779	0.646	0.705	0.148
2FSD	A	0.718	0.545	0.188	0.263
2FT0	A	0.793	0.125	0.107	0.126
2FT1	A	0.654	0.521	0.380	0.300
2FUG	A	0.826	0.299	0.523	0.059
2FUG	B	0.674	0.388	0.605	0.153
2FUG	G	0.662	0.459	0.596	0.204
2FV2	A	0.846	0.200	0.192	0.087
2FY9	A	0.796	0.854	0.875	0.385
2FYZ	A	0.719	0.949	0.725	0.778
2FZF	A	0.848	0.381	0.421	0.080
2G0B	B	0.831	0.444	0.414	0.105
2G30	A	0.850	0.200	0.115	0.106
2G38	A	0.636	0.758	0.556	0.455
2G38	B	0.711	0.415	0.395	0.197
2G3M	A	0.881	0.228	0.433	0.045
2G7O	A	0.191	0.500	0.018	0.818
2GAG	B	0.816	0.139	0.647	0.019
2GAG	C	0.774	0.380	0.613	0.086
2GAG	D	0.538	0.595	0.500	0.510
2GD7	A	0.514	0.878	0.483	0.793
2GDG	A	0.772	0.630	0.763	0.132
2GDQ	A	0.879	0.154	0.148	0.066
2GE7	A	0.645	0.638	0.685	0.347

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2GEC	A	0.761	0.778	0.326	0.242
2GFI	A	0.871	0.264	0.438	0.047
2GFP	A	0.877	0.667	0.043	0.121
2GFT	A	0.959	0.000	0.000	0.034
2GGS	A	0.901	0.067	0.071	0.050
2GH8	A	0.818	0.422	0.244	0.151
2GHT	A	0.822	0.375	0.100	0.157
2GHV	C	0.836	0.722	0.342	0.152
2GIA	B	0.589	0.459	0.298	0.367
2GIB	A	0.629	0.721	0.564	0.444
2GIY	A	0.804	0.583	0.359	0.161
2GJ2	A	0.557	0.556	0.139	0.443
2GLD	A	0.477	0.636	0.117	0.541
2GLX	A	0.898	0.083	0.143	0.039
2GMF	A	0.818	0.667	0.087	0.178
2GMH	A	0.924	0.000	0.000	0.042
2GMN	A	0.962	0.250	0.333	0.016
2GOY	A	0.748	0.200	0.250	0.132
2GR8	A	0.654	0.620	0.795	0.286
2GRU	A	0.929	0.286	0.571	0.018
2GSC	B	0.744	0.512	0.710	0.122
2GT1	A	0.876	0.273	0.086	0.103
2GTD	A	0.847	0.532	0.786	0.048
2GU9	A	0.694	0.536	0.417	0.253
2GUD	B	0.760	0.543	0.758	0.107
2GUZ	A	0.620	0.686	0.600	0.444
2GUZ	F	0.631	0.722	0.650	0.483
2GW6	A	0.642	0.400	0.146	0.324
2GZ1	A	0.826	0.417	0.481	0.091
2GZ4	A	0.750	0.511	0.471	0.176
2GZB	A	0.847	0.286	0.211	0.101
2H0Q	A	0.828	0.256	0.385	0.072
2H1C	A	0.734	0.444	0.229	0.223
2H1E	A	0.614	0.520	0.200	0.369
2H3H	B	0.938	0.333	0.462	0.024
2H6B	B	0.745	0.687	0.541	0.232
2H6F	A	0.832	0.672	0.534	0.132

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2H88	B	0.674	0.333	0.625	0.121
2H88	C	0.612	0.868	0.495	0.547
2H88	D	0.683	0.963	0.456	0.419
2H8G	A	0.862	0.531	0.472	0.089
2H9A	A	0.859	0.368	0.318	0.087
2H9A	B	0.893	0.372	0.727	0.023
2H9D	C	0.622	0.632	0.692	0.390
2HBA	A	0.365	0.846	0.262	0.795
2HBV	A	0.879	0.340	0.640	0.032
2HCV	A	0.812	0.346	0.766	0.035
2HDW	A	0.850	0.244	0.370	0.061
2HEK	A	0.911	0.429	0.300	0.060
2HF9	A	0.877	0.300	0.333	0.063
2HFN	A	0.711	0.200	0.421	0.101
2HH7	A	0.400	0.750	0.057	0.617
2HJ3	A	0.713	0.600	0.364	0.259
2HMV	A	0.799	0.594	0.559	0.140
2HNU	A	0.630	0.471	0.276	0.328
2HOX	B	0.817	0.318	0.574	0.058
2HQS	E	0.738	0.500	0.357	0.207
2HQT	B	0.725	0.348	0.308	0.186
2HQX	A	0.544	0.125	0.029	0.415
2HRA	A	0.789	0.286	0.125	0.169
2HRV	A	0.748	0.211	0.167	0.167
2HU9	A	0.785	0.333	0.273	0.143
2HY5	A	0.769	0.438	0.538	0.122
2HY5	B	0.621	0.233	0.206	0.265
2HY5	C	0.703	0.259	0.412	0.135
2I0X	A	0.571	0.500	0.056	0.425
2I1O	A	0.846	0.556	0.082	0.147
2I2Q	A	0.721	0.500	0.029	0.275
2I39	A	0.690	0.433	0.406	0.221
2I46	A	0.651	0.480	0.231	0.315
2I74	A	0.761	0.063	0.034	0.171
2I79	A	0.719	0.225	0.346	0.130
2I7D	A	0.850	0.333	0.318	0.087
2I8T	A	0.725	0.217	0.179	0.183

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2I99	A	0.881	0.267	0.348	0.053
2I9U	A	0.912	0.222	0.471	0.023
2IB0	A	0.704	0.250	0.083	0.254
2IC6	A	0.338	0.500	0.085	0.683
2ICY	B	0.894	0.000	0.000	0.103
2ID5	B	0.833	0.194	0.137	0.110
2IG3	A	0.764	0.533	0.258	0.205
2IG8	A	0.676	0.489	0.511	0.232
2IGI	A	0.817	0.387	0.462	0.094
2II3	D	0.829	0.709	0.619	0.134
2IMI	A	0.873	0.367	0.550	0.047
2INC	A	0.772	0.198	0.526	0.054
2INC	B	0.773	0.375	0.566	0.095
2INC	C	0.627	0.708	0.415	0.407
2INP	A	0.755	0.297	0.541	0.086
2INP	C	0.777	0.375	0.766	0.050
2INP	E	0.771	0.605	0.722	0.133
2IP2	A	0.767	0.492	0.382	0.173
2IPB	A	0.831	0.241	0.318	0.079
2IPI	B	0.919	0.185	0.278	0.032
2IRU	A	0.881	0.538	0.200	0.103
2ISK	A	0.685	0.505	0.771	0.143
2IU5	A	0.771	0.278	0.152	0.174
2IU8	A	0.676	0.535	0.445	0.267
2IUM	A	0.791	0.308	0.414	0.099
2IUT	A	0.824	0.174	0.070	0.138
2IVF	A	0.900	0.206	0.542	0.029
2IVF	B	0.757	0.370	0.741	0.061
2IVF	C	0.893	0.563	0.360	0.081
2IWV	A	0.679	0.500	0.079	0.312
2IXD	A	0.866	0.407	0.423	0.073
2IYG	A	0.706	0.471	0.258	0.250
2IYK	A	0.768	0.333	0.029	0.224
2IZW	A	0.775	0.406	0.382	0.144
2IZZ	A	0.716	0.409	0.590	0.137
2J04	A	0.862	0.308	0.145	0.114
2J0N	A	0.754	0.214	0.231	0.140

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2J1N	A	0.734	0.229	0.296	0.138
2J2J	A	0.692	0.328	0.618	0.110
2J4D	A	0.913	0.105	0.091	0.048
2J5D	A	0.333	1.000	0.286	0.909
2J6L	B	0.853	0.263	0.600	0.026
2J6P	A	0.759	0.387	0.429	0.140
2J9O	A	0.928	0.333	0.188	0.051
2JBV	A	0.947	0.053	0.111	0.019
2JCB	B	0.868	0.000	0.000	0.094
2JD3	A	0.700	0.833	0.678	0.452
2JD4	B	0.907	0.167	0.032	0.081
2JDA	A	0.820	0.100	0.059	0.124
2JE0	A	0.872	0.333	0.118	0.105
2JEE	A	0.615	0.791	0.618	0.600
2JIG	A	0.870	0.278	0.250	0.076
2JOD	A	0.604	0.615	0.333	0.400
2JRA	A	0.418	0.857	0.407	0.897
2JSC	A	0.656	0.852	0.442	0.420
2JWA	A	0.295	0.889	0.211	0.857
2JWK	A	0.595	0.688	0.306	0.431
2K29	A	0.740	0.946	0.761	0.846
2NLU	A	0.590	0.827	0.573	0.667
2NN4	A	0.710	0.818	0.360	0.314
2NNU	A	0.725	0.500	0.109	0.261
2NP9	A	0.835	0.286	0.348	0.082
2NPI	C	0.864	1.000	0.850	0.600
2NQ2	C	0.753	0.373	0.388	0.150
2NQR	A	0.800	0.509	0.341	0.155
2NT0	A	0.926	0.143	0.235	0.032
2NTE	A	0.829	0.000	0.000	0.121
2NTK	B	0.842	0.290	0.474	0.058
2NTP	A	0.883	0.167	0.028	0.104
2NW8	A	0.778	0.519	0.459	0.156
2NWI	A	0.713	0.379	0.282	0.214
2NX4	A	0.658	0.235	0.308	0.190
2NYA	A	0.962	0.231	0.214	0.026
2NYG	A	0.848	0.222	0.231	0.082

