

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Prediction of Potential Host-Pathogen Protein Interactions by Structure

Permalink

<https://escholarship.org/uc/item/70f914xc>

Author

Davis, Fred Pejman

Publication Date

2007-05-07

Peer reviewed|Thesis/dissertation

Prediction of Potential Host-Pathogen Protein Interactions by Structure

by

Fred Pejman Davis

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2007

by

Fred Pejman Davis

To my parents, Kathy and Steve.

Acknowledgments

I owe the greatest debt of gratitude to my parents, Kathy and Steve. Their sacrifices, efforts, and encouragement are responsible for whatever accomplishments I may claim. Thanks to my brother Mike who welcomed me to New York City, my sisters, Kelly and Sara, who welcomed me to the Bay area, and everyone else who has helped me to enjoy these fine places. Thanks to the members of the Sali group I've had the pleasure of sharing the lab with, whose hospitality and kindness have made my years here enjoyable. They have shared their wisdom, sparred scientifically, and reminded me to 'be positive' when my frustrations grew overwhelming. Thanks to Dr. Andrej Sali for guidance, encouragement, and allowing me the freedom to pursue my interests. Andrej has assembled a bright and diverse group of scientists that make the lab a stimulating and enjoyable environment. Thanks to the members of my thesis committee, Drs. Tanja Kortemme and Brian Shoichet, for their helpful scientific comments and suggestions, as well as guiding me through the various steps and checkpoints in the graduate school process. Thanks to the members of my oral examination(s) committee, and the students that helped me practice, for ensuring that I had at least a passing knowledge of biophysical principles. Lastly, I am grateful for the creative efforts of Thelonious Monk, the Velvet Underground, Mastodon, John Coltrane, and the other fine musicians that have made the hours I spent behind the computer terminal almost enjoyable.

Chapter 2 is a reprint of the material as it appears in Davis and Sali, 2005 [39]. Andrej Sali directed and supervised the research that forms the basis of the chapter.

Chapter 3 is a reprint of the material as it appears in Davis, Braberg, Shen, Pieper,

Sali, and Madhusudhan, 2006 [40]. I designed the study together with M.S. Madhusudhan and Andrej Sali. Hannes Braberg implemented preliminary versions of the algorithm. Min-yi Shen aided in the design of the statistical potential. Ursula Pieper developed the MODBASE interface for visualization of the predictions. I wrote the manuscript together with Andrej Sali.

Chapter 4 describes work that will be submitted for publication. I designed the study with help from James H. McKerrow. David T. Barkan implemented and ran the sequence-based interaction predictions. Narayanan Eswar developed and ran the automated comparative protein structure modeling pipeline for the genomes described in the study. I analyzed the results with help from James H. McKerrow and Andrej Sali. I wrote the manuscript together with Andrej Sali.

This work is comparable to work for a standard thesis awarded by the University of California, San Francisco.

Abstract

Prediction of Potential Host–Pathogen Protein Interactions by Structure

by

Fred Pejman Davis

Proteins function through interactions with other biomolecules. Here I describe a series of tools developed and applied to study potential interactions between host and pathogen proteins. First, I describe a comprehensive relational database of structurally defined interfaces between pairs of protein domains, PIBASE. A diverse set of geometric, physico-chemical, and topologic properties are calculated to describe each complex, its domains, interfaces, and binding sites (<http://salilab.org/pibase>). This database allows a range of observations, from the atomistic detail of individual interfaces, to the structural organization of protein interaction space. Next, I present a comparative modeling method that uses experimentally determined structures of protein complexes as templates to predict the composition of protein complexes. Candidate complexes are assessed by comparative modeling of the components and subsequent assessment by a statistical potential derived from binary domain interfaces in PIBASE. Moreover, the predicted complexes were also filtered using functional annotation and sub-cellular localization data. The protocol was validated using experimentally observed interactions in *Saccharomyces cerevisiae* (<http://salilab.org/modbase>). Finally, I present a global computational protocol that generates testable predictions of potential host–pathogen protein interactions. The proto-

col first scans the total genomes for host and pathogen proteins with similarity to known protein complexes, then assesses these putative interactions, using structure if available, and finally filters these using biological context, such as the stage-specific expression of pathogen proteins and tissue expression of host proteins. The technique was applied to a set of ten pathogens, including species of mycobacterium, apicomplexa, and kinetoplastida, responsible for “neglected” human diseases. The method was assessed by (i) comparison to a set of known host–pathogen interactions, (ii) comparison to genomics data describing host and pathogen genes involved in infection, and (iii) analysis of the functional properties of the human proteins predicted to interact with pathogen proteins. The predictions include interactions known from previously characterized mechanisms, such as cytoadhesion and protease inhibition, as well as suspected interactions in hypothesized pathways, such as apoptotic pathways (<http://salilab.org/hostpathogen>). These results suggest that comparative protein structure modeling in combination with genomic and proteomic data can be a valuable tool for the study of inter-specific protein interactions.

Contents

List of Tables	xi
List of Figures	xii
1 Introduction	1
1.1 Introduction for non-scientists	1
1.2 The protein sequence to structure to function relationship	6
1.3 Protein–Protein Interactions	10
1.4 Host–Pathogen Interactions	12
1.5 Outline	13
2 PIBASE: A Comprehensive Database of Structurally Defined Protein Interfaces	14
2.1 Introduction	15
2.2 Methods and Results	19
2.2.1 Sources of protein structures and their classification	19
2.2.2 Detection of domain-domain interfaces	19
2.2.3 Properties of complexes, domains, interfaces, and binding sites	21
2.2.4 Redundancy Removal and Clustering	23
2.2.5 Implementation	25
2.2.6 Accessibility	26
2.2.7 Composition of PIBASE	27
2.3 Discussion	30
2.4 Supplementary Information: Hamann Distance Function	34
2.5 Supplementary Information: Topological Fingerprints	34
2.5.1 Complexes	34
2.5.2 Binding Sites and Interfaces	35
2.6 Supplementary Information: Kd-trees algorithm	35
3 Protein Complex Compositions Predicted by Structural Similarity	41
3.1 Introduction	43
3.2 Methods	45

3.2.1	Prediction Algorithm	45
3.2.2	Construction of Statistical Potentials	48
3.2.3	Benchmarking of Statistical Potentials	49
3.2.4	Validation of complex prediction	50
3.2.5	Binding Mode Selection	51
3.2.6	Data Sources	51
3.2.7	Technology	54
3.3	Results	55
3.3.1	Benchmark	55
3.3.2	Predictions	56
3.3.3	Validation	57
3.3.4	Comparison to other computational methods	58
3.3.5	Alternate Binding Modes	61
3.3.6	Co-complexed domains	62
3.4	Discussion	62
3.4.1	Accuracy	63
3.4.2	Importance of Structure	65
3.4.3	Alternative Binding Modes	66
3.4.4	Network Specificities	67
3.4.5	Extension of Known Co-complexed Domain Superfamilies	68
3.4.6	Future Directions	68
4	Host–Pathogen Protein Interactions Predicted by Structure	71
4.1	Introduction	73
4.2	Results	76
4.2.1	Detecting sequence and structure similarities	77
4.2.2	Identifying pairs of proteins with similarity to known complexes	78
4.2.3	Assessing the sequence or structural basis of the potential interactions	79
4.2.4	Applying biological and network-level filters	79
4.2.5	Assessment	80
4.2.6	Assessment I: Comparison of predicted and known host–pathogen protein interactions	81
4.2.7	Assessment II: Comparison to gene expression and essentiality data	86
4.2.8	Assessment III - Functional overview of predicted potential interactions	87
4.3	Discussion	93
4.3.1	Specific examples of potential interactions	93
4.3.2	Enrichment of potential interactions with actual interactions	100
4.3.3	Limitations in coverage	101
4.3.4	Errors in accuracy	102
4.3.5	Other computational methods	105
4.3.6	Potential impact	106
4.4	Materials and Methods	107
4.4.1	Detecting sequence and structure similarities	107

4.4.2	Identifying pairs of proteins with similarity to known interactions and assessing the sequence or structural basis of the potential interactions	109
4.4.3	Applying biological and network-level filters	110
4.4.4	Assessment: Functional overview of predicted complexes	111
4.4.5	Assessment: Comparison to gene expression and essentiality data . .	112
5	Conclusion	115
5.1	Summary	115
5.2	Future Directions	117
5.2.1	Improvements in coverage	117
5.2.2	Improvements in accuracy	118
5.3	Role of computation and structure in the investigation of inter-specific biomolecular networks	121

List of Tables

2.1	PIBASE content.	27
2.2	PIBASE properties.	37
4.1	Interaction template and biological data coverage of the genomes analyzed.	78
4.2	Potential interaction set reduction by assessment and filtering.	81
4.3	Comparison of predicted and known host–pathogen protein interactions. . .	85
4.4	Comparison of predictions to experimental observations of proteins involved in infection.	87
4.5	Functional annotation of human proteins predicted to interact with <i>M. tuberculosis</i>	92
4.6	Biological data characterizing host and pathogen proteins.	113

List of Figures

2.1	PIBASE Build Procedure.	18
2.2	Interactions in SCOP space.	28
2.3	PIBASE Interface Property Distributions.	30
2.4	Distribution of buried solvent accessible surface area in interacting SCOP domain pairs.	36
2.5	Topology properties calculated for each structure.	38
2.6	PIBASE Interface Property Distributions.	40
3.1	Prediction Logic Overview.	46
3.2	Assessment of statistical potentials.	54
3.3	<i>S. cerevisiae</i> predictions.	59
3.4	Experimental overlap of <i>S. cerevisiae</i> predictions.	60
3.5	Selection among alternate binding modes.	61
3.6	Co-complexed domain superfamilies.	62
4.1	Prediction protocol	77
4.2	Example of a validated prediction: falcipain-2 – cystatin-A.	94
4.3	Examples of predicted potential interactions.	95

Chapter 1

Introduction

This thesis is organized as follows. I first present an overview of the work that begins with an introduction for non-scientists and is followed by an introduction for scientists (Chapter 1). The body of the work then follows in three parts: PIBASE: A Comprehensive Database of Structurally Defined Protein Interface (Chapter 2), Protein Complex Compositions Predicted by Structural Similarity (Chapter 3), and Potential Host–Pathogen Protein Interactions Predicted by Structure (Chapter 4). I conclude by discussing future directions for these specific methods and for the role of computation and structure in the investigation of inter-specific biomolecular networks (Chapter 5).

1.1 Introduction for non-scientists

Proteins are one important kind of molecule in our bodies that carry out many of the different tasks required for life. In fact, all forms of life, from microscopic bacteria to the largest grey whale, have unique sets of proteins that are responsible for different tasks.

Some proteins interact with other proteins, with the DNA in your genome, or with ‘small molecules’ such as pharmaceutical drugs. For example, hemoglobin is a protein that binds to the oxygen in air that you breathe into your lungs and carries it through blood vessels to other cells in your body that need oxygen.

Each protein is made up of a string of building blocks called amino acid residues. There are 20 basic kinds of amino acid residues, but they can be decorated in different ways to make more individual kinds. The specific order and kind of amino acid residues in a protein is called its *amino acid sequence*. This linear chain of amino acid residues arranges itself in three-dimensions to form a unique three-dimensional shape. Every protein usually has one preferred 3D structure, although sometimes this preference changes when the protein is in different environments. This 3D structure determines how the protein interacts with other molecules in the body, similar to how the shape of Lego® connectors determines what Lego® blocks can connect to one another.

Tools have been created to take pictures of the 3D structures of proteins, when they are alone and when they are interacting with other molecules. This field of biology is called *structural biology*. Different tools can be used to study the structure of pictures at different resolutions. At the highest resolution, these structures describe almost exact positions of individual atoms in the proteins. Structural biologists around the world have determined the structures of thousands of proteins. When these biologists work in universities, and sometimes when they work for companies, they have made these structures freely available to the public.

Determining the structures of proteins is very expensive and requires a significant

amount of time and effort. However, actually determining the structure by experiment is not always necessary to learn about the structure of a protein. When the amino acid sequences of two proteins are similar, their 3D structures are also likely to be very similar. This relationship has sparked the development of computer programs that predict the 3D structure of one protein given the structure of another protein with a similar sequence. This procedure is called *comparative protein structure modeling*.

Many proteins carry out their functions by interacting with other proteins. Some proteins come together almost permanently to form a molecular machine. In other cases, pairs of proteins interact briefly to pass a message along. The structure of a protein determines what other proteins it interacts with and how these interactions occur. There are many ways to study protein–protein interactions that can teach us about the structure of the complex and strength of the interaction. The structure of a protein complex can be determined by the same kind of techniques used for individual proteins. As they do for individual proteins, the structures of complexes also help determine what kinds of functions the interactions mediate.

In addition to the interactions that occur between proteins from a single species, interactions are possible between proteins from different species. For example, interactions between human proteins and pathogen proteins, such as those from bacteria or viruses, occur during infection. These interactions are important for both the pathogen’s invasion of the human and for the human’s immune response against the pathogen. Knowledge of these host–pathogen interactions are important for two reasons. First, these interactions are key to understanding the molecular process of infection. Second, these interactions

highlight possible ways that infection can be treated. If a drug can be designed to inhibit these interactions, then the infection process may be halted.

In the work I present in this thesis, I have made predictions of interactions between human proteins and proteins from a set of microbes that cause human diseases including leprosy, tuberculosis, cryptosporidiosis, malaria, toxoplasmosis, Chaga's disease, African sleeping sickness, and leishmaniasis. Individual protein interactions have been observed between human proteins and microbial proteins, but a comprehensive survey of these interactions has not been performed for any infectious disease. There are experimental challenges that make studying these types of interactions more complicated than studying interactions within an organism. For this reason, computation can be a useful tool, since it does not suffer from the experimental difficulties of studying a human pathogen in the lab. In addition, computation doesn't expose laboratory workers to the dangers of studying microbes that cause human infectious diseases.

This work was done in three parts. First, I will describe a database that I built of protein interfaces, or the parts of proteins that interact with other proteins. I collected these interfaces out of the database of all protein structures and described them in different ways that capture the chemistry and geometry of the interfaces. These include properties such as the size of the area that the two proteins contact one another and what kinds of amino acids interact with one another. I've made this database freely available on the internet to help biologists interested in a specific protein or protein-protein interactions in general (<http://salilab.org/pibase>).

Second, I describe a comparative modeling procedure that uses experimentally

determined structures of protein complexes to predict protein-protein interactions. If two proteins are similar to a pair that have been previously observed to interact, then this is a clue they could also interact. I made a table of how often certain amino acids interact with one another across the protein interfaces in the database I described earlier. Then, I used this list to score how likely it is that two proteins interact with one another, given the types of amino acids that they contain. I used this program to predict protein-protein interactions in *Saccharomyces cerevisiae*, a yeast that is one of the most common organisms used to study molecular biology. Thousands of protein-protein interactions have been identified in this species, making it a good test case to assess the performance of our method. I've made these predictions freely available on the internet to help biologists who are interested in *S. cerevisiae* protein interactions (<http://salilab.org/modbase>).

Finally, I discuss how I used these tools to predict interactions between human proteins and proteins from micro-organisms that cause human infectious diseases. In addition to structures of the human and pathogen proteins, I also used information about the disease. For example, tuberculosis affects lung tissue. Therefore, if a human protein exists in lung tissue then it is more likely to encounter proteins from the bacteria that causes tuberculosis. Sometimes, it is even known where on a bacteria a certain protein exists. For example, some proteins occur on the surface of the bacteria, and so are more likely to encounter human proteins, than if they occurred deep inside the bacteria. We make thousands of predictions for the ten different infectious diseases. It is difficult to assess how good our predictions are since not much is known about the physical interactions that occur between human and pathogen proteins. However, many of our predictions make sense given what is

known about the different diseases, and in some cases help explain observed effects of the diseases. For example, we predict an interaction between a human protein that is known to be involved in the immune system and a protein, from the organism that causes malaria, that is known to stimulate the immune system. I've made these predictions available on the internet, so that scientists that are interested in a specific disease, or specific human or pathogen protein, can get ideas about potential interactions that may be involved in the infection process. (<http://salilab.org/hostpathogen>).

1.2 The protein sequence to structure to function relationship

Proteins function *in vivo* by interacting with other biomolecules, including small molecules, nucleic acids, as well as other proteins. The three-dimensional structure of a protein, which is governed by its primary amino acid sequence as well as its environment, determines the nature of the biomolecular interactions it engages in. Thus, knowledge of a protein's structure provides insight into its biological function.

The atomic structure of proteins can be modeled using experimentally determined information, such as the diffraction pattern observed by X-ray crystallography or the nuclear Overhauser effects (NOE) observed by nuclear magnetic resonance (NMR) spectroscopy. Both of these experimental methods however require a substantial amount of experimental resources and time. X-ray crystallography requires the expression, purification, and most importantly, crystallization of a protein sample in order to be subsequently analyzed under

an X-ray beam-line. Technological advances have been made in all of the steps required for protein crystallography, partly as a result of the large-scale structural genomics efforts. These advances include more efficient expression systems and robotic sampling of crystallization conditions. In addition, synchrotron light sources which emit brighter X-ray beams enable the collection of higher resolution data, in a shorter span of time, than conventional X-ray sources. NMR has also been used in the structural genomics efforts, and methods have been similarly improved. NMR also has practical lower limits in sample quality, such as the purity and concentration, and requires experimental effort in radio-labeling the protein to be studied. Recent advances in NMR spectroscopy have extended its applicability to larger complexes, for example the GroEL-GroES complex [57].

Both X-ray crystallography and NMR spectroscopy have practical upper bounds on the size of the complex to be structurally characterized. Other experimental techniques are applicable beyond these limits, although they typically exhibit lower resolution. These techniques include electron cryo-microscopy, electron cryo-tomography, single-angle x-ray scattering. Traditional biochemical techniques, such as co-immunoprecipitation, gold immuno-labeling, and biophysical techniques, such as analytical ultracentrifugation, can also provide useful information about the individual positions of the proteins in the complex, as well as the overall shape of the macromolecular assembly.

Computational methods play a pivotal role in structural biology; their contribution can be logically organized into two, related and non-exclusive, categories: (1) analyzing experimental data and (2) leveraging available experimental data to help describe systems for which experimental information is not available. First, computation has become essential

for the analysis of experimentally observed data, such as diffraction maps or NOE data. Structural modeling by both X-ray crystallography and NMR spectroscopy make use of molecular mechanics force fields to build structural models. Computational methods are also useful in integrating multiple sources of structural information. For example, density fitting algorithms are used to combine the atomic structure of an individual protein with the electron cryo-microscopy density map of larger multi-protein assemblies that are not amenable to X-ray crystallography [221, 222]. Other methods have also been developed that use many additional sources of information, such as indirect information gleaned from biophysical and biochemical studies, in order to help determine the structure of multi-protein assemblies [5].

The broad goal of traditional protein structural biology, as well as the more recent large-scale structural genomics efforts, is to determine the structure of proteins and their complexes, in order to gain a mechanistic understanding into protein function. Computational methods leverage these experimentally determined protein structures in order to expand the structural coverage of protein space, by applying available experimental data towards systems for which experimental information is not available. Comparative protein structure modeling has been developed to predict the structure of a protein based on its sequence similarity to a protein whose structure has been experimentally determined. *De novo* modeling techniques have also been developed that extract general rules of protein structure from the set of available experimental protein structures, and then apply these to predict the structure of proteins for which experimental information is not available. Computational methods have also been developed to predict the structures of multi-protein

assemblies. These include comparative modeling techniques, which infer the structure of an interaction given the structure of a complex between closely related proteins, as well as *de novo* protein docking methods, which aim to predict the structure of a complex given the structure of its components. In this thesis I will focus on a comparative modeling technique that I developed to predict physical protein interactions using structure.

The three-dimensional structure of a protein or protein assembly provides insight into its biological function. Several computational methods have been developed to help determine the function of a protein given its three-dimensional structure. Methods have been developed to identify functional active sites, such as catalytic sites, small molecule binding sites, and protein binding sites. Several methods also combine the structure with the sequence conservation in the family of homologs to identify regions that are likely to be functionally relevant. In addition to identifying sites that are likely to engage in interactions, structure can also be used to predict likely interaction partners, including small molecule ligands, protein partners, and DNA sequences. From quantum-level investigations into the catalytic mechanisms employed by enzymes that catalyze the reactions of life, to the atomic level detail of molecular recognition, structure enables the study of protein function at a variety of resolutions. In this thesis, I will focus on structural insights into physical protein–protein interactions, with a brief mention of protein–ligand interactions in this context.

In addition to providing mechanistic insight into function, the structure of a protein or protein assembly is an important morphological attribute for the study of protein evolution. Structure can be more useful than sequence in this respect, because similarities in tertiary structure can be detected over greater evolutionary timescales than similarities

in the primary amino acid sequence [35]. Thus, structure can often identify putative evolutionary relationships that are not obvious from the sequence alone. For example, structural modeling was used to identify remote structural relationships between components of the nuclear pore complex and coated vesicles, suggesting an ancient evolutionary relationship between these two compartmentalization systems [43, 44].

1.3 Protein–Protein Interactions

It has been estimated that each protein in the *Homo sapiens* proteome interacts with approximately 5-10 other proteins [81]. These include transient interactions as well as multi-protein assemblies that act as molecular machines [6]. Protein–protein interactions can be identified by a variety of techniques, including both traditional low-throughput and high-throughput techniques. Biochemical techniques, such as immunoprecipitation and tandem-affinity purification (TAP), can be used to isolate protein complexes formed *in vivo*. These complexes can then be further resolved by gel electrophoresis techniques such as two-dimensional PAGE. Mass spectroscopy (MS) can then be used to identify the proteins that form these complexes. The combination of TAP and MS (TAP-MS) has been used to comprehensively characterize physical protein interactions in the *Saccharomyces cerevisiae* proteome [66, 115]. Another method used to identify protein interactions in a high-throughput fashion is the yeast-two-hybrid (Y2H) genetic technique. Genome-wide Y2H has been used to characterize the protein–protein interaction networks in a variety of organisms, including *Saccharomyces cerevisiae* [93, 225] and *Homo sapiens* [185]. However, genome-wide Y2H suffers from an estimated 50% false positive and 90% false negative rates

[81].