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2NYT	B	0.726	0.261	0.146	0.210
2NZX	C	0.823	0.200	0.080	0.139
2O09	A	0.857	0.000	0.000	0.109
2O1C	A	0.653	0.000	0.000	0.284
2O3S	A	0.845	0.250	0.057	0.135
2O4V	A	0.786	0.341	0.452	0.103
2O5F	B	0.858	0.364	0.200	0.106
2O70	A	0.800	0.103	0.300	0.051
2O7G	A	0.659	0.692	0.257	0.347
2O7M	A	0.732	0.400	0.216	0.218
2O8M	B	0.688	0.333	0.263	0.226
2O8X	A	0.508	0.889	0.364	0.651
2OAR	A	0.672	0.890	0.663	0.635
2OAU	A	0.614	0.622	0.397	0.389
2OCT	A	0.629	0.900	0.529	0.561
2ODD	A	0.420	0.833	0.270	0.711
2ODF	A	0.905	0.143	0.143	0.050
2OFK	A	0.896	0.300	0.200	0.070
2OGK	B	0.799	0.364	0.160	0.164
2OHC	A	0.820	0.424	0.298	0.129
2OHW	A	0.914	0.000	0.000	0.056
2OKX	A	0.926	0.250	0.227	0.041
2OMD	A	0.763	0.621	0.462	0.198
2OPE	A	0.700	0.250	0.192	0.210
2OQ2	C	0.899	0.308	0.500	0.035
2OQY	A	0.826	0.103	0.316	0.041
2OR2	A	0.902	0.100	0.105	0.048
2ORM	A	0.612	0.692	0.659	0.500
2ORY	B	0.904	0.214	0.353	0.035
2OSZ	A	0.523	0.714	0.446	0.608
2OU1	C	0.425	0.759	0.386	0.795
2OWA	A	0.734	0.389	0.233	0.209
2OX6	A	0.747	0.444	0.205	0.215
2OXG	F	0.752	0.636	0.583	0.197
2OYY	A	0.549	0.792	0.413	0.574
2P04	A	0.757	0.905	0.442	0.279
2P0M	A	0.923	0.100	0.087	0.050

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2P1J	A	0.738	0.761	0.522	0.271
2P1M	B	0.869	0.520	0.217	0.114
2P2O	A	0.761	0.000	0.000	0.000
2P38	A	0.710	0.500	0.133	0.273
2P3R	A	0.884	0.300	0.343	0.058
2P4W	A	0.624	0.662	0.485	0.397
2P4Z	A	0.902	0.318	0.368	0.047
2P54	A	0.891	0.300	0.120	0.086
2P5T	A	0.674	0.667	0.636	0.320
2P6P	A	0.919	0.200	0.462	0.020
2P90	A	0.822	0.488	0.447	0.115
2PA7	A	0.689	0.556	0.435	0.263
2PA8	D	0.750	0.464	0.203	0.216
2PA8	L	0.728	0.697	0.605	0.254
2PBX	A	0.817	0.400	0.323	0.122
2PD2	A	0.778	0.000	0.000	0.184
2PEZ	A	0.858	0.467	0.292	0.106
2PGD	A	0.797	0.667	0.029	0.157
2PI2	E	0.692	0.594	0.452	0.271
2PKD	D	0.673	0.538	0.378	0.284
2PL2	A	0.804	0.500	0.132	0.179
2PMV	B	0.869	0.120	0.188	0.054
2POK	B	0.857	0.466	0.466	0.082
2POS	A	0.745	0.250	0.100	0.209
2PQR	A	0.622	0.471	0.457	0.297
2PR1	A	0.757	0.471	0.222	0.207
2PS1	A	0.835	0.406	0.419	0.094
2PSO	A	0.650	0.125	0.016	0.328
2PT7	A	0.827	0.474	0.692	0.065
2PT7	G	0.745	0.667	0.391	0.237
2PTT	A	0.631	0.643	0.214	0.371
2PTT	B	0.676	0.444	0.242	0.278
2PUZ	A	0.891	0.353	0.353	0.059
2PVP	A	0.811	0.423	0.190	0.156
2PWJ	A	0.840	0.320	0.471	0.066
2PX0	A	0.837	0.200	0.444	0.046
2PYG	A	0.898	0.000	0.000	0.064

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2PYW	A	0.856	0.143	0.143	0.079
2PZH	C	0.735	0.639	0.511	0.229
2Q00	B	0.787	0.182	0.105	0.153
2Q0O	A	0.774	0.380	0.463	0.119
2Q0O	C	0.640	0.743	0.542	0.431
2Q2I	A	0.661	0.550	0.524	0.278
2Q46	A	0.877	0.000	0.000	0.119
2Q52	A	0.893	0.375	0.150	0.086
2Q7A	A	0.829	0.667	0.074	0.168
2Q7M	A	0.683	0.500	0.523	0.226
2Q87	A	0.757	0.250	0.150	0.179
2Q8N	C	0.802	0.167	0.027	0.167
2QBU	A	0.842	0.525	0.553	0.090
2QCU	A	0.898	0.083	0.029	0.078
2QCX	B	0.858	0.333	0.280	0.088
2QDL	A	0.786	0.600	0.250	0.194
2QEB	A	0.889	0.333	0.143	0.087
2QEE	G	0.824	0.268	0.629	0.039
2QF4	A	0.800	0.000	0.000	0.150
2QFC	A	0.736	0.632	0.150	0.257
2QFD	A	0.727	0.308	0.143	0.222
2QFI	A	0.703	0.000	0.000	0.287
2QJT	B	0.881	0.636	0.298	0.102
2QKL	B	0.744	0.750	0.455	0.257
2QKP	C	0.507	0.521	0.357	0.500
2QLC	A	0.786	0.280	0.438	0.089
2QMM	A	0.846	0.276	0.471	0.054
2QT3	A	0.880	0.205	0.320	0.047
2QTS	A	0.729	0.438	0.416	0.184
2QU7	B	0.876	0.179	0.313	0.045
2QUL	A	0.914	0.394	0.722	0.019
2QV6	A	0.779	0.457	0.640	0.098
2QVJ	A	0.774	0.402	0.418	0.136
2QXV	A	0.858	0.100	0.115	0.071
2QYA	A	0.722	0.483	0.452	0.198
2QYX	A	0.864	0.278	0.227	0.084
2QZ8	A	0.712	0.656	0.656	0.247

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2R0H	A	0.788	0.185	0.294	0.090
2R15	A	0.749	0.286	0.087	0.218
2R4F	A	0.750	0.412	0.553	0.124
2R5O	B	0.795	0.516	0.457	0.141
2R5U	A	0.694	0.556	0.319	0.274
2R6J	A	0.887	0.083	0.040	0.081
2R7A	A	0.882	0.375	0.375	0.065
2RCF	A	0.598	0.688	0.489	0.460
2RDE	A	0.732	0.345	0.196	0.210
2RE9	A	0.665	0.509	0.482	0.261
2RJI	A	0.726	0.364	0.200	0.219
2RJZ	A	0.726	0.676	0.500	0.255
2RL8	B	0.795	0.263	0.227	0.129
2SCP	A	0.851	0.150	0.250	0.058
2SPC	A	0.604	0.860	0.551	0.625
2SQC	A	0.963	0.200	0.214	0.026
2TBV	B	0.797	0.417	0.185	0.168
2UUZ	A	0.539	0.719	0.418	0.561
2UVL	A	0.670	0.682	0.385	0.333
2UWI	A	0.630	0.500	0.128	0.357
2UX0	A	0.679	0.524	0.244	0.293
2UXU	A	0.743	0.326	0.368	0.147
2V0O	A	0.718	0.443	0.723	0.108
2V1O	A	0.745	0.625	0.521	0.211
2V2G	A	0.832	0.482	0.771	0.049
2V3S	A	0.677	0.250	0.074	0.284
2V66	B	0.595	0.750	0.687	0.743
2V76	C	0.706	0.400	0.069	0.278
2VE3	A	0.926	0.267	0.160	0.050
2VEO	A	0.967	0.167	0.333	0.010
2VG0	A	0.828	0.343	0.429	0.083
2VHH	B	0.810	0.452	0.758	0.055
2VL6	A	0.753	0.510	0.385	0.190
2VLB	A	0.877	0.174	0.286	0.047
2VLG	C	0.608	0.529	0.220	0.376
2VLQ	A	0.643	0.650	0.361	0.359
2VLQ	B	0.664	0.000	0.000	0.219