Protein–protein interactions have been studied using a variety of genetic, biochemical, biophysical, and bioinformatic techniques [62]. These techniques provide information at a variety of spatial and temporal resolutions. At the highest spatial resolution, structural techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy provide detailed information about the atomic interactions that mediate biomolecular recognition. However, these methods suffer from an upper-limit on the size of complexes they can analyze. For this reason, electron cryo-microscopy and cryo-tomography are useful to study larger complexes, although they provide a lower resolution than X-ray or NMR.

While three-dimensional structures provide insight into the atomic detail of molecular recognition, they often do not enable the study of the thermodynamics and kinetics of interaction. For example, it is difficult to estimate the strength of an interaction using only the structures of protein complexes. Biophysical techniques are useful in characterizing the thermodynamics and kinetics of biomolecular interactions, characterizing parameters including heat capacities, entropy, enthalpy, equilibrium association constants, and the association and dissociation rate constants.

As they have for structure-determination techniques, computational tools have proven useful both in processing the data generated by large-scale experimental surveys, as well as leveraging this data to predict protein interactions that have not been experimentally observed. In this thesis I will develop a series of computational tools that aim to identify protein interactions based on structure.

1.4 Host–Pathogen Interactions

Infectious diseases account for over 25% of annual global deaths [242]. Over the past thirty years, this burden has been amplified by newly emerging diseases, such as HIV and SARS, as well as diseases such as tuberculosis and malaria, that are re-emerging as drug resistance becomes widespread. Recent technological advances, such as whole genome sequencing, genomic, and proteomic experiments have been applied to infectious diseases, increasing our understanding of the basic biology behind infection. Within one year of the 2003 SARS outbreak, the causative agent was identified as a new coronavirus by DNA microarray technology, the whole genome was sequenced, the structures of key proteins proposed by comparative modeling, and potential inhibitors identified by molecular docking. However, in general, this basic biology data has yet to have a significant impact on public health.

Elucidation of the interaction network between host and pathogen biomolecules is important for two reasons. First, the details of this network will illuminate the molecular basis of infectious diseases. Second, this network will highlight potential targets for chemotherapeutic or immunization strategies. Individual interactions between host and pathogen proteins have been identified using traditional biochemical and genetic techniques that focus on one protein or pathway. However, high-throughput techniques, such as TAP-MS, have not yet been used to identify direct physical interactions between host and pathogen proteins. Thus, the network of interactions between host and pathogen proteins remains largely uncharacterized.

The sparsity of experimental information makes computational tools especially

valuable for suggesting possible host–pathogen protein interactions. Specifically, this is a niche where computational tools may precede comprehensive experimental characterization, and provide valuable clues or starting points for subsequent experimental follow-up. Ultimately, the elucidation of the physical interactions that underly host–pathogen interaction will help the treatment and prevention of infectious disease.

1.5 Outline

Here, I will present a series of studies that work towards the prediction of potential interactions between host and pathogen proteins using structural information. First, I will present a database of protein structure interfaces that I developed to aide in the investigation of protein-protein interactions (<http://salilab.org/pibase>). I will discuss the development of the database, highlight some interesting observations, and discuss current applications. Second, I will present a comparative modeling method that uses experimentally determined structures of protein complexes as templates to predict physical protein–protein interactions. The predictions made for protein interactions in *S. cerevisiae* are publicly available through the MODBASE database (<http://salilab.org/modbase>). Finally, I will describe the use of these structural tools in conjunction with genomic and proteomic data to predict potential host–pathogen protein interactions formed in a set of ten human infectious diseases (<http://salilab.org/hostpathogen>). I conclude with a general discussion of the current, and future, roles of computation and structure in the investigation of inter-specific biomolecular networks.

Chapter 2

PIBASE: A Comprehensive Database of Structurally Defined Protein Interfaces

This chapter has been previously published [39].

Abstract

In recent years, the Protein Data Bank (PDB) has experienced rapid growth. To maximize the utility of the high resolution protein-protein interaction data stored in the PDB, we have developed PIBASE, a comprehensive relational database of structurally defined interfaces between pairs of protein domains. It is composed of binary interfaces extracted from structures in the PDB and the Probable Quaternary Structure (PQS) server using domain assignments from the Structural Classification of Proteins (SCOP) and CATH fold classification systems. PIBASE currently contains 158,915 interacting domain pairs between 105,061 domains from 2,125 SCOP families. A diverse set of geometric, physicochemical, and topologic properties are calculated for each complex, its domains, interfaces, and binding sites. A subset of the interface properties are used to remove interface redundancy within PDB entries, resulting in 20,912 distinct domain-domain interfaces. The complexes are grouped into 989 topological classes based on their patterns of domain-domain contacts. The binary interfaces and their corresponding binding sites are categorized into 18,755 and 30,975 topological classes, respectively, based on the topology of secondary structure elements. The utility of the database is illustrated by outlining several current applications. The database is accessible *via* the world wide web at <http://salilab.org/pibase>.

2.1 Introduction

Proteins do not act in isolation, but rather through interactions with molecules in their spatio-temporal environment that includes small molecules, nucleic acids, as well as

other proteins [6]. Therefore, the structures of individual proteins are often uninformative of biological function if taken out of context. Recent experimental advances have addressed this problem by enabling studies of protein interactions along two frontiers [186, 190]: (1) large-scale detection of protein-protein interactions [58, 65, 88, 93, 225] and (2) structure determination of protein complexes [187]. To maximize their utility, these experiments require informatics resources to store, organize, visualize, analyze, and disseminate the data. The objective is to understand the evolution and physics of protein interactions and to develop predictive models of protein structure and function.

Experimentally determined structures of protein complexes are deposited in the Protein Data Bank (PDB) [24]. The protein structure database is growing rapidly, in part due to the recent structural genomics effort [24, 48]. The PDB currently holds approximately 28,000 structures. Each entry contains on average 2.2 protein chains, and each chain contains on average 2.1 domains. Domains are considered the basal unit of protein structure, function, and evolution [169]. These units fold independently, often mediate a specific biological function, and combine modularly to form larger proteins. Several approaches to the definition of domain boundaries in proteins have been developed based on sequence and structure [232]. The Structural Classification of Proteins (SCOP) [148] and CATH [155] are two commonly used structure-based domain definition and classification systems.

Biologically relevant quaternary states are proposed for crystallographic protein structures by the Probable Quaternary Structure (PQS) server [83]. The server applies crystallographic and non-crystallographic symmetry operations to the PDB structure, and

then assesses the validity of each chain interface using a set of empirically derived cutoffs for properties such as buried solvent accessible surface area, buried number of residues, hydrogen bonding, and salt bridges. The PDB and the PQS are sources of the highest resolution protein-protein interaction data.

The structures of protein subunit interfaces have long been studied using collections of protein chain and domain interfaces [11, 15, 37, 87, 94, 98, 108, 160, 224]. Numerous analyses have used datasets of protein chain interfaces extracted from the PDB to investigate properties such as residue type propensities, sequence conservation, and structure conservation at protein interfaces [11, 29, 98, 99, 154, 228]. These studies of interface properties have given valuable insights into the physics and evolution of protein interactions.

In this paper, we describe a comprehensive relational database of structurally defined domain-domain interfaces. We annotate them by a diverse set of geometric, physicochemical, and topologic properties that characterize the structure of the protein complexes from the level of the complex to the atomic level details of each interface. A subset of these properties are used to remove interface redundancy as well as categorize the complexes, the interfaces, and the binding sites into topological classes. This multi-level characterization allows queries that span a range from properties of specific interfaces to proteome level views of interactions. The motivation in developing PIBASE has been to create a comprehensive repository of information characterizing the structure of protein complexes at a range of size scales using a diverse set of descriptors.

The construction of the database is described first, detailing the data sources, interface definition, properties computed, interface redundancy removal, and clustering (Meth-

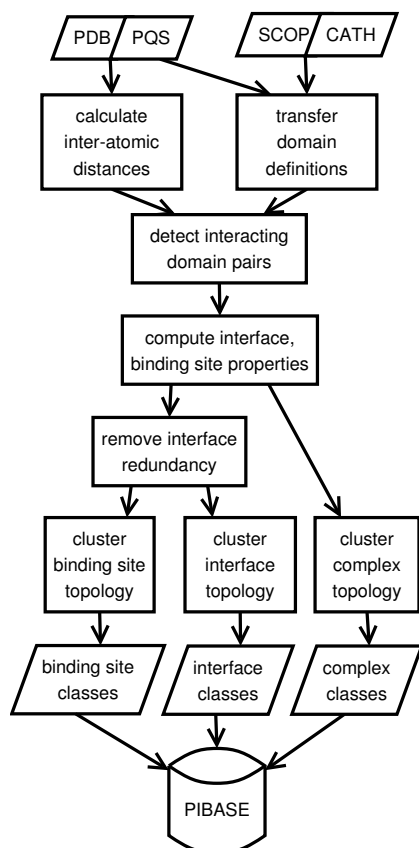


Figure 2.1: PIBASE Build Procedure. Briefly, inter-atomic distances calculated from PDB and PQS structures are combined with domain definitions from SCOP and CATH to generate a list of interacting domain pairs. A set of properties are then computed for all the interfaces and binding sites. A subset of these properties are then used to remove interface redundancy within structures associated with each PDB code. The binding sites, interfaces, and complexes are then clustered using their topological properties.

ods). We then discuss the composition of the database, describing the contents as well as the distributions of several of the computed interface properties (Results). Finally, we conclude with a brief discussion of several of our current applications of PIBASE (Discussion).

2.2 Methods and Results

2.2.1 Sources of protein structures and their classification

Two types of input data were used: protein structures and domain definitions. Structures were obtained from the Protein Data Bank [24] and the Probable Quaternary Structure (PQS) server [83]. For those structures determined by NMR spectroscopy, the first model in the ensemble was used.

Domain definitions for the PDB structures were obtained from the Structural Classification of Proteins (SCOP) [148] and CATH [155] classification systems. A mapping was generated between PDB and PQS chains that allowed domain definitions to be transferred from the PDB to its associated PQS entries. Approximately 1.5% of PQS entries contain chains with sequence changes, chain breaks, or chain mergers relative to their parent PDB structure. These differences occur for reasons such as missing density in the PDB structure. Domain definitions were not generated for these PQS entries as the chain mapping is difficult and inexact.

2.2.2 Detection of domain-domain interfaces

The list of binary interfaces was generated by a three step procedure (Figure 2.1). First, inter-atomic distances were calculated for all structures using a user specified distance cutoff. A cutoff of 6.05 Å was chosen unless specified otherwise, to allow contacts made *via* water molecules [180, 181]. The inter-atomic distances were then combined with the domain definitions to create a list of all domain pairs that share at least one inter-atomic distance below the specified distance threshold. This list of interacting domain pairs serves

as the core of PIBASE. Buried solvent accessible surface area (below) was also computed for each interacting domain pair and a minimum cutoff on the burial was imposed to yield the list of interfaces. Unless specified otherwise, a cutoff of 300 \AA^2 was used, as justified below.

It is often difficult to ascertain the biologically relevant quaternary state solely based on crystallographic information [31]. Previous studies have attempted to determine the biological relevance of observed chain contacts by two alternative strategies. The first is to define empirical thresholds on a set of interface properties (such as change in solvent accessible surface area, number of buried residues, etc) that are able to distinguish true biological interfaces from crystal packing artifacts [83]. The second strategy is to analyze the conservation of residues at the observed interface, with the hypothesis that a biological interface would be more conserved than a crystal packing artifact [49, 228]. While validation of both strategies is difficult, error rates of 4.3-20% have been reported [49, 83, 228]. The first strategy of empirical cutoffs has been implemented in an automated fashion in the PQS server [83]. PQS uses a threshold of 400 \AA^2 as one of the factors in determining biological relevance of a chain-chain interface. Unless otherwise mentioned, we used a lower threshold of 300 \AA^2 , which removed roughly 45% of the interacting SCOP domain pairs (Supplementary Figure 2.4). However, all interacting domain pairs are stored and all analysis can be performed easily with any choice of cutoff.

2.2.3 Properties of complexes, domains, interfaces, and binding sites

For each processed structure, PIBASE contains a set of properties at four different levels: the complex, its constituent domains, binary interfaces, and binding sites (the two halves of an interface) (Supplementary Table 2.2). These properties are described next.

Complexes. A *domain connectivity graph* was generated for each structure, describing the pattern of domain-domain contacts. In this graph, a domain is a node, and a binary domain interface is an edge (Supplementary Figure 2.6(b)). This graph captures the arrangement of the individual domains in the complex. A crude linear representation of this graph is then computed and used as a topological fingerprint to group the structures into topological classes (Supplementary Information: Topological Fingerprints, Supplementary Figure 2.6(d)). While degeneracy exists in this representation (*ie*, two distinct topologies may have the same sorted edge list), it is useful as both a query and crude clustering term.

Domains. *Solvent accessible surface area* was computed for each domain using a probe radius of 1.4 Å with the algorithm of Richmond and Richards as implemented in MODELLER [179, 189]. *Secondary structure* assignments are made by DSSP [100]. *Classification codes* (*eg*, class, fold, superfamily, and family) from the domain assignment system used, SCOP or CATH, are also associated with each domain.

Interfaces and Binding Sites. The interface and binding site properties compose the majority of PIBASE content. A subset of these properties were used for redundancy removal and clustering of the interfaces. The interface properties can be grouped into two categories: non-contact and contact. Non-contact properties (properties 1-8 in the Interface column of Supplementary Table 2.2(c)) are properties that are a sum or a union of the

properties in the corresponding domains. For instance, the number of residues presented at the interface is computed as a sum of the number of residues presented by each of the two binding sites. The contact properties (properties 9-17 in the Interface column of Supplementary Table 2.2(c)) implicitly capture the interface orientation. These properties can not be defined independently by the two binding sites. The binding site properties fall into two categories: non-topology and topology. Non-topology properties (properties 1-8 in the Binding Site column of Supplementary Table 2.2(c)) describe the size and physicochemical properties of the binding site. Topology properties (properties 9-14 in the Binding Site column of Supplementary Table 2.2(c)) describe the local structure of the binding site.

The *change in solvent accessible surface area* (defined as $\Delta\text{SASA}_{AB} = \text{SASA}_A + \text{SASA}_B - \text{SASA}_{AB}$), *number of residues*, and *number of secondary structure elements* describe the extent of the interface. The *residue types present*, *secondary structure types present*, and *change in polar solvent accessible surface area* are more fine-grained properties describing the actual physical structures present and their chemical composition. Two measures of binding site continuity were computed. The *number of structural patches* was determined by counting the number of connected components in a graph representation of binding site residues where an edge was placed between residues within 6 Å of each other. The *number of sequence segments* was counted to describe the sequence continuity of each binding site. The continuity properties of the interfaces were calculated by summing the properties from their corresponding binding sites.

The *number of residue pairs*, *number of secondary structure element contacts*, and *number of inter-atomic contacts* describe a combination of the extent and complexity of

the interface. The *residue contact types*, *secondary structure contact types*, and *secondary structure topology* capture the nature and complexity of the interface. The inter-atomic contacts are further categorized into *Van der Waals contacts*, *hydrogen bonds*, *salt bridges*, and *disulfide bridges* based on distance criteria (H-bond criteria as defined by JOY [142]; disulfide bridge defined when two Cys S atoms are closer than 3.0 Å).

As for the domain connectivity graphs, a crude linear representation of the secondary structure topology graph was generated for use as a topological fingerprint to group the structures into topological classes (Supplementary Information: Topological Fingerprints, Supplementary Figure 2.6(c), 2.6(d)).

2.2.4 Redundancy Removal and Clustering

Two types of clustering were performed on the domain-domain interfaces. The first procedure, redundancy removal, aims to provide a non-redundant set of interfaces for analysis by addressing the issue of duplicate interface structures. The second procedure, clustering, aims to group together similar interfaces to aid in the understanding of interface diversity. Although both procedures involve clustering, they serve different purposes.

Removal of Redundancy of domain-domain interfaces

Redundancy, in the form of duplicate interface structures, exists for several reasons: redundancy within PDB entries, interfaces duplicated in derived PQS structures, and redundancy across different PDB entries. The first two types of redundancy are explicitly addressed by hierarchically clustering interfaces associated with each PDB code using

a distance function that combines the following properties: types of residue-residue contacts present (represented as a 210-bit vector, aa), buried solvent accessible surface area ($\Delta SASA$), and the number of residues in the interface ($numres$) (Eq 2.1). The bit vectors were compared using the Hamann distance measure, $dist_{hamann}$, a rescaled and reversed version of the traditional Hamann similarity coefficient, sim_{hamann} , developed for use in plant systematics (Supplementary Information: Eq 2.3, 2.3) [78]. The resulting dendrogram was cut into clusters using a strict threshold of 0.1. This cutoff corresponds to maximum differences of 10% in the buried surface areas, 10 % in the numbers of residues, or 0.1 Hamann distance in the residue-residue type contact vectors. The cluster membership of each interacting domain pair is stored in PIBASE. The clustering was performed on the interacting domain pairs list, prior to the buried surface area filter. This procedure identified approximately 75% of the domain pairs as redundant (Table 2.1).

$$\begin{aligned}
 d_{A,B} = & \frac{1}{3}(\text{dist}_{hamann}(aa_A, aa_B) \\
 & + (1 - \frac{\min(\Delta SASA_A, \Delta SASA_B)}{\max(\Delta SASA_A, \Delta SASA_B)}) \\
 & + (1 - \frac{\min(numres_A, numres_B)}{\max(numres_A, numres_B)}))
 \end{aligned}
 \tag{2.1}$$

The third type of redundancy, duplicate interfaces across PDB entries, can also be addressed in a similar fashion, but is not yet implemented. While the minimal three property set was found to be effective at recognizing interface similarity within a PDB file, a different and likely larger set of interface properties are required for a more general interface similarity measure. However, the choice of specific properties to use depends heavily on the

definition of redundancy, and the intended application. As such, we leave this clustering up to the user, while providing the appropriate tools (*eg*, properties, clustering algorithms).

Clustering of complexes, interfaces, and binding sites

The topological fingerprints were used to group the non-redundant complexes, interfaces, and binding sites into discrete topological classes. The complexes were grouped according to their domain connectivity (Supplementary Figure 2.6(b)), while the interfaces and binding sites were grouped according to the topology of their secondary structure elements (Supplementary Figure 2.6(c)). The clustering reveals 989, 18,755, and 30,975 topological groups of complexes, interfaces, and binding sites, respectively.

In the current implementation, groups are formed by members with identical topological fingerprints. A more refined distance metric for topology fingerprints would be useful in describing a continuous gradation of topology similarity. However, such a clustering will depend on a specific application, and is therefore beyond the scope of the current paper.

2.2.5 Implementation

PIBASE was implemented using the MySQL relational database system (<http://www.mysql.com>). It was built by a set of Perl programs using the DBI interface to communicate with the MySQL system. Most properties were computed with MODELLER [188]. Secondary structure assignments were made by DSSP [100]. Inter-atomic distances were computed using an in-house ANSI C implementation, called kd-contacts, of the median kd-trees algorithm [23, 61]. The kd-trees algorithm, a commonly used computational

geometry algorithm, performs nearest neighbor queries by first building a tree in $O(n \log n)$ time, and then querying it in $O(n^{1-1/d} + k)$ time, where n is the number of data points in the d -dimensional space, and k is the number of reported points (Supplementary Information: kd-trees algorithm). This approach is much faster than the naive approach of all vs all distance calculation ($O(n^2)$). The logarithmic scaling allows rapid calculation of contact maps even for large structures with tens of thousands of atoms, such as PQS entries of virus capsids.

The clustering of the distance matrix for redundancy removal was performed using an in-house Perl library. The calculations were done in a parallel fashion on 50 2.6 GHz Pentium IV processors in approximately 15 hours. The database is updated automatically with every SCOP and CATH release.

2.2.6 Accessibility

PIBASE is accessible *via* the world wide web at <http://salilab.org/pibase>. The interface allows the user to query the database by PDB codes, complex topology fingerprints, and domain classification codes. The range of possible queries will expand as users request additional functionality.

While a web interface is well suited for standard queries with relatively simple conditions, a programming interface can be more useful for complex queries. A Perl library, used in the construction of the database, will be released shortly, allowing complex queries to be performed without the complexity of directly accessing the underlying MySQL structures. In addition, the contents of the database tables, as well as a schema describing the logical

<i>Structures</i>		
Structures (PDB & PQS)	38,940	
Associated PDB codes	20,740	
	<i>SCOP</i>	<i>CATH</i>
Domains	120,110	103,246
<i>Interfaces</i>		
Interacting domain pairs	158,915	138,286
Interfaces ($\Delta\text{SASA} \geq 300 \text{ \AA}^2$)	86,127	76,746
<i>Redundancy</i>		
Interacting domain pairs unique within structure file	77,105	
Interacting domain pairs unique within PDB code	41,493	
Unique and $\Delta\text{SASA} \geq 300 \text{ \AA}^2$	20,912	

Table 2.1: PIBASE content. Unless otherwise noted, all the numbers shown represent data obtained from both PDB and PQS structures. The number of interacting domain pairs are shown using both SCOP and CATH definitions. The interface clustering was performed only on the SCOP pairs. The ΔSASA filter of 300 \AA^2 removes $\sim 45\%$ of the interacting domain pairs. The redundancy removal procedure flags $\sim 75\%$ of the interfaces as redundant.

relationships between the tables, are available for download.

2.2.7 Composition of PIBASE

Briefly, PIBASE currently contains 158,915 interacting domain pairs between 105,061 domains from 2,125 SCOP families. More interface structures are available between domains from the same SCOP family (1405 homo-family pairs) than different SCOP families (982 hetero-family pairs) (Figure 2.2(a)). Of a total of 2,567 families in the SCOP classification, interface structures are available for 1,946 of them.