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2VO9	A	0.723	0.182	0.059	0.234
2VOK	A	0.823	0.071	0.048	0.116
2VOU	A	0.934	0.300	0.136	0.050
2VQ7	A	0.903	0.351	0.433	0.043
2VSG	A	0.791	0.453	0.582	0.103
2VUG	A	0.839	0.381	0.320	0.103
2XAT	A	0.731	0.500	0.018	0.267
2YVA	A	0.751	0.500	0.354	0.195
2YVD	A	0.884	0.026	0.071	0.033
2YVE	A	0.760	0.387	0.343	0.160
2YVR	A	0.489	0.700	0.259	0.571
2YVS	A	0.849	0.667	0.429	0.125
2YXO	A	0.901	0.050	0.125	0.029
2YXZ	A	0.836	0.418	0.561	0.072
2YY0	A	0.605	0.957	0.611	0.933
2YY7	A	0.897	0.056	0.063	0.051
2YYS	A	0.901	0.133	0.118	0.056
2YYV	A	0.870	0.053	0.083	0.054
2YYY	A	0.880	0.263	0.435	0.043
2Z0A	A	0.597	0.739	0.425	0.469
2Z0J	A	0.861	0.036	0.143	0.029
2Z0T	A	0.697	0.059	0.056	0.185
2Z15	A	0.782	0.421	0.348	0.150
2Z1Y	B	0.810	0.263	0.513	0.060
2Z5A	A	0.621	0.451	0.371	0.310
2Z5A	B	0.800	0.396	0.288	0.142
2Z69	B	0.767	0.611	0.512	0.184
2Z6R	A	0.841	0.432	0.528	0.077
2Z73	A	0.803	0.143	0.016	0.184
2Z8F	A	0.973	0.167	0.143	0.015
2ZBT	A	0.833	0.567	0.630	0.093
2ZBT	C	0.793	0.238	0.294	0.105
2ZBT	D	0.828	0.571	0.519	0.116
2ZDH	A	0.925	0.333	0.500	0.027
2ZFZ	A	0.584	0.217	0.263	0.259
2ZGY	A	0.881	0.444	0.222	0.093
2ZIH	A	0.798	0.447	0.404	0.132

Protein	Chain	Accuracy	Sensitivity	Specificity	FAR
2ZJD	A	0.711	0.700	0.179	0.288
3B42	A	0.651	0.744	0.460	0.391
3B4R	A	0.697	0.227	0.093	0.250
3B5H	A	0.768	0.273	0.188	0.164
3B8F	A	0.730	0.581	0.556	0.204
3B9O	A	0.871	0.308	0.645	0.030
3BBZ	A	0.438	0.500	0.148	0.575
3BF7	A	0.898	0.000	0.000	0.050
3BFQ	G	0.750	0.281	0.474	0.100
3BJK	A	0.793	0.579	0.629	0.127
3BK6	A	0.665	0.667	0.299	0.336
3BLJ	B	0.818	0.278	0.179	0.128
3BOF	A	0.911	0.000	0.000	0.048
3BPJ	A	0.657	0.956	0.662	0.880
3BQS	A	0.671	0.500	0.143	0.312
3BRV	A	0.641	1.000	0.641	1.000
3BRV	B	0.590	0.886	0.596	0.808
3BS7	A	0.560	0.857	0.158	0.471
3BU8	A	0.711	0.258	0.182	0.208
3BY6	A	0.627	0.605	0.442	0.363
3C7B	A	0.739	0.470	0.700	0.112
3C7B	B	0.763	0.530	0.832	0.075
3C8I	A	0.732	0.455	0.152	0.241
3C9H	B	0.900	0.222	0.069	0.082
3C9U	A	0.814	0.625	0.493	0.143
3CHB	D	0.650	0.563	0.643	0.273
3CJH	A	0.712	0.917	0.733	0.750
3CJH	B	0.559	0.875	0.560	0.815
3CJS	B	0.653	0.667	0.276	0.350
3CLW	B	0.907	0.083	0.077	0.058
3EIP	A	0.607	0.889	0.200	0.427
3GRS	A	0.848	1.000	0.016	0.140
3TDT	A	0.752	0.500	0.044	0.243
7AHL	A	0.744	0.513	0.756	0.107

Appendix C: Classifier Accuracy Comparison for Several Methods (described in Section 4.3)

Protein	Topological	Cons-PPISP	SPPIDER	PPI-Pred	Promate
1A0C-A	0.789	0.751	0.801	0.664	0.712
1B67-A	0.561	0.515	0.456	0.588	0.544
1CHK-A	0.924	0.903	0.962	0.807	0.971
1DCI-A	0.711	0.778	0.771	0.749	0.724
1EEX-G	0.681	0.745	0.759	0.613	0.642
1F1M-A	0.750	0.691	0.802	0.846	0.741
1FSE-A	0.754	0.552	0.567	0.731	0.761
1GU7-A	0.928	0.953	0.909	0.780	0.945
1HF8-A	0.962	0.867	0.897	0.905	0.962
1IGQ-A	0.635	0.667	0.648	0.741	0.593
1J2R-A	0.801	0.814	0.830	0.761	0.723
1JOC-A	0.661	0.707	0.715	0.537	0.650
1KMT-A	0.846	0.870	0.877	0.920	0.920
1LJ2-A	0.606	0.613	0.585	0.547	0.566
1MKK-A	0.604	0.699	0.656	0.645	0.753
1NC5-A	0.983	0.989	0.994	0.882	0.978
1O5L-A	0.780	0.806	0.488	0.837	0.876
1ORJ-A	0.782	0.825	0.762	0.802	0.802
1PIX-A	0.837	0.836	0.800	0.853	0.831
1QLM-A	0.984	0.930	0.915	0.883	0.981
1R9D-A	0.994	0.950	0.968	0.887	0.969
1S1D-A	0.943	0.905	0.934	0.858	0.946
1SU1-A	0.813	0.766	0.799	0.821	0.761
1T56-A	0.937	0.933	0.617	0.731	0.922
1U1S-A	0.734	0.712	0.712	0.652	0.652
1UYP-A	0.935	0.926	0.882	0.843	0.912
1VLG-H	0.840	0.787	0.835	0.841	0.768
1WHI-A	0.883	0.803	0.656	0.844	0.984
1WWL-A	0.935	0.893	0.847	0.701	0.909
1XT5-A	0.932	0.837	0.896	0.778	0.881
1YF2-A	0.882	0.932	0.920	0.892	0.962
1Z6N-A	0.939	0.988	0.976	0.873	0.940
1ZVT-B	0.873	0.858	0.862	0.833	0.886
2ASK-A	0.657	0.614	0.485	0.653	0.663
2BJI-A	0.919	0.920	0.901	0.909	0.927
2CHC-A	0.711	0.708	0.667	0.732	0.702
2DDR-C	0.970	0.956	0.919	0.842	0.980
2DYJ-A	0.820	0.615	0.527	0.648	0.780
2EAB-A	0.986	0.983	0.990	0.920	0.992
2F6M-B	0.676	0.682	0.664	0.617	0.710
2FNU-A	0.892	0.898	0.901	0.861	0.869

Protein	Topological	Cons-PPISP	SPPIDER	PPI-Pred	Promate
2GE7-A	0.590	0.645	0.794	0.579	0.542
2GUZ-F	0.556	0.738	0.646	0.631	0.523
2HMV-A	0.803	0.856	0.647	0.806	0.806
2IG3-A	0.832	0.882	0.732	0.661	0.811
2IZW-A	0.841	0.702	0.539	0.803	0.820
2NNU-A	0.818	0.870	0.645	0.830	0.940
2OAU-A	0.734	0.713	0.535	0.697	0.709
2P04-A	0.848	0.822	0.645	0.897	0.776
2PTT-A	0.792	0.650	0.767	0.670	0.757
2QFD-A	0.798	0.860	0.843	0.851	0.851
2RJI-A	0.841	0.679	0.798	0.738	0.786
2VO9-A	0.877	0.899	0.791	0.743	0.824
2Z15-A	0.838	0.790	0.849	0.790	0.807
3BS7-A	0.822	0.907	0.867	0.773	0.800
1A0C-A	0.789	0.751	0.801	0.664	0.712
1B67-A	0.561	0.515	0.456	0.588	0.544
1CHK-A	0.924	0.903	0.962	0.807	0.971
1DCI-A	0.711	0.778	0.771	0.749	0.724
1EEX-G	0.681	0.745	0.759	0.613	0.642
1F1M-A	0.750	0.691	0.802	0.846	0.741
1FSE-A	0.754	0.552	0.567	0.731	0.761
1GU7-A	0.928	0.953	0.909	0.780	0.945
1HF8-A	0.962	0.867	0.897	0.905	0.962
1IGQ-A	0.635	0.667	0.648	0.741	0.593
1J2R-A	0.801	0.814	0.830	0.761	0.723
1JOC-A	0.661	0.707	0.715	0.537	0.650
1KMT-A	0.846	0.870	0.877	0.920	0.920
1LJ2-A	0.606	0.613	0.585	0.547	0.566
1MKK-A	0.604	0.699	0.656	0.645	0.753
1NC5-A	0.983	0.989	0.994	0.882	0.978
1O5L-A	0.780	0.806	0.488	0.837	0.876
1ORJ-A	0.782	0.825	0.762	0.802	0.802
1PIX-A	0.837	0.836	0.800	0.853	0.831
1QLM-A	0.984	0.930	0.915	0.883	0.981
1R9D-A	0.994	0.950	0.968	0.887	0.969
1S1D-A	0.943	0.905	0.934	0.858	0.946
1SU1-A	0.813	0.766	0.799	0.821	0.761
1T56-A	0.937	0.933	0.617	0.731	0.922
1U1S-A	0.734	0.712	0.712	0.652	0.652
1UYP-A	0.935	0.926	0.882	0.843	0.912
1VLG-H	0.840	0.787	0.835	0.841	0.768
1WHI-A	0.883	0.803	0.656	0.844	0.984
1WWL-A	0.935	0.893	0.847	0.701	0.909
1XT5-A	0.932	0.837	0.896	0.778	0.881