Visually, it is obvious that the distribution of partner structural similarities is non-uniform (Figure 2.2(b)). To investigate this distribution further, we compared the observed distribution of partner structural similarities to a random model in which all SCOP fam-

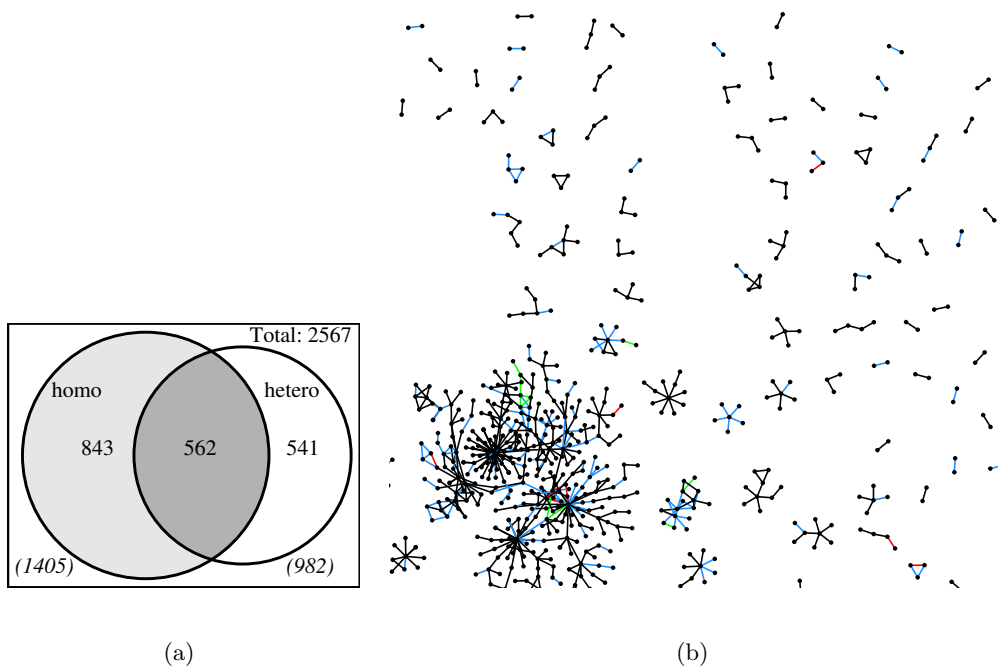
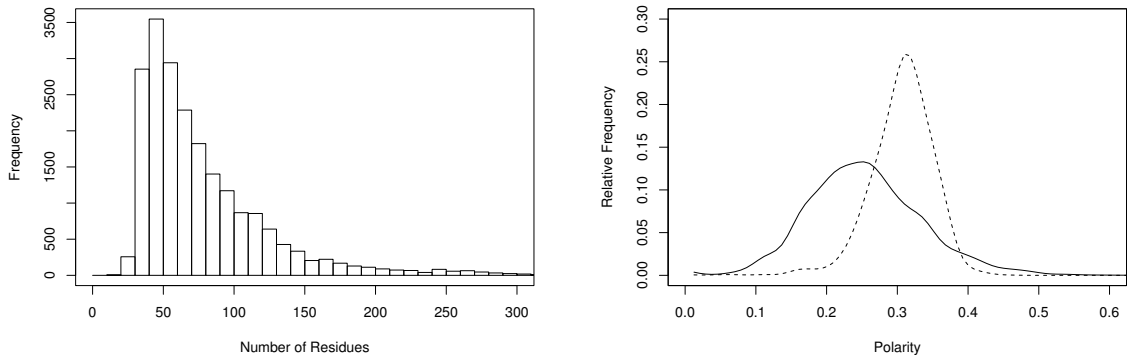


Figure 2.2: Interactions in SCOP space. (a) Venn diagram of interaction coverage in SCOP space. Regular font represents numbers of SCOP families. Italics represents numbers of interfaces (*eg*, 1405 homo-family SCOP family pairs). (b) Partial SCOP domain family interactome. Nodes represent SCOP families. Edges represent structurally observed interfaces. Edge color represents the SCOP similarity of the interacting nodes (Red = same superfamily, Green = same fold, Blue = same class, Black = no similarity). Only interfaces between different SCOP families are shown. Graph layout by LGL [3].

ilies interact with all other SCOP families. We found that interactions between domains with similarities only at the superfamily level are overrepresented (~ 5 -fold). Interactions between domains from the same fold are almost twice as abundant as expected from the random model. Interactions between domains from the same SCOP class are approximately the same as expected. The interfaces between structurally dissimilar domains are underrepresented (~ 2 -fold). In summary, the structures of interacting domain pairs currently available are weighted towards partners with the most structural similarity along the SCOP hierarchy. However, it is difficult to conclude from this observation that actual protein interaction networks behave in this manner, as the observed preferences likely reflect both an actual non-uniform distribution of structural similarity between interacting partners and sampling bias in the PDB.

Interfaces are observed to be mostly continuous in structure, but very segmented in sequence. On average, ~ 78 residues are presented to the interface on ~ 34 secondary structure elements, or ~ 23 continuous sequence segments (Figure 2.3(a), Supplementary Figure 2.6(a), 2.6(b)). However, each interface usually involves only 2 structurally continuous patches, one contributed by each binding site (Supplementary Figure 2.6(c)). As expected, the sequence discontinuity is directly proportional to the buried surface area of the interface ($r^2 = 0.71$, Supplementary Figure 2.6(d)).

An example of the physicochemical properties that can be analyzed using PIBASE is the polarity of interfaces (Figure 2.3(b)). The polarity of each interface was defined by the fraction of the buried solvent accessible surface area that was contributed by polar atoms (N, O). The interface polarity exhibits a broad distribution with a mean of $\sim 25\%$. This



(a) Number of residues at the interface.

(b) Interface *vs* Whole Domain Surface Polarity.

Figure 2.3: PIBASE Interface Property Distributions. All the plots are calculated from the non-redundant set of interfaces. (a) Number of residues at the interface. Maximum 862 residues (not shown). (b) Interface polarity (solid) compared to whole domain surface polarity (dashed). The polarity of an interface is defined by the fraction of buried surface area contributed by polar atoms (N, O). Similarly, the polarity of the whole domain surface is defined by the fraction of solvent accessible surface area contributed by polar atoms. The whole domain surface polarities were calculated from all SCOP domains.

distribution is slightly more hydrophobic than the whole domain surface, which exhibits a narrower distribution with a mean of $\sim 30\%$. The polarity of the whole domain surface was similarly defined, as the fraction of the solvent accessible surface area contributed by polar atoms.

2.3 Discussion

We presented a comprehensive database of structurally characterized protein complexes. We first described its construction (Figure 2.1), followed by its content (Tables 2.1, Figure 2.2). From domain topology to secondary structure topologies at individual interfaces, we presented groupings at size scales from the entire complex to individual interfaces.

We also described the distributions of a subset of the interface properties stored in PIBASE (Figure 2.3).

Several collections of protein chain and domain interfaces have been recently reported [87, 98, 108, 160, 224]. A SCOP domain family interactome was published that supplemented SCOP interfaces extracted from the PDB with those observed in yeast protein interaction data [160]. This resource allowed the proposal of possible evolutionary reasons for the observed repertoire of family-family intermolecular and intra-molecular interactions. More recently, a collection has been created of non-redundant high-resolution structures of protein chain pairs extracted from the PDB [108]. The interfaces were clustered using geometric hashing [152], a sequence order-independent structural superposition algorithm, which allowed the detection of conserved interface architectures across different fold types. The datasets reported vary widely in their size and breadth of descriptors, as expected from the different types of analysis they were designed for.

The main goals in developing PIBASE have been completeness of its domain interface coverage as well as diversity of the descriptors calculated at various scales. Though it contains a comprehensive set of interfaces, filters can easily be applied to focus on a specific type of interface or on those with a given minimum experimental resolution. The explicit topological clustering, previously developed for fold classification [241], is unique in its application to protein complexes.

The completeness of PIBASE makes it suited for investigations into the structure of protein interactions, as well as for benchmarking methods such as protein-protein docking. To illustrate its utility as a general purpose bioinformatics resource, we list here several

current applications in our group.

The interfaces stored in PIBASE have been used as templates for the prediction of protein interaction partners [167]. Candidate interaction partners are generated by detecting pairs of proteins from the same genome that contain domains for which an interface has been observed. These candidate interactions are then assessed by building comparative models of the individual proteins and scoring their putative interface using a statistical potential that captures residue type contact preferences at interfaces. This method predicts not only interaction partners, but also binding modes. Similar schemes have been previously reported [9, 126]. The interaction predictions have been deposited in MODBASE [167].

The spatial localization of protein binding sites in PIBASE has been analyzed (Korkin, Davis, Sali, in preparation). Localization is a measure that describes the degree of overlap of the binding sites observed for a given protein domain family. The lower the localization, the more scattered the distribution of binding sites. A range of localization values are observed for domain families. Many families exhibit a higher localization than expected by random (*eg*, obligate homo-dimeric enzymes such as alkaline phosphatase), while others exhibit a lower localization than expected by random (*eg*, highly divergent families such as C-type lectins).

The binding sites stored in PIBASE are also used by LS-SNP, a large-scale structural annotation of human single nucleotide polymorphisms (SNPs) [103]. This analysis combines multiple types of sequence and structure information, including protein binding sites, to predict whether an observed SNP is functionally deleterious.

Lastly, PIBASE has been integrated into an automated structure annotation sys-

tem in the DBAli [137] and MODBASE [167] databases. As structural genomics efforts are rapidly determining protein structures, it becomes important to annotate them using automated methods which leverage existing knowledge. For a given input protein structure, a structural alignment program MAMMOTH [157] is used to find similarities to known protein structures, and the SALIGN module of MODELLER (Madhusudhan, Marti-Renom, Eswar, Sali, in preparation) is applied to prepare multiple alignments of similar protein structures. The query protein structure can then inherit numerous properties from the similar characterized structures, including ligand binding sites from LIGBASE [213], and binding partners from PIBASE.

The modular and relational design of PIBASE allows easy cross-referencing to other databases of protein structure, sequence, and function. Work is currently underway to cross-reference binary protein interaction databases such as BIND [18], using MODBASE structural annotation of the interacting proteins [167]. Through further integration with the plethora of high quality databases, PIBASE will become a valuable resource for the structural biology community.

Acknowledgements

We would like to thank members of the Sali laboratory for valuable comments and suggestions, in particular Frank Alber, Maya Topf, Damien Devos, MS Madhusudhan, Mike F. Kim, Dmitri Korkin, Eswar Narayanan, and Ursula Pieper. We acknowledge funding by NSF (EIA-0325004), as well as computer hardware gifts from Sun, Intel, and IBM. FPD acknowledges support from a Howard Hughes Medical Institute predoctoral fellowship.

2.4 Supplementary Information: Hamann Distance Function

Bit vectors were compared using the Hamann distance measure, $\text{dist}_{\text{hamann}}$, a rescaled and reversed version of the traditional Hamann similarity coefficient, $\text{sim}_{\text{hamann}}$, developed for use in plant systematics [78]. To compare bit vectors X and Y of length m , the numbers of matches and mismatches for on and off states are first counted. The Hamann distance function is then computed as detailed below (Eq 2.3).

	$X_i = 1$	$X_i = 0$
$Y_i = 1$	n_a	n_b
$Y_i = 0$	n_c	n_d

$$\text{sim}_{\text{hamann}} = \frac{(n_a + n_d) - (n_b + n_c)}{m} = [-1, 1] \quad (2.2)$$

$$\text{dist}_{\text{hamann}} = 1 - \frac{(\text{sim}_{\text{hamann}} + 1)}{2} = [0, 1] \quad (2.3)$$

2.5 Supplementary Information: Topological Fingerprints

2.5.1 Complexes

A *domain connectivity graph* is generated for each structure, describing the pattern of domain-domain contacts. In this graph, a domain is a node, and a binary domain interface is an edge (Figure 2(b)). This graph captures the arrangement of the individual domains in the complex. A crude linear representation of this graph is then computed. This string is generated by labeling the nodes with their degree (*ie*, the number of nodes they are connected to). The edges are then listed using the labels of their composite nodes. The edge list is sorted lexically and the resulting string is used as a crude topological fingerprint

to group the structures into topological classes (Figure 2(d)). While degeneracy exists in this representation (*ie*, two distinct topologies may have the same sorted edge list), it is useful as both a query and crude clustering term.

2.5.2 Binding Sites and Interfaces

Similar to the domain connectivity graphs, a crude linear representation of the secondary structure topology graph was generated for use as a topological fingerprint. This string is generated by labeling each node on the secondary structure topology graph with their degree and their secondary structure type. The edges are then listed using the labels of their composite nodes. The edge list is sorted lexically and the resulting string is used as a crude fingerprint to group the structures into *topological classes* (Figure 2(c), 2(d)).

2.6 Supplementary Information: Kd-trees algorithm

Inter-atomic distances were computed using an in-house ANSI C implementation, called kd-contacts, of the median kd-trees algorithm [23, 61]. The kd-trees algorithm, a commonly used computational geometry algorithm, performs nearest neighbor queries by first building a tree in $O(n \log n)$ time, and then querying it in $O(n^{1-(1/d)} + k)$ time, where n is the number of data points in the d -dimensional space, and k is the number of reported points. This approach is much faster than the naive approach of all *vs* all distance calculation ($O(n^2)$). The logarithmic scaling allows rapid calculation of contact maps even for large structures with tens of thousands of atoms, such as PQS entries of virus capsids. Briefly, the algorithm begins by building a binary tree that recursively decomposes the d -

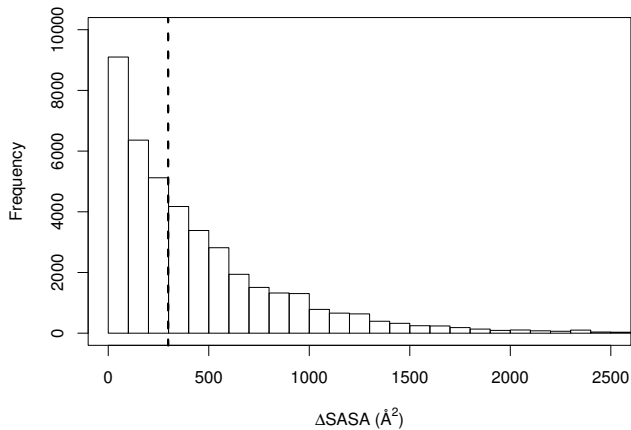
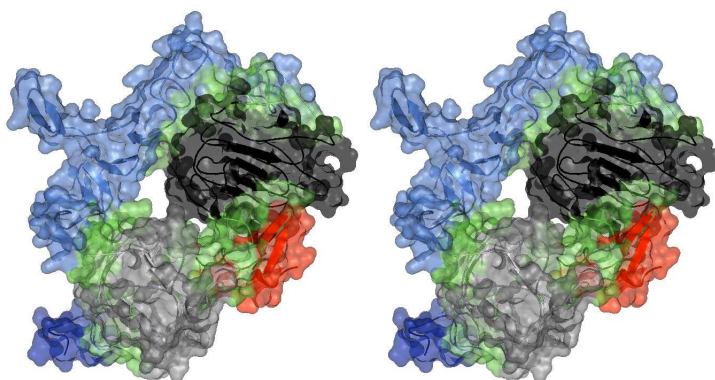


Figure 2.4: Distribution of buried solvent accessible surface area in interacting SCOP domain pairs. The distribution is calculated from the non-redundant set of interacting domain pairs (Methods). The largest interface is 7225 \AA^2 (not shown). The dashed vertical lines shows the cutoff on the solvent accessible surface area used to obtain interfaces for subsequent annotation and analysis.

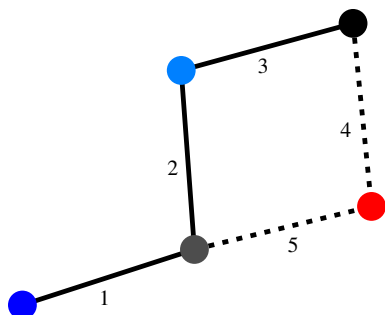
dimensional (here, $d = 3$) input space on alternate axes along the median partition. At each branch the splitting value is stored, which allows a unique bounding box of points to be associated with each node in the tree. The result is a non-uniform hypercube binning of d -dimensional space, which is adapted to the actual distribution of data points. A nearest neighbor query, given a query point and radius, is then performed by traversing only those branches that may possibly contain a nearest neighbor according to their bounding box definitions. This procedure allows rapid elimination of entire branches of the tree, leaving only those bins that may contain a nearest neighbor. Distance calculations are required only for the points in these candidate bins.

(a) <i>Complexes</i>			
1	Domain connectivity graph		
(b) <i>Domains</i>			
1	Solvent accessible surface area		
2	Secondary structure assignment		
3	Domain classification code		
(c) <i>Interfaces and Binding Sites</i>			
		Interface	Binding Site
1	Buried solvent accessible surface area (Δ SASA)	✓	
2	Buried polar solvent accessible surface area (Δ SASA _{polar})	✓	
3	Number of residues	✓	✓
4	Residue types present	✓	✓
5	Number of secondary structure elements	✓	✓
6	Secondary structure types present	✓	✓
7	Number of structural patches	✓	✓
8	Number of sequence segments	✓	✓
9	Number of residue contacts	✓	✓
10	Residue contact types	✓	✓
11	Inter-atomic contacts (distance binned) ($\leq 4.5, 5, 5.5, 6.05$ Å)	✓	✓
12	Number of secondary structure element contacts	✓	✓
13	Secondary structure type contacts	✓	✓
14	Secondary structure topology	✓	✓
15	Number of H bonds	✓	
16	Number of salt bridges	✓	
17	Number of disulfide bridges	✓	

Table 2.2: PIBASE properties. The computed properties characterize each complex at four levels: the overall complex, its constituent domains, interfaces, and binding sites. A subset of the properties are later used to remove interface redundancy and cluster the complexes, interfaces, and binding sites into topological classes.

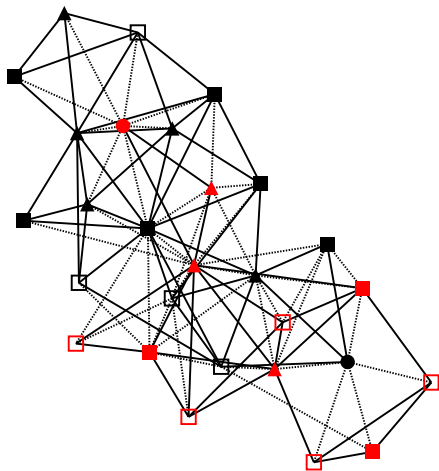


(a) Structure (PQS livo_1.mmol).



(b) Complex topology.

Figure 2.5: Topology properties calculated for each structure. Color represents the SCOP classification of the domains (blue, g.3.9.1; red, g.3.11.1; black, c.10.2.5). (a) Ribbon and surface representation of the complex of human epidermal growth factor and receptor extracellular domains (PQS entry livo_1.mmol). The interfaces are colored in green. (b) Graph representation of the domain connectivity. Solid lines represent intra-chain interfaces; Dashed lines represent inter-chain interfaces.

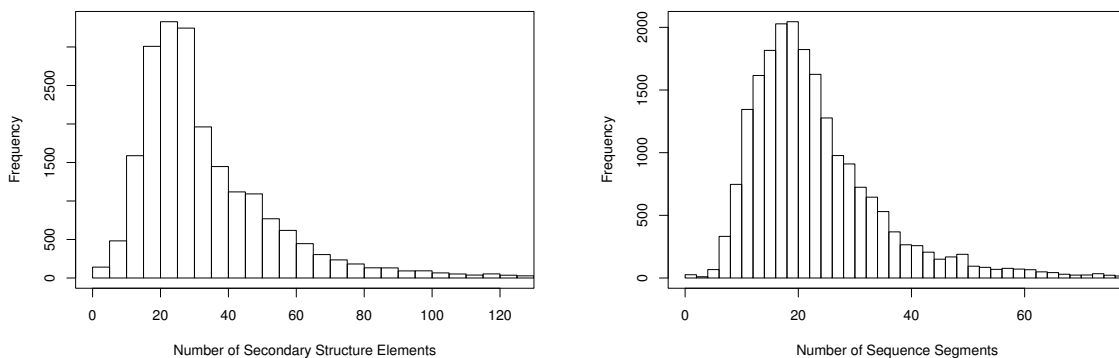


(c) Interface and binding site topologies for interface number 4, as defined in (b).

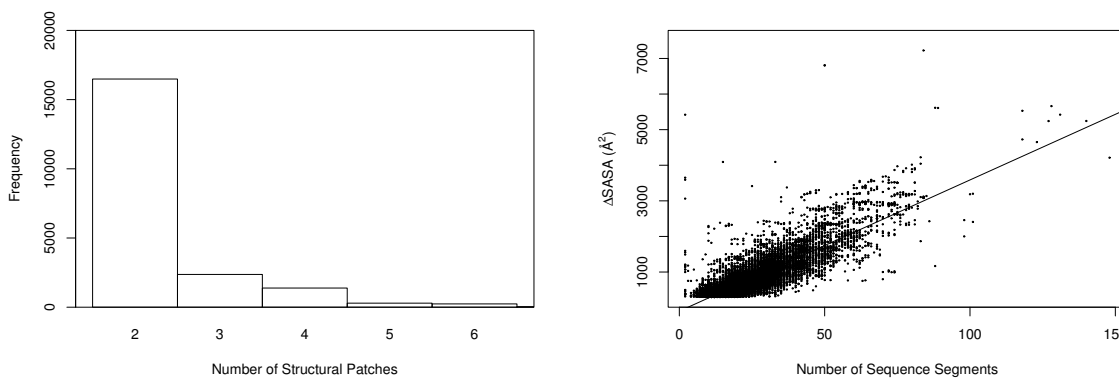
complex	3-2.3-2.3-1.2-2.2-2
interface 4	3_-2T.4B-3_.4T-2_.4T-3T.4T-4B.5B-2T.5B-2T.5H-1_.5H-1_.5H-2T.5H-3T.5H-4B.5T-1_.5T-2_.5T-3_.5T-5B.5T-5T.5_.1_.5_-2.5_.5B.5_.5T.6B-2_.6B-3T.6B-4B.6B-5B.6B-5T.7B-1T.7B-2B.7B-2T.7B-4T.7B-5T.7B-5_.7B-6B.8H-1B.8H-1B.8H-1B.8H-1T.8H-1_.8H-2B.8H-2T.8H-5T
red binding site (interface 4)	2H-2B.3_-2T.3_-2T.3_-3_.4T-2_.4T-3_.4_-4T.4_-4T.4_-4_.6B-3_.6B-4T.6B-4T.6B-4_.6B-4_.8B-2B.8B-2H.8B-2_.8B-4T.8B-4T.8B-4_.8B-4_.8B-6B

(d) Topological Fingerprints.