Appendix D: RuleFit Feature Selection

Feature Importance for initial feature selection for docking algorithm (described in Section 5.1.1)

Feature	**RuleFit Importance**
Mean interface residue tetrahedrality	100.0
Mean interface residue tetrahedrality / mean residue tetrahedrality	46.7
Ratio of interactions of class aromatic-small	44.2
Mean interface residue T5	42.1
Number of interface residues for protein A	41.8
Total number of interface residues	38.7
Ratio of interface / total residues for protein B	34.8
Ratio of interactions of class hydrophobic-aromatic	33.9
Mean interface residue volume / mean residue volume	33.2
Mean interface residue potential	30.6
Mean conservation of interface residues	26.7
Ratio of interface to total number of CYS residues	24.3
Mean interface residue potential / mean residue potential	21.9
Raton of interactions of class positively charged-negatively charged	21.7
Mean volume for T0 simplices	20.4
Number of interface residues for protein B	20.2
Total volume of simplices that cross interface / total volume of both chains	18.8
Ratio of interface to total number of SER residues	17.4
Ratio of interface to total number of TRP residues	17.4
Ratio of interface to total number of PRO residues	17.2
Ratio of interface to total number of VAL residues	15.1
Mean interface residue conservation / mean conservation of all residues	14.5
Mean interface residue T4	14.4
Ratio of interface to total number of LEU residues	13.6
Ratio of interface to total number of GLN residues	13.4
Ratio of interface to total number of PHE residues	13.1
Mean interface residue T5 / mean residue T5	11.6
Ratio of interface to total number of TYR residues	11.6
Ratio of interface to total number of GLY residues	11.6
Ratio of interactions of class hydrophobic-small	10.4
Ratio of interface to total number of ASN residues	10.3
Ratio of interactions of class positively charged-small	10.2
Ratio of interface to total number of LYS residues	9.6
Ratio of interactions of class aromatic-polar	9.5
Ratio of interactions of class polar-small	9.4
Ratio of interface / total residues for protein A	8.7
Mean interface residue T3	8.4
Mean interface residue T4 / mean residue T4	8.3

Feature	RuleFit Importance
Ratio of interactions of class hydrophobic-positively charged	8.3
Ratio of interactions of class positively charged-polar	7.5
Mean volume for T2 simplices	7.4
Volume of simplices that cross interface	6.6
Ratio of interaction of class hydrophobic-polar	6.3
Ratio of interface to total number of ARG residues	6.3
Ratio of interface to total number of GLU residues	6.1
Mean interface residue T1 / mean residue T1	6.0
Ratio of interface to total number of THR residues	5.5
Ratio of interface to total number of ASP residues	5.1
Ratio of interactions of class aromatic-positively charged	5.1
Mean potential for the conformation	4.9
Mean interface residue T3 / mean residue T3	4.7
Mean interface residue T2 / mean residue T2	4.4
Mean volume for T5 simplices	4.1
Mean volume for T3 simplices	4.1
Mean volume for T1 simplices	4.1
Total potential	4.0
Mean number of simplices interface residues participate in	4.0
Ratio of interactions of class negatively charged-small	3.7
Mean volume for T4 simplices	3.4
Ratio of interactions of class hydrophobic-negatively charged	3.4
Mean interface residue T1	3.3
Ratio of interface to total number of ALA residues	2.8
Mean interface residue T0 / mean residue T1	2.7
Mean interface residue Total / mean residue Total	2.5
Ratio of interface to total number of ILE residues	2.5
Ratio of interface / total residues for the conformation	2.4
Mean interface residue Volume	2.0
Ratio of interaction of class aromatic-negatively charged	1.9
Mean interface residue T2	1.8
Ratio of interface to total number of MET residues	1.6
Ratio of interactions of class negatively charged-polar	0.8
Ratio of interface to total number of HIS residues	0.7
Mean interface residue T0	0.1

Feature Importance for docking algorithm on data set with additional data (described in Section 5.1.2)

Feature	RuleFit Importance
Mean interface residue tetrahedrality	100.0
Mean interface residue tetrahedrality / mean residue tetrahedrality	46.9
Mean interface residue T5	42.7
Number of interface residues for protein A	40.8
Ratio of interactions of class aromatic-small	38.3
Ratio of interactions of class hydrophobic-aromatic	34.4
Total number of interface residues	33.0
Mean interface residue potential	31.8
Ratio of interface / total residues for protein B	31.3
Mean conservation of interface residues	26.5
Mean interface residue volume / mean residue volume	25.8
Number of interface residues for protein B	25.5
Ratio of interface to total number of CYS residues	24.9
Mean interface residue potential / mean residue potential	19.3
Ratio of interface to total number of PRO residues	18.6
Mean volume for T0 simplices	18.0
Raton of interactions of class positively charged-negatively charged	17.5
Ratio of interface to total number of TRP residues	17.3
Ratio of interface to total number of SER residues	16.7
Mean interface residue T4	15.2
Ratio of interface to total number of GLN residues	13.0
Ratio of interface to total number of VAL residues	12.8
Total volume of simplices that cross interface / total volume of both chains	12.8
Ratio of interface to total number of GLY residues	11.8
Volume of simplices that cross interface	11.8
Ratio of interactions of class positively charged-small	10.7
Ratio of interface to total number of LEU residues	10.6
Ratio of interface to total number of ASN residues	10.3
Mean interface residue T4 / mean residue T4	10.2
Mean interface residue T3	10.1
Ratio of interactions of class hydrophobic-small	9.4
Mean interface residue T5 / mean residue T5	9.4
Mean interface residue conservation / mean conservation of all residues	8.6
Mean volume for T5 simplices	8.5
Ratio of interface to total number of PHE residues	8.3
Ratio of interactions of class polar-small	7.7
Ratio of interface to total number of TYR residues	7.0
Mean interface residue T1 / mean residue T1	6.9
Ratio of interactions of class positively charged-polar	6.6
Total potential	6.3

Feature	RuleFit Importance
Ratio of interface to total number of ASP residues	6.2
Ratio of interactions of class hydrophobic-positively charged	6.2
Ratio of interface to total number of LYS residues	6.1
Ratio of interface / total residues for protein A	5.7
Mean potential for the conformation	5.2
Ratio of interactions of class aromatic-polar	5.2
Ratio of interface to total number of THR residues	5.1
Mean volume for T2 simplices	5.0
Ratio of interactions of class aromatic-positively charged	5.0
Mean interface residue T3 / mean residue T3	4.6
Ratio of interface / total residues for the conformation	4.0
Ratio of interaction of class hydrophobic-polar	3.8
Ratio of interface to total number of ARG residues	3.6
Mean interface residue Volume	3.5
Ratio of interactions of class hydrophobic-negatively charged	3.5
Mean volume for T1 simplices	3.5
Mean volume for T3 simplices	3.4
Ratio of interface to total number of GLU residues	3.1
Mean interface residue T2 / mean residue T2	2.8
Mean interface residue T1	2.6
Mean interface residue Total / mean residue Total	2.4
Mean interface residue T0 / mean residue T1	2.4
Mean volume for T4 simplices	2.3
Mean number of simplices interface residues participate in	2.1
Ratio of interface to total number of ILE residues	1.8
Ratio of interface to total number of ALA residues	1.6
Ratio of interface to total number of HIS residues	1.6
Ratio of interactions of class negatively charged-small	1.3
Ratio of interaction of class aromatic-negatively charged	1.1
Ratio of interactions of class negatively charged-polar	0.7
Ratio of interface to total number of MET residues	0.7
Mean interface residue T0	0.6
Mean interface residue T2	0.5

Feature Importance for docking algorithm on data subset with enzyme-inhibitor data (described in Section 5.1.3)