Figure 2.5: Topology properties calculated for each structure (cont.). Color represents the SCOP classification of the domains (blue, g.3.9.1; red, g.3.11.1; black, c.10.2.5). (c) Graph representation of the binding site and interface topology for interface 4, as defined in (b). The interface topology is defined by the subgraph containing only dashed edges. The topologies of the two interacting binding sites are defined by the two disconnected subgraphs containing only solid edges. Node shapes represent secondary structure type (triangle is β sheet, circle is α helix, filled box is loop, and open box is unassigned). (d) The topological fingerprints are listed for the overall complex, interface number 4, as defined in (b), and one of its corresponding binding sites. The characters in the interface and binding site fingerprints represent secondary structure type (B is β sheet, H is α helix, T is Loop, and _ = unassigned). Structure visualization by PyMOL (<http://pymol.sourceforge.net>); Graph layout by LGL [3].



(a) Number of secondary structure elements at the interface. (b) Number of sequence segments at the interface.



(c) Number of structural patches at the interface. (d) Number of Sequence Segments *vs* Buried Surface Area.

Figure 2.6: PIBASE Interface Property Distributions. All the plots are calculated from the non-redundant set of interfaces. (a) Number of secondary structure elements at the interface. Maximum 253 elements (not shown). (b) Number of continuous sequence segments at the interface. Maximum 148 segments (not shown). (c) Number of structural patches at the interface. Maximum 43 structural patches (not shown). (d) Number of sequence segments *vs* buried surface area ($r^2 = 0.71$).

Chapter 3

Protein Complex Compositions

Predicted by Structural Similarity

This chapter has been previously published ([40])

Abstract

Proteins function through interactions with other molecules. Thus, the network of physical interactions among proteins is of great interest to both experimental and computational biologists. Here we present structure-based predictions of 3,387 binary and 1,234 higher order protein complexes in *Saccharomyces cerevisiae* involving 924 and 195 proteins, respectively. To generate candidate complexes, comparative models of individual proteins were built and combined together using complexes of known structure as templates. These candidate complexes were then assessed using a statistical potential, derived from binary domain interfaces in PIBASE (<http://salilab.org/pibase>). The statistical potential discriminated a benchmark set of 100 interface structures from a set of sequence-randomized negative examples with a false positive rate of 3% and a true positive rate of 97%. Moreover, the predicted complexes were also filtered using functional annotation and sub-cellular localization data. The ability of the method to select the correct binding mode among alternatives is demonstrated for three camelid VHH domain — porcine α -amylase interactions. We also highlight the prediction of co-complexed domain superfamilies that are not present in template complexes. Through integration with MODBASE, the application of the method to proteomes that are less well characterized than that of *S. cerevisiae* will contribute to expansion of the structural and functional coverage of protein interaction space. The predicted complexes are deposited in MODBASE (<http://salilab.org/modbase>).

3.1 Introduction

Recent developments in high-throughput screening have generated large data sets identifying protein complexes. The *Saccharomyces cerevisiae* proteome has been especially well characterized through yeast-two-hybrid (Y2H) [93, 225] and tandem affinity purification (TAP) experiments [65, 66, 88]. Experimentally observed interactions, resulting from both high-throughput and traditional low-throughput methodologies, are deposited in databases such as the Biomolecular Interaction Network Database (BIND, 18) and the Database of Interacting Proteins (DIP, 192).

Concomitant with these experimental advances, a spate of computational techniques to predict protein-protein interactions have also been developed. Several approaches based on protein sequence, structure, function, and genomic features have been described [191]. In an effort to reduce the prediction errors, several methods integrate multiple types of experimentally determined information and theoretical considerations [96, 121, 127].

Structure-based methods have been developed for the prediction of binary protein interactions. InterPreTS [9] uses a statistical potential derived from known hetero-dimer structures and MULTIPROSPECTOR [125] relies on threading to score pairs of proteins that are similar to binary interactions of known structure. In addition to predicting new interactions, structure-based methods can also annotate interactions that have been previously observed experimentally. A recent study used computational methods in conjunction with experimentally determined complex compositions and electron density maps from negative-stain electron cryo-microscopy to generate structural models of yeast complexes [12]. In a similar vein, structural knowledge has been used to predict the domains that are

most likely to mediate binary protein interactions [153].

Here, we describe predictions of proteins that form complexes in *S. cerevisiae* based on similarity to complexes whose atomic structures have been solved experimentally. First, comparative models of conceivable complexes are built and then assessed by a specialized statistical potential. The high-confidence interactions can be additionally filtered by examining orthogonal sources of information including sub-cellular localization and functional annotation.

The current study is unique primarily in its prediction of structural models for higher-order complexes as well as homomeric complexes. Computational methods have been developed to infer higher-order complexes from binary protein interaction networks [17, 207], but they do not explicitly use structural knowledge. Previous studies have also focused primarily on the prediction of heterodimers, though homodimerization is biologically prevalent and functionally significant [133]. We show that the multiple structure-based assessment steps, from the initial fold assignment, to the interaction prediction, enables our method to achieve a higher coverage, and presumably accuracy, than methods based solely on sequence similarity.

We begin by describing the approach and benchmarking the method. Predictions are then presented for proteins in *S. cerevisiae* and validated against experimentally observed complexes. We highlight the performance of the protocol in the selection of the correct binding mode when multiple template interface structures are available and discuss newly predicted co-complexed superfamilies. Finally, we conclude with a brief discussion of potential applications of the method in light of the ultimate goal of full structural coverage

of interaction space.

3.2 Methods

3.2.1 Prediction Algorithm

Candidate complexes are first generated, then assessed, and finally filtered by orthogonal biological information (Fig. 3.1(a)).

Candidate Complex Generation

Pairs of *S. cerevisiae* proteins were identified as potential interaction partners if they were assigned SCOP domains belonging to superfamilies for which an interaction structure exists in PIBASE (Fig. 3.1(b)) [39]. In some superfamilies, such as the ARM superfamily (SCOP a.118.1), the lengths of the member domains vary widely. Because alignments between structures of different lengths are difficult, a threshold was placed on the relative sizes of the target and template domains - the shorter of the two domains must be at least 60% of the length of the longer domain. In addition, the target domains were required to be aligned with the template domains in sufficient number of positions such that the corresponding template residues formed at least 50 % of the template interface contacts.

Protein Data Bank (PDB) [24] structures that contained more than two domains were used as templates for the prediction of higher-order complexes with more than two proteins. Target domains that were assessed to interact through the interface modes in a given PDB structure were listed as candidate members of a complex. Each complex was

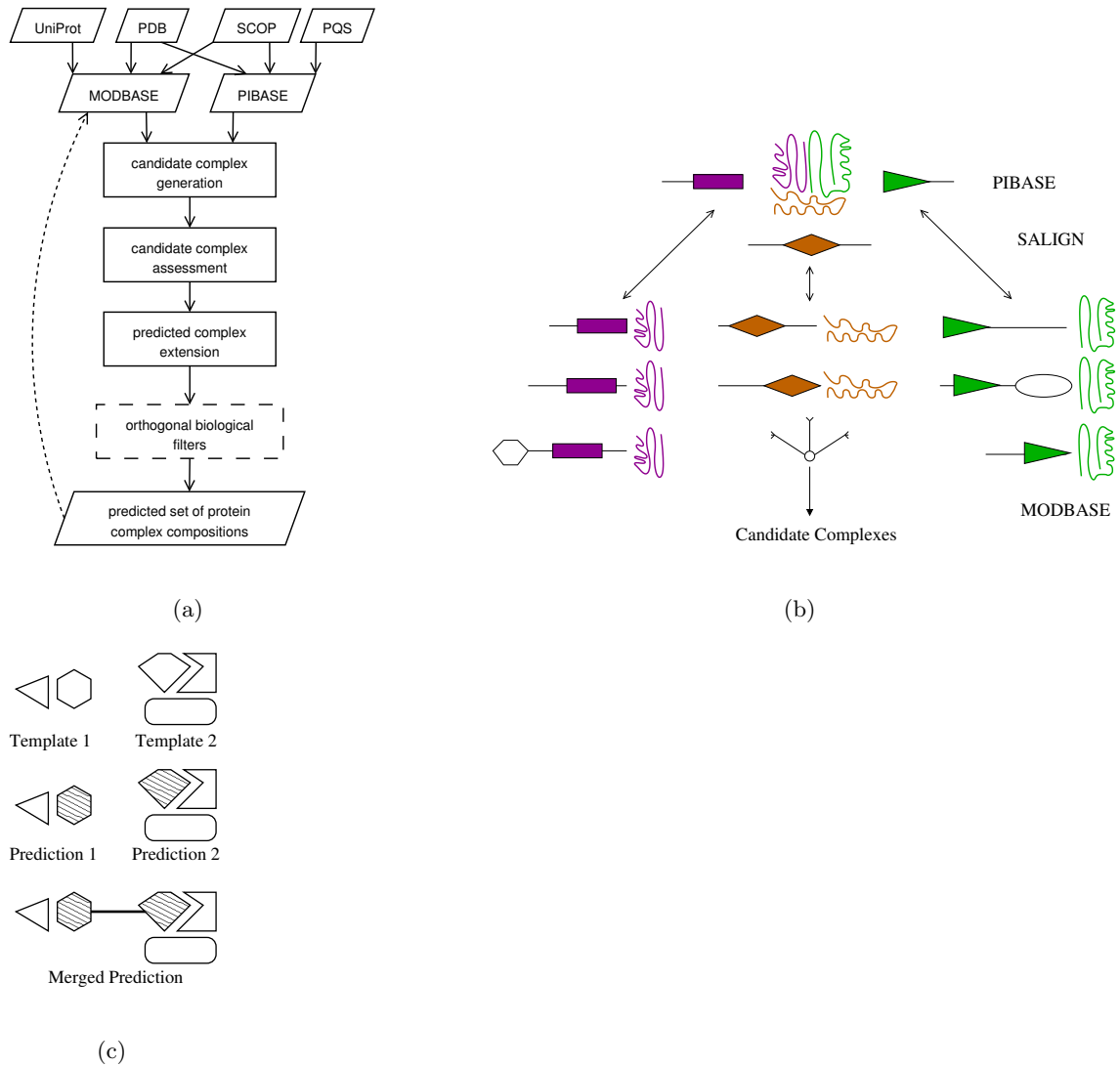


Figure 3.1: Prediction Logic Overview. (a) Prediction Flowchart. Groups of protein sequences modeled with SCOP domains observed to form a complex in PIBASE are listed as candidate complexes. These candidate complexes are then assessed by a statistical potential. Interactions that score above a Z-score threshold are filtered using sub-cellular localization and functional annotation. The resultant predictions are deposited in MODBASE. (b) Candidate Complex Generation. Comparative models of target domains are structurally aligned to templates of known structure in PIBASE using the SALIGN module of MODELLER. Putative interface residues are identified from the alignment. (c) Predicted complexes are merged if they contain different domains of a single target protein.

then scored with the worst of the Z-scores for the interacting domain pairs it contained, as described below. This was done to provide a conservative estimate of complex quality based on the lowest scoring constituent interface. Predicted complexes were merged if they contained different domains of a single target protein. In effect, the covalent link between the domains served as a ‘bridge’ between predicted complexes that were based on different templates (Fig. 3.1(c)).

Assessment of Candidate Complexes

Each candidate interaction pair was scored by assessing the agreement between the target sequences and the template interface structure using a statistical potential derived from binary interface structures in PIBASE.

First, residue contacts across the interface were calculated for the template interface and grouped into classes based on the main chain or side chain participation of each residue. Next, the MODBASE models of each candidate interaction partner were structurally aligned against the corresponding domains in the template interface using the SALIGN module of MODELLER [188]. Finally, the residue correspondences defined by the alignments were used to score the candidate partner sequences against the template interface contacts using the statistical potential, as described below.