Feature	RuleFit Importance
Ratio of interface to total number of SER residues	96.8
Ratio of interface / total residues for protein B	85.0
Mean interface residue tetrahedrality	85.0
Ratio of interface to total number of CYS residues	80.4
Mean interface residue T4	71.1
Mean conservation of interface residues	69.0
Total number of interface residues	68.2
Ratio of interactions of class hydrophobic-aromatic	60.7
Mean interface residue T5	53.8
Mean interface residue T4 / mean residue T4	51.6
Ratio of interface to total number of PHE residues	47.8
Mean interface residue T5 / mean residue T5	43.2
Mean interface residue tetrahedrality / mean residue tetrahedrality	38.7
Ratio of interface to total number of GLN residues	32.2
Ratio of interface to total number of TRP residues	28.8
Mean interface residue T3	26.0
Ratio of interface to total number of ARG residues	25.2
Mean volume for T0 simplices	24.6
Number of interface residues for protein A	23.8
Ratio of interface to total number of HIS residues	21.6
Mean interface residue volume / mean residue volume	19.5
Mean interface residue conservation / mean conservation of all residues	18.5
Volume of simplices that cross interface	18.1
Mean interface residue potential	15.9
Mean volume for T5 simplices	15.6
Ratio of interface to total number of LEU residues	15.4
Ratio of interface to total number of GLY residues	14.9
Ratio of interface to total number of LYS residues	14.5
Ratio of interface to total number of ASN residues	14.2
Ratio of interface to total number of GLU residues	14.0
Mean interface residue T1 / mean residue T1	14.0
Total volume of simplices that cross interface / total volume of both chains	13.9
Ratio of interface to total number of ALA residues	13.6
Mean volume for T1 simplices	13.1
Raton of interactions of class positively charged-negatively charged	12.9
Ratio of interface to total number of ILE residues	12.9
Mean interface residue T3 / mean residue T3	10.9
Ratio of interactions of class hydrophobic-small	10.5
Total potential	9.5
Ratio of interactions of class aromatic-small	9.1

Feature	RuleFit Importance
Ratio of interactions of class polar-small	8.8
Ratio of interface to total number of ASP residues	8.7
Ratio of interface to total number of MET residues	8.1
Ratio of interaction of class hydrophobic-polar	7.6
Mean number of simplices interface residues participate in	6.3
Ratio of interactions of class aromatic-positively charged	6.1
Mean interface residue potential / mean residue potential	6.0
Mean volume for T2 simplices	6.0
Ratio of interactions of class hydrophobic-positively charged	5.7
Ratio of interactions of class aromatic-polar	5.6
Ratio of interactions of class positively charged-small	5.6
Ratio of interactions of class hydrophobic-negatively charged	5.1
Mean volume for T3 simplices	4.9
Number of interface residues for protein B	4.8
Mean interface residue T1	3.8
Ratio of interaction of class aromatic-negatively charged	3.6
Ratio of interface to total number of PRO residues	3.5
Ratio of interface to total number of THR residues	3.4
Ratio of interactions of class positively charged-polar	2.8
Mean interface residue Total / mean residue Total	2.6
Mean interface residue Volume	2.6
Mean interface residue T2	2.5
Ratio of interactions of class negatively charged-small	2.2
Ratio of interface / total residues for protein A	2.2
Mean volume for T4 simplices	2.0
Mean interface residue T2 / mean residue T2	1.8
Ratio of interface to total number of VAL residues	1.7
Mean potential for the conformation	1.6
Ratio of interface to total number of TYR residues	1.6
Mean interface residue T0	1.1
Mean interface residue T0 / mean residue T1	0.6
Ratio of interactions of class negatively charged-polar	0.5
Ratio of interface / total residues for the conformation	0.5

Feature Importance for docking algorithm on data subset with antibody-antigen data (described in Section 5.1.4)

Feature	**RuleFit Importance**
Ratio of interactions of class aromatic-small	98.9
Ratio of interface to total number of ASN residues	87.8
Mean interface residue tetrahedrality	71.4
Mean conservation of interface residues	69.3
Mean interface residue tetrahedrality / mean residue tetrahedrality	62.9
Mean interface residue conservation / mean conservation of all residues	61.0
Ratio of interface to total number of ILE residues	60.8
Ratio of interface to total number of ASP residues	51.8
Ratio of interface to total number of GLU residues	51.2
Ratio of interface to total number of GLN residues	47.4
Mean volume for T1 simplices	47.1
Mean volume for T2 simplices	44.7
Ratio of interface to total number of PRO residues	42.1
Ratio of interface to total number of THR residues	36.8
Ratio of interface to total number of VAL residues	32.6
Ratio of interaction of class aromatic-negatively charged	31.7
Ratio of interface to total number of TRP residues	30.8
Total volume of simplices that cross interface / total volume of both chains	27.7
Total number of interface residues	25.8
Ratio of interactions of class aromatic-positively charged	25.7
Ratio of interface to total number of PHE residues	25.4
Ratio of interface to total number of TYR residues	24.7
Ratio of interface to total number of SER residues	23.1
Mean interface residue T5	22.6
Number of interface residues for protein B	20.6
Mean volume for T0 simplices	19.1
Mean interface residue T2	19.0
Ratio of interactions of class aromatic-polar	18.6
Mean volume for T5 simplices	17.9
Ratio of interactions of class hydrophobic-aromatic	16.9
Mean interface residue T1	16.7
Ratio of interactions of class positively charged-polar	16.3
Ratio of interactions of class negatively charged-polar	16.1
Ratio of interface to total number of MET residues	16.0
Ratio of interface to total number of HIS residues	15.9
Raton of interactions of class positively charged-negatively charged	14.8
Mean interface residue T2 / mean residue T2	14.6
Ratio of interactions of class hydrophobic-small	14.4
Ratio of interface to total number of LEU residues	14.3
Ratio of interactions of class negatively charged-small	14.2

Feature	RuleFit Importance
Ratio of interactions of class polar-small	14.0
Mean interface residue T1 / mean residue T1	13.7
Number of interface residues for protein A	13.2
Mean interface residue T5 / mean residue T5	12.6
Mean interface residue volume / mean residue volume	11.6
Mean interface residue T0 / mean residue T1	11.4
Ratio of interactions of class hydrophobic-positively charged	11.2
Mean interface residue Total / mean residue Total	10.7
Mean interface residue T4	10.6
Ratio of interface to total number of ALA residues	10.0
Ratio of interactions of class hydrophobic-negatively charged	9.7
Mean volume for T4 simplices	9.7
Ratio of interface to total number of GLY residues	9.2
Mean number of simplices interface residues participate in	8.5
Mean interface residue T4 / mean residue T4	8.4
Ratio of interface to total number of LYS residues	8.1
Ratio of interface to total number of ARG residues	7.6
Mean interface residue T3 / mean residue T3	5.9
Mean interface residue T0	5.8
Mean interface residue potential / mean residue potential	5.3
Ratio of interaction of class hydrophobic-polar	5.2
Ratio of interactions of class positively charged-small	4.9
Ratio of interface to total number of CYS residues	4.6
Mean volume for T3 simplices	4.1
Mean potential for the conformation	3.6
Ratio of interface / total residues for protein B	3.4
Mean interface residue potential	2.7
Mean interface residue T3	2.5
Ratio of interface / total residues for protein A	2.5
Total potential	2.2
Mean interface residue Volume	2.0
Ratio of interface / total residues for the conformation	0.9
Volume of simplices that cross interface	0.1

Appendix E: Classification Results

Classification performance for docking algorithm on initial data set (described in Section 5.2.1)

Protein	**Position of native conformation**	**RMS of predicted top confirmation**	**fraction of correct receptor residues**	**fraction of correct ligand residues**	**Highest ranked confirmation with RMS $\leq$ 5 Å**
1A0O	553	12.59	0.400	0.667	149
1ACB	158	15.41	0.500	0.417	52
1AHW	42	31.3	0.700	0.350	12
1ATN	375	18.08	0.647	0.238	214
1AVW	41	20.98	0.333	0.286	29
1AVZ	206	16.43	0.500	0.833	156
1BQL	316	23.19	0.688	0.308	193
1BRC	364	12.31	0.625	0.556	298
1BRS	316	17.41	0.867	0.714	21
1BTH	24	18.3	0.464	0.900	24
1BVK	237	19.77	0.769	0.143	13
1CGI	64	1.89	0.889	0.875	1
1CHO	197	15.22	0.588	0.615	77
1CSE	52	13.58	0.360	0.667	30
1DFJ	1	0	1.000	1.000	1
1DQJ	8	5.23	0.800	0.750	8
1EFU	1	0	1.000	1.000	1
1EO8	241	16.89	0.800	0.688	241
1FBI	88	21.12	0.611	0.143	3
1FIN	9	12.02	0.767	0.467	5
1FQ1	297	18.43	0.353	0.478	38
1FSS	134	14.14	0.400	0.688	35
1GLA	573	29.41	0.846	0.625	341
1GOT	105	13.68	0.034	0.333	90
1IAI	14	4.97	0.842	1.000	1
1IGC	221	23.64	0.214	0.462	75
1JHL	203	20.49	0.769	0.182	139
1MAH	47	15.11	0.500	0.706	6

Protein	Position of native conformation	RMS of predicted top confirmation	fraction of correct receptor residues	fraction of correct ligand residues	Highest ranked confirmation with RMS $\leq$ 5 Å
1MDA	394	13.06	1.000	0.667	51
1MEL	2	7.93	0.154	0.100	2
1MLC	165	16.73	0.250	0.400	122
1NCA	2	22.26	0.619	0.813	2
1NMB	146	14.91	0.556	0.538	55
1PPE	7	6.92	0.682	0.917	2
1QFU	59	10.89	0.611	0.750	19
1SPB	20	9.92	0.353	0.824	20
1STF	126	0.89	0.952	0.933	1
1TAB	155	6.69	0.579	0.900	5
1TGS	8	7.09	0.417	0.467	6
1UDI	86	3.61	0.714	0.875	1
1UGH	9	0.48	0.952	0.895	1
1WEJ	116	14.43	0.769	0.545	116
1WQ1	1	0	1.000	1.000	1
2BTF	185	9.29	0.571	0.762	9
2JEL	30	23.21	0.588	0.467	7
2KAI	231	12.13	0.650	0.429	23
2PCC	507	13.8	0.778	0.900	220
2PTC	187	12.44	0.600	0.231	151
2SIC	162	15.21	0.571	0.167	41
2SNI	81	15.82	0.286	0.545	22
2TEC	143	21.51	0.333	0.583	27
2VIR	50	25.93	0.294	0.571	16
3HHR	1	0	1.000	1.000	1
4HTC	13	5.87	0.382	0.593	2

Classification performance for docking algorithm after re-ranking top 200 conformations (described in Section 5.2.1)

Protein	RMS of predicted top conformation after re-ranking	highest confirmation	position after re-ranking	original position
1A0O	12.59	16	160	149
1ACB	15.41	1	158	158
1AHW	31.3	1	40	42
1ATN	18.08	114	193	198
1AVW	20.98	1	37	41
1AVZ	16.43	11	153	156
1BQL	23.19	31	180	193
1BRC	17.06	195	191	186
1BRS	17.41	2	114	132
1BTH	18.3	1	28	24
1BVK	19.77	3	82	65
1CGI	1.89	1	63	64
1CHO	15.22	1	177	197
1CSE	13.58	1	49	52
1DFJ	0	1	1	1
1DQJ	5.23	1	5	8
1EFU	0	1	1	1
1EO8	16.89	17	161	165
1FBI	21.12	1	70	88
1FIN	12.02	1	9	9
1FQ1	18.43	8	173	181
1FSS	14.14	1	184	134
1GLA	29.41	88	160	164
1GOT	13.68	1	118	105
1IAI	4.97	1	13	14
1IGC	23.64	3	173	166
1JHL	20.49	14	128	139
1MAH	15.11	1	64	47
1MDA	14.66	29	185	181
1MEL	7.93	1	3	2
1MLC	16.73	1	173	165
1NCA	22.26	1	3	2

Protein	RMS of predicted top conformation after re-ranking	highest confirmation	position after re-ranking	original position
1NMB	14.91	1	138	146
1PPE	6.92	1	8	7
1QFU	10.89	1	61	59
1SPB	9.92	1	15	20
1STF	0.89	1	130	126
1TAB	6.69	1	148	155
1TGS	7.09	1	12	8
1UDI	3.61	1	86	86
1UGH	0.48	1	10	9
1WEJ	14.43	1	104	116
1WQ1	0	1	1	1
2BTF	9.29	1	182	185
2JEL	23.21	1	22	30
2KAI	12.13	2	151	150
2PCC	13.8	28	179	168
2PTC	19.16	1	183	187
2SIC	15.21	1	155	162
2SNI	15.82	1	80	81
2TEC	21.51	1	152	143
2VIR	25.92	1	46	50
3HHR	0	1	1	1
4HTC	5.87	1	18	13

REFERENCES

REFERENCES

[1] Ajay, Murcko, MA. "Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes," *Journal of Medicinal Chemistry* 38(26): 4953-4967, 1995.

[2] Ansari, S, Helms, V. "Statistical analysis of predominantly transient protein-protein interfaces," *Proteins* 61: 344-355, 2005.

[3] Aytuna, AS, Gursoy, A, Keskin, O. "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," *Bioinformatics* 21: 2850-2855, 2005.

[4] Baldi, P, Brunak, S, Chauvin, Y, Andersen, CAF. "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics* 16: 412-424, 2000.

[5] Barenboim, M, Jamison, DC, Vaisman, II. "Statistical geometry approach to the study of functional effects of human nonsynonymous SNPs," *Human Mutation* 26: 471-476, 2005.

[6] Bashford, D. "An object-oriented programming suite for electrostatic effects in biological molecules," In *Scientific Computing in Object-Oriented Parallel Environments*, volume 1343 of Lecture Notes in Computer Science, Yutaka Ishikawa, Rodney R. Oldehoeft, John V.W. Reynders, and Marydell Tholburn, editors, pages 233-240, Berlin. ISCOPE97, Springer, 1997.

[7] Baxter, CA, Murray, CW, Clark, DE, Westhead, DR, Eldridge, MD. "Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity," *Proteins* 33: 367-382, 1998.

[8] Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN, Bourne, PE. "The Protein Data Bank," *Nucleic Acids Research* 28: 235-242, 2000.

[9] Bernaer, J, Poupon, A, Aze, J, Janin, J. "A docking analysis of the statistical physics of protein-protein recognition," *Physical Biology* 2: S17-S23, 2005.

[10] Bernal, JD. "A Geometrical Approach to the Structure of Liquids," *Nature* 183: 141-147, 1959.

[11] Bonvin, AM. "Flexible protein-protein docking," *Current Opinion in Structural Biology* 16: 194-200, 2006.

[12] Bordner, AJ, Abagyan, R. "Statistical analysis and prediction of protein-protein interfaces," *Proteins* 60: 353-366, 2005.

[13] Bostick, D, Vaisman, II. "A new topological method to measure protein structure similarity," *Biochemical and Biophysical Research Communications* 304: 320-325, 2003.

[14] Bostick, DL, Shen, M, Vaisman. II. "A simple topological representation of protein structure: implications for new, fast, and robust structural classification," *Proteins: Structure, Function, and Bioinformatics* 56: 487-501, 2004.

[15] Bradford, JR, Westhead, DR. "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics* 21: 1487-1494, 2005.

[16] Breiman, L. "Random Forests," *Machine Learning* 45: 5-32, 2001.

[17] Budavari, S. (1989) *The Merck Index*. Merck & Co., New York.

[18] Bueno, M, Camacho, CJ. "Acidic groups docked to well defined wetted pockets at the core of the binding interface: a tale of scoring and missing protein interactions in CAPRI," *Proteins* 69: 786-792, 2007.

[19] Burgoyne, NJ, Jackson, RM. "Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces," *Bioinformatics* 22: 1335-1342, 2006.

[20] Caffrey, DR, Somaroo, S, Hughes, JD, Mintseris, J, Huang, ES. "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?," *Protein Sci* 13: 190-202, 2004.

[21] Camacho, CJ, Vajda, S. "Protein-protein association kinetics and protein docking," *Current Opinion in Structural Biology* 12: 36-40, 2002.

[22] Charifson, PS, Corkery, JJ, Murcko, MA, Walters, WP. "Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins," *J. Med. Chem* 42: 5100-5109, 1999.

[23] Chen, HL, Zhou, HX. "Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data," *Proteins* 61: 21-35, 2005.

[24] Chothia, C, Janin, J. "Principles of protein-protein recognition," *Nature* 256: 705-708, 1975.

[25] Chothia, C. "Hydrophobic bonding and accessible surface area in proteins," *Nature* 256: 705-708, 1975.

[26] Comeau, SR, Gatchell, DW, Vajda, S, Camacho, CJ. "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics* 20(1): 45-50, 2004.

[27] Comeau, SR, Kozakov, D, Brenke, R, Shen, Y, Beglove, D, Vajda, S. "ClusPro: Performance in CAPRI rounds 6-11 and the new server," *Proteins* 69: 781-785, 2007.

[28] Dalgaard, P. *Introductory Statistics with R.* Springer, 2002.

[29] Dandekar, T, Snel, B, Huynen, M, Bork, P. "Conservation of gene order: a finger print of proteins that physically interact," *Trends Biochem Sci* 23: 324-328, 1998.

[30] de Vries, SJ, Bonvin, AMJJ. "Intramolecular surface contacts contain information about protein-protein interface regions," *Bioinformatics* 22: 2094-2098, 2006.

[31] de Vries, SJ, van Dijk, AD, Krzeminski, M, van Dijk, M, Thureau, A, Hsu, V, Wassenaar, T, Bonvin, AM. "HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets," *Proteins* 69: 726-733, 2007.

[32] Dominguez, C, Boelens, R, Bonvin, AM. "HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information," *J. Am. Chem. Soc.* 125: 1731-1737, 2003.

[33] Espadaler, J, Romero-Isart, O, Jackson, RM, Oliva, B. "Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships," *Bioinformatics* 21: 3360-3368, 2005.

[34] Fariselli, P, Pazos, F, Valencia, A, Casadio, R. "Prediction of protein-protein interaction sites in heterocomplexes with neural networks," *Eur J Biochem* 269: 1356-1361, 2002.

[35] Fernandez-Recio, J, Abagyan, R, Totrov, M. "Improving CAPRI Predictions: Optimized Desolvaiton for Rigid-Body Docking," *Proteins* 60: 308-313, 2005.

[36] Fernandez-Recio, J, Totrov, M, Abagyan, R. "Identification of protein-protein interaction sites from docking energy landscapes," *J Mol Biol* 335: 843-865, 2004.

[37] Fernandez-Recio, J, Totrov, M, Skorodumov, C, Abagyan, R. "Optimal docking area: a new method for predicting protein-protein interaction sites," *Proteins* 58: 134-143, 2005.

[38] Finney, JL. "Modeling the structures of amorphous metals and alloys," *Nature* 266: 309-314, 1977.

[39] Finney, JL. "Random Packings and the Structure of Simple Liquids. I. The Geometry of Random Close Packing," *Proc. R. Soc.* A319: 479-493, 1970.

[40] Fitzjohn, PW, Bates, PA. "Guided Docking: First Step to Locate Potential Binding Sites," *Proteins* 52: 28-32, 2003.

[41] Friedman, JH, Popescu, B E. "Predictive Learning via Rule Ensembles." Technical Report, Stanford University, 2005.

[42] Gabb, HA, Jackson, RM, Sternberg, MJ. "Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information," *J. Mol. Biol.* 272: 106-120, 1997.

[43] Gallet, X, Charloteaux, B, Thomas, A, Brasseur, R. "A fast method to predict protein interaction sites from sequences," *J Mol Biol* 302: 917-926, 2000.

[44] Gao, Y, Douguet, D, Tovchigrechko, A, Vakser, IA. "DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking," *Proteins* 69: 845-851, 2007.

[45] Glaser, F, Steinberg, DM, Bakser, IA, Ben-Tal, N. "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins* 43: 89-102, 2001.

[46] Gong, XQ, Chang, S, Zhang, QH, Li, CH, Shen, LZ, Ma, XH, Wang, MH, Liu, B, He, HQ, Chen, WZ, Wang, CX. "A filter enhanced sampling and combinatorial scoring study for protein docking in CAPRI," *Proteins* 69: 859-865, 2007.

[47] Gottschalk, KE, Neuvirth, H, Schreiber, G. "A novel method for scoring of docked protein complexes using predicted protein-protein binding sites," *Protein Engineering, Design & Selection* 17(2): 183-189, 2004.

[48] Gray, JJ, Moughon, S, Wang, C, Schueler-Furman, O, Kuhlman, B, Rohl, CA, Baker, D. "Protein-Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations," *J. Mol. Biol.* 331: 281-299, 2003.

[49] Grosdidier, S, Pons, C, Solernou, A, Fernandez-Recio, J. "Prediction and scoring of docking poses with pyDock," *Proteins* 69: 852-858, 2007.

[50] Guharoy, M, Chakrabarti, P. (2007) Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics* 23(15): 1909-1918.

[51] Gunther, S, May, P, Hoppe, A, Frommel, C, Preissner, R. "Docking without docking ISEARCH-Prediction of interactions using known interfaces," *Proteins* 69: 839-844, 2007.

[52] Halperin, I, Wolfson, H, Nussinov, R. "Protein-Protein Interactions: Coupling of Structurally Conserved Residues and of Hot Spots across Interfaces. Implications for Docking," *Structure* 12: 1027-1038, 2004.

[53] Heifetz, A, Pal, S, Smith, GR. "Protein-protein docking: Progress in CAPRI rounds 6-12 using a combination of methods: The introduction of steered solvated molecular dynamics," *Proteins* 69: 816-822, 2007.

[54] Ho, TK. "Random Decision Forest," *Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition* 278-282, 1995.

[55] Hoffmann, D, Kramer, B, Washio, T, Steinmetzer, T, Rarey, M, Lenguaer, T. "Two-Stage Method for Protein-Ligand Docking," *J. Med. Chem.* 42: 4422-4433, 1999.

[56] Hu, Z, Ma, B, Wolfson, H, Nussinov, R. "Conservation of polar residues as hot spots at protein interfaces," *Proteins* 39: 331-342, 2000.

[57] Hwang, H, Pierce, B, Mintseris, J, Janin, J, Weng, Z. “Protein-Protein Docking Benchmark version 3.0,” *Proteins* 73(3):705-9, 2008.

[58] Jackson, RM, Gabb, HA, Sternberg, MJ. “Rapid Refinement of Protein Interfaces Incorporating Solvation: Application to the Docking Problem,” *J. Mol. Biol.* 276: 265-285, 1998.

[59] Jackson, RM, Sternberg, MJ. “A Continuum Model for Protein-Protein Interactions: Application to the Docking Problem,” *J. Mol. Biol.* 250: 258-275, 1995.

[60] Janin, J, Clothia, C. "The structure of protein-protein recognition sites," *J Biol Chem* 265: 16027-16030, 1990.

[61] Janin, J. “The targets of CAPRI rounds 6-12,” *Proteins* 69(4): 697-872, 2007.

[62] Jones, G, Willett, P, Glen, RC, Leach, AR, Taylor, R. “Development and Validation of a Genetic Algorithm for Flexibile Docking,” *J. Mol. Biol.* 267: 727-748, 1997.

[63] Jones, S, Marin, A, Thornton, JM. "Protein domain interfaces: characterization and comparison with oligomeric protein interfaces," *Protein Eng* 13: 77-82, 2000.

[64] Jones, S, Thornton, JM. "Prediction of protein-protein interaction sites using patch analysis," *J Mol Biol* 272: 133-143, 1997.

[65] Jones, S, Thornton, JM. "Prediction of protein-protein interaction sites using surface patches," *J Mol Biol* 272: 121-132, 1997.

[66] Jones, S, Thornton, JM. "Principles of protein-protein interactions," *PNAS* 93: 13-20, 1996.

[67] Jones, S, Thornton, JM. "Protein-protein interactions: a review of protein dimer structure," *Prog in Biophys & Mol Biol* 63: 31-65, 1995.

[68] Kabsch, W, Sander, C. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers* 22(12):2577-637, 1983.

[69] Kanamori, E, Murakami, Y, Tsuchiya, Y, Standley, DM, Nakamura, H, Kinoshita, K. “Docking of protein molecular surfaces with evolutionary trace analysis,” *Proteins* 69: 832-838, 2007.

[70] Keskin, O, Ma, B, Nussinov, R. "Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues," *J Mol Biol* 345: 1281-1294, 2005.

[71] Keskin, O, Tsai, CJ, Wolfson, H, Nussinov, R. "A new, structurally non-redundant diverse data set of protein-protein interfaces and its implications," *Protein Sci* 13: 1043-1055, 2004.

[72] Kohlbache, O, Burchardt, A, Moll, A, Hildebrandt, A, Bayer, P, Lenhor, HP. "Structure prediction of protein complexes by an NMR-based protein docking algorithm," *Journal of Biomolecular NMR* 20: 15-21, 2001.

[73] Koike, A, Takagi, T. "Prediction of protein-protein interaction sites using support vector machines," *Protein Eng Des sel* 17: 165-173, 2004.

[74] Koolman, J, Rohm, KH. *Colour Atlas of Biochemistry*, Thieme, Stuttgart, 1996.

[75] Kortemme, T, Baker, D. "Computational design of protein-protein interactions," *Curr Opin Struct Biol* 8: 91-97, 2004.

[76] Kortemme, T, Morozov, AV, Baker, D. "An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes," *J. Mol. Biol.* 326: 1239-1259, 2003.

[77] Kozakov, D, Brenke, R, Comeau, SR, Vajda, S. "PIPER: An FFT-based protein docking program with pairwise," *Proteins* 65(2): 392-406, 2006.

[78] Krol, M, Chaleil, RA, Tournier, AL, Bates, PA. "Implicit flexibility in protein docking: Cross-docking and local refinement," *Proteins* 69: 750-757, 2007.

[79] Kyte, J, Doolittle, R. "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.* 157: 105-132, 1982.

[80] Landau, M, Mayrose, I, Rosenberg, Y, Glaser, F, Martz, E, Pupko, T, Ben-Tal, N. "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information," *Nucl. Acids Res.* 33:W299-W302, 2005.

[81] Larsen, TA, Olson, AJ, Goodsell, DS. "Morphology of protein-protein interfaces," *Structure* 6: 421-427, 1998.

[82] Laskowski, RA, Chistyakov, VV, Thornton, J M. "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids," *Nucleic Acids Res.*, 33: D266-D268, 2005.

[83] Law, DS, Ten Eyck, LF, Katzenelson, O, Tsigelny, I, Roberts, VA, Pique, ME, Mitchell, JC. "Finding Needles in Haystacks: Reranking DOT Results by Using Shape Complementarity, Cluster Analysis, and Biological Information," *Proteins* 52: 33-40, 2003.

[84] Lensink, MF, Mendez, R, Wodak, SJ. "Docking and scoring protein complexes: CAPRI 3rd Edition," *Proteins* 69: 704-718, 2007.

[85] Li, X, Keskin, O, Ma, B, Nussinov, R, Liang, J. "Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organize in the unbound states: implications for docking," *J Mol Biol* 344: 781-795, 2004.

[86] Li, L, Chen, R, Weng, Z. "RDOCK: Refinement of Rigid-body Protein Docking Predictions," *Proteins* 53: 693-707, 2003.

[87] Li, N, Sun, Z, Jiang, F. "SOFTDOCK application to protein-protein interaction benchmark and CAPRI," *Proteins* 69: 801-808, 2007.

[88] Liu, S, Gao, Y, Vakser, IA. "Dockground protein-protein docking decoy set," *Bioinformatics* 24(22):2634-2635, 2008.

[89] Lo Conte, L, Chothia, C, Janin, J. "The atomic structure of protein-protein recognition sites," *J Mol Biol* 285: 2177-2198, 1999.

[90] London, N, Schueler-Furman, O. "Assessing the energy landscape of CAPRI targets by FunHunt," *Proteins* 69: 809-815, 2007.

[91] Lu, L, Lu, H, Skolnick, J. "Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading," *Proteins* 49: 350-364, 2002.

[92] Ma, B, Elkayam, T, Wolfson, H, Nussinov, R. "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces," *PNAS* 100: 5772-5777, 2003.

[93] Mandell, JG, Roberts, VA, Pique, ME, Kotlovyi, V, Mitchell, JC, Nelson, E, Tsigelny, I, Ten Eyck, LF. "Protein docking using continuum electrostatics and geometric fit, *Protein Engineering* 14(2): 105-113, 2001.

[94] May, A, Zacharias, M. "Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility," *Proteins* 69: 774-780, 2007.

[95] McCoy, AJ, Epa, VA, Colman, PM. "Electrostatic complementarity at protein/protein interfaces," *J Mol Biol* 268: 570-584, 1997.

[96] Moont, G, Gabb, HA, Sternberg, MJ. "Use of Pair Potentials Across Protein Interfaces I Screening Predicted Docked Complexes," *Proteins* 35: 364-373, 1999.

[97] Morris, GM, Goodsell, DS, Halliday, RS, Huey, R, Hart, WE, Belew, RK, Olson, AJ. "Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function," *J. Computational Chemistry* 19: 1639, 1998.

[98] Moustakas, DT, Lang, PT, Pegg, S, Pettersen, E, Kuntz, ID, Brooijmans, N, Rizzo, RC. "Development and validation of a modular, extensible docking program: DOCK 5," *J Comput Aided Mol Des* 20: 601-619, 2006.

[99] Muegge, I, Martin, YC. "A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach," *J. Med. Chem.* 42: 791-804, 1999.

[100] Neuvirth, H, Ras, R, Schreiber, G. "ProMate: a structure based prediction program to identify the location of protein-protein binding sites," *J Mol Biol* 338: 181-199, 2004.

[101] Nissink, JW, Murray, C, Hartshorn, M, Verdonk, ML, Cole, JC, Taylor, R. "A New Test Set for Validating Predictions of Protein-Ligand Interaction," *Proteins* 49: 457-471, 2002.

[102] Norel, R, Petrey, D, Wolfson, HJ, Nussinov, R. "Examination of Shape Complementarity in Docking of Unbound Proteins," *Proteins* 36: 307-317, 1999.

[103] Ofran, Y, Mysore, V, Rost, B. "Prediction of DNA-binding residues from sequence," *Bioinformatics* 23(13): i347-i353, 2007.

[104] Palma, PN, Krippahl, L, Wampler, JE, Moura, JJ. "BiGGER: A New (Soft) Docking Algorithm for Predicting Protein Interactions," *Proteins* 39: 372-384, 2000.

[105] Pang, YP, Perola, E, Xu, K, Prendergast, FG. "EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases," *J Comput Chem* 22(15):1750-1771, 2001.

[106] Pierce, B, Weng, Z. "ZRANK: Reranking Protein Docking Predictions With an Optimized Energy Function," *Proteins* 67: 1078-1086, 2007.

[107] Pons, C, Grosdidier, S, Solernou, A, Perez-Cano, L, Fernandez-Recio, J. "Present and future challenges and limitation in protein-protein docking," *Proteins* 78: 95-108, 2010.

[108] Porollo, A, Meller, J. "Prediction-based fingerprints of protein-protein interactions," *Proteins: Structure, Function and Bioinformatics* 66: 630-645, 2007.

[109] Qin, S, Zhou, HX. "A holistic approach to protein docking," *Proteins* 69: 743-749, 2007.

[110] Rey, A, Skolnick, J. "Efficient algorithm for the reconstruction of a protein backbone from the alpha-carbon coordinates," *J. Comput. Chem.* 13: 443, 1992.

[111] Rousseeuw, PJ, Leroy, AM. *Robust regression and outlier detection*, Wiley, 1987.

[112] Salwinski, I, Eisenberg, D. "Computational methods of analysis of protein-protein interactions," *Curr Opin Struct Biol* 13: 377-382, 2003.

[113] Sander, C, Schneider, R. "Database of homology-derived protein structures," *Proteins, Structure, Function & Genetics* 9:56-68, 1991.

[114] Schneidman-Duhovny, D, Nussinov, R, Wolfson, HJ. "Automatic prediction of protein interactions with large scale motion," *Proteins* 69: 764-773, 2007.

[115] Sheinerman, FB, Honig, G. "On the role of electrostatic interactions in the design of protein-protein interfaces," *J Mol Biol* 318: 161-177, 2002.

[116] Singh, RK, Tropsha, A, Vaisman, II. "Delaunay Tessellation of Proteins: Four Body Nearest Neighbor Propensities of Amino Acid Residues," *J. Comput. Biol.* 3: 213-222, 1996.

[117] Tame, JR. "Scoring functions: A view from the bench," *Journal of Computer-Aided Molecular Design* 13: 99-108, 1999.

[118] Taylor, T, Rivers, M, Wilson, G, Vaisman, II. "New Method for Protein Secondary Structure Assignment Based on a Simple Topological Descriptor," *Proteins: Structure, Function, and Bioinformatics* 60: 513-54, 2005.

[119] Terashi, G, Takeda-Shitaka, M, Kanou, K, Iwadate, M, Takaya, D, Umeyama, H. "The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation," *Proteins* 69: 866-872, 2007.

[120] Timberlake, KC. *Chemistry – 5th Edition*, Harper-Collins Publishers Inc, NY, 1992.

[121] Tsai, CJ, Lin, SL, Wolfson, HJ, Nussinov, R. "Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect," *Protein Sci* 6: 53-64, 1997.

[122] Vaisman, II, Tropsha, A, Zheng, W. "Computational Preferences in Quadruplets of Nearest Neighbor Residues in Protein Structure: Statistical Geometry Analysis," *Proceedings of the IEEE Symposia on Intelligence and Systems:* 163-168, 1998.

[123] Verdonk, ML, Cole, JC, Hartshorn, MJ, Murray, CW, Taylor, RD. "Improved Protein-Ligand Docking Using GOLD," *Proteins* 52: 609-623, 2003.

[124] Wang, C, Schueler-Furman, O, Andre, I, London, N, Fleishman, SJ, Bradley, P, Qian, B, Baker, D. "RosettaDock in CAPRI rounds 6-12," *Proteins* 69: 758-763, 2007.

[125] Wang, K, Fain, B, Levitt, M, Samudrala, R. "Improved protein structure selection using decoy-dependent discriminatory functions," *BMC Structural Biology* 4: 8-27, 2004.

[126] Weng, Z, Vajda, S, Delisi, C. "Prediction of protein complexes using empirical free energy functions," *Protein Science* 5: 614-626, 1996.

[127] Wiehe, K, Pierce, B, Tong, WW, Hwang, H, Mintseris, J, Weng, Z. "The performance of ZDOCK and ZRANK in rounds 6-11 of CAPRI," *Proteins* 69: 719-725, 2007.

[128] Witten, IH, Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[129] Wu, CH, Apweiler, R, Bairoch, A, Natale, DA, Barker, WC, Boeckmann, B, Ferro, S, Gasteiger, E, Huang, H, Lopez, R, Magrane, M, Martin, M., Mazumder, R, O'Donovan, C, Redaschi, N, Suzek, B. "The Universal Protein Resource (UniProt): an expanding universe of protein information," *Nucleic Acids Res.* 34: D187-191, 2006.

[130] Yan, C, Dobbs, D, Honavar, V. "A two stage classifier for identification of protein-protein interface residues," *Bioinformatics* 20: i371-i378, 2004.

[131] Yan, C, Dobbs, D, Honavar, V. "Identification of surface residues involved in protein-protein interaction - a support vector machine approach," In Abraham A, Franke K, and Koppen M, editors, *Intelligent Systems Design and Applications (ISDA-03)*: 53-62, 2003.

[132] Yao, H, Kristensen, DM, Mihalek, I, Sowa, ME, Shaw, C, Kimmel, M, Kavraki, L, Lichtarge, O. "An accurate, sensitive, and scalable method to identify functional sites in protein structures," *J. Mol Biol* 326: 255-261, 2003.

[133] Young, L, Jernigan RL, Covell DG. "A role for surface hydrophobicity in protein-protein recognition," *Protein Sci* 3: 717-729, 1994.

[134] Zhou, HX, Shan, Y. "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins* 44: 336-343, 2001.

[135] Zhou, HX, Qin, S. “Interaction-site prediction for protein complexes: a critical assessment,” *Bioinformatics* 23(17): 2203-2209, 2007.

CURRICULUM VITAE

Olivia Peters has maintained an interest in biotechnology and computational biology throughout her career. Her research interests have included the application of biologically-inspired signal processing techniques to the design of communications devices, novel algorithms for the analysis of microarray data, and computational modeling of inter-cellular processes. Ms. Peters has a B.S. in Electrical Engineering (University of Virginia, 1999), an M.S. in Biomedical Engineering (University of Virginia, 2000).