A Z-score was calculated to assess the significance of the raw statistical potential score, by consideration of the mean and standard deviation of the statistical potential scores for 1,000 sequences where all amino acids in the target domain sequences were shuffled. Sequence randomization has been previously shown to perform comparably to a more physical

model involving structural sampling in the context of fold assessment [141].

Orthogonal Biological Information

Orthogonal biological support for each predicted complex was provided by sub-cellular localization and gene ontology functional annotation of their components, obtained from the YeastGFP [68] and SGD databases [46], respectively. The number of shared localization and function terms were computed for both experimental and predicted complexes. If all pairs of proteins in a complex shared at least one function or localization term, the complex was flagged as co-functioning or co-localized, respectively.

3.2.2 Construction of Statistical Potentials

A series of statistical potentials was built using the binary domain interfaces in PIBASE extracted from structures at or above 2.5 Å resolution, randomly excluding 100 benchmark interfaces. Twenty-four statistical potentials were built using different values of three parameters: the contacting atom types (main chain - main chain, main chain - side chain, side chain - side chain, or all), the relative location of the contacting residues (inter- or intra- domain), and the distance threshold for contact participation (4, 6, or 8 Å):

$$g_{ij} = \frac{\sum_{p=1}^N \Delta n_{ij}^{(p)}(R_o) \text{cifa}_{ci,cj} n_p}{\sum_{p=1}^N n_{ij}^{(p)} \max(\text{cifa}_{i,j})} \quad (3.1)$$

$$\text{cifa}_{x,y} = \min \left(\frac{\text{interacting atoms}_x}{\text{atoms}_x}, \frac{\text{interacting atoms}_y}{\text{atoms}_y} \right)$$

$$n_{ij}^{(p)} = \begin{cases} n_i^{(p)} n_j^{(p)} & \text{intra-domain potential,} \\ n_i^{(d1)} n_j^{(d2)} + n_i^{(d2)} n_j^{(d1)} & \text{inter-domain potential.} \end{cases}$$

$$w_{ij} = -\ln \left[\frac{g_{ij}}{\frac{1}{400} \sum_{k=1}^{20} \sum_{l=1}^{20} g_{kl}} \right] \quad (3.2)$$

Each of the $\Delta n_{ij}^{(p)}(R_o)$ residue pairs of type i and j in protein p that occurred within the distance threshold R_o was weighted by $cifa$, the minimum of the fraction of total atoms (of the type specified in the potential) in each residue that fell within the distance threshold (Eqn. 3.1), and n_p , the number of residues in the protein. This count for each residue type pair was normalized by $n_{ij}^{(p)}$, the total number of possible contacts of that type in each protein, weighted by $\max(cifa_{ij})$. In the case of the inter-domain potential, $n_{ij}^{(p)}$ was computed by taking into account the occurrence of each residue type in each domain individually. Finally, the score for each residue type pair was normalized by the sum of the scores observed for all residue type pairs (Eqn. 3.2).

3.2.3 Benchmarking of Statistical Potentials

Performance on the benchmark set of 100 randomly selected interface structures, that were excluded during construction of the potentials, was used to compare the 24 statistical potentials. Eighty-four of these benchmark interfaces occur between domains from the same family. This is representative of all interfaces in PIBASE, 8,877 (15.5%) of which are between domains from different families, and 48,257 (84.5%) of which are between domains from the same family. The sequences of the benchmark interfaces were scored against their structures and a Z-score was calculated, as described above. Receiver-

operator curves (ROC) were built to describe the observed false-positive and true-positive rates at different Z-score thresholds. The ROC curves were then integrated to calculate the area under the curve (AUC). The AUC represents the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, with 0.5 corresponding to a random prediction, and 1 to a perfect classifier [56].

To investigate the effect of variation in the benchmark set on each of the ROC curves, 20 jack-knife trials were performed where 20 randomly selected structures were removed and the ROC curves recomputed using the remaining 80 structures. Standard deviations of the areas under the resulting ROC curves were then calculated.

3.2.4 Validation of complex prediction

The predicted interactions were validated in two ways. First, the predicted *S. cerevisiae* complexes were compared to the experimentally determined complexes in the BIND database [18] and those recently reported by Gavin *et al.*, referred to as Cellzome [66]. The binary interactions were compared by counting the overlap of the predictions with the interactions in the BIND and Cellzome sets. The Cellzome set consisted of pairs of proteins that were deemed highly reliable in forming partnerships based on their computed ‘socio-affinity’ score [66].

Second, the higher order complexes were compared between the predicted and experimental sets by counting how many of the predicted complexes were equivalent to, or were subcomplexes of, experimentally determined complexes. Since the predictions are based on known structures, the sizes of the predicted complexes are far smaller than those

obtained by biochemical methods such as tandem affinity purification methods. For this reason, we elected not to use a metric that explicitly penalizes size differences (*eg*, the metric defined in 17).

3.2.5 Binding Mode Selection

The ability of the potential to select the proper binding mode when multiple template interfaces of different orientation are available was assessed. The test cases used were the structures of camelid VHH domains AMB7, AMD10, and AMD9 bound to porcine pancreatic α -amylase (PPA) (PDB codes 1kxt, 1kxv, and 1kxq, respectively). All three modes were evaluated for each VHH-PPA complex using the interface statistical potential.

3.2.6 Data Sources

The prediction algorithm uses three types of data: (i) target protein sequences among which complexes are to be predicted, (ii) structures of protein complexes to be used as templates, and (iii) a list of the locations and types of structural domains in the target and template proteins (Fig. 3.1(a)).

Target Proteins

S. cerevisiae protein sequences were obtained from MODBASE, a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure [168]. The models were calculated by MODPIPE [52], an automated modeling pipeline that relies on MODELLER for fold assignment, sequence-structure alignment, model building, and model assessment [188]. 6,600

S. cerevisiae proteins were processed, resulting in 9,464 models for 3,440 sequences. 2,659 sequences had at least one reliable model (5,387 reliable models in total). A model is considered reliable when the model score, derived from statistical potentials, is higher than a cutoff of 0.7 [141]. A reliable model has a greater than 95% probability of having at least 30% of C α atoms within 3.5 Å of their correct positions. 3,376 sequences had at least one reliable fold assignment (8,935 reliable folds in total). A fold assignment is considered reliable when the model is based on a PSI-BLAST match to a template with an e-value smaller than 0.0001.

Structural Domain Annotation

The domain definitions for PDB structures were obtained from the SCOP database (ver 1.69) that classifies each domain using a four level hierarchy, class, fold, superfamily, and family [148]. The location and types of domains in the target protein sequences were then predicted using the SCOP annotation of their MODBASE templates, as follows. Domain boundaries were first assigned based on the MODBASE alignment of each target protein to its structural template. Each target domain was required to have at least 70% of the residues in its template domain to receive the domain assignment. Next, if the target domain had greater than 30% sequence identity to the template domain and the MODBASE structural model was assessed to be reliable, the target domain received the template's SCOP classification at the family level. If the sequence identity was less than 30% and a reliable model was built or if the sequence identity was greater than 30% but MODBASE deemed only a reliable fold assignment, the superfamily was assigned. The

remaining domains received the template domain's SCOP classification at the fold level, and were not used in the interaction prediction.

For those target proteins for which multiple models were available in MODBASE, a tiling procedure combined the domain assignments for each model into a non-overlapping set of domain boundaries that maximized the coverage length and classification detail in the SCOP hierarchy.

Template Complexes

Structures of template complexes were retrieved from PIBASE, a comprehensive relational database of structurally defined protein interfaces [39]. It currently includes 209,961 structures of interactions between 2,613 SCOP domain families. The ASTEROIDS component of the SCOP ASTRAL compendium was used to cluster the interfaces, reducing the computational expense of the predictions [32]. The ASTEROIDS alignments, available for SCOP classes a-g, were used together with the interface contacts stored in PIBASE to cluster all interface structures that shared pairs of SCOP families. When at least 75% of the pairwise residue contacts in one interface also occurred between residues that were aligned in another interface, the two interfaces were merged into a single cluster. The clustering reduced the 79,428 domain interfaces between pairs of domains in the SCOP classes a-g to 21,791 representative interfaces. These interfaces were filtered using a threshold of at least 1,000 interatomic contacts resulting in a set of interfaces of significant size. Thresholds similar to this, which roughly corresponds to a buried surface area of 400 \AA^2 , have been previously used to filter crystallographic artifacts from biologically relevant interfaces[83].

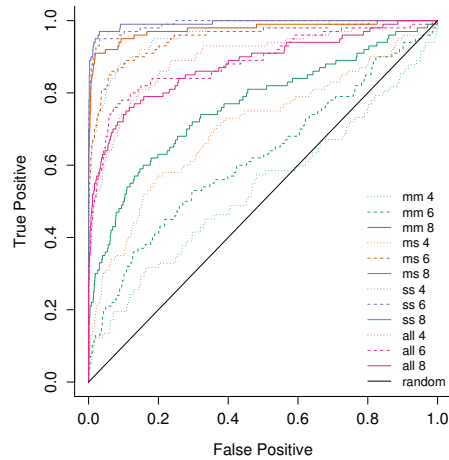


Figure 3.2: Assessment of statistical potentials. Receiver operator curves (ROC) are shown for the inter-domain potential performance on the benchmark set of complexes.

The final set of template binary interfaces contained 5,275 structures, including both inter-molecular and intramolecular interfaces.

3.2.7 Technology

The prediction system was implemented as a Perl module and an integrated set of Perl scripts, except for the inter-atomic contacts calculator written in ANSI C [39]. The SALIGN module of MODELLER [188] was used to generate model template alignments. The Perl DBI interface was used to access the MODBASE and PIBASE MySQL databases (<http://www.mysql.com>). The calculations were done in a parallel fashion on 50 3.0 GHz Pentium IV processors, taking 20 hours for the yeast genome. The predictions are accessible *via* the MODBASE web interface (<http://salilab.org/modbase>).

3.3 Results

3.3.1 Benchmark

The statistical potentials were tested using the benchmark set of 100 complexes, and their performance compared using receiver operator curves (ROC) (Methods). The highest power of discriminating between the native and non-native interfaces was achieved by the statistical potential built from side chain - side chain contacts across the interfaces at a threshold of 8 Å, corresponding to the extent of the first residue shell (Fig. 3.2). The ROC curve for this potential had an area under the curve (AUC) of 0.993, and at the optimal Z-score threshold of -1.7 had true positive and false positive rates of 97% and 3%, respectively. Clear performance trends were observed for the parameters sampled in the potential construction. The general trend of increased performance at the 8 Å threshold over the lower thresholds is likely due to a more complete description of interactions within the first residue shell. The inter-domain potential always performed better than the corresponding intra-domain potential, when all other parameters were equivalent (data not shown). The side chain - side chain (SS) potential performed better than the corresponding main chain - side chain (MS) potential, which in turn performed better than the corresponding main chain - main chain (MM) potential. At 6 Å and 8 Å, the all atom-type potential performed better than only the MM potential. At 4 Å, the all atom-type potential performed better than both MS and MM potentials. The range of performances, generated by varying the other parameters (*ie*, atom type, inter- or intra- domain), was widest at the 4 Å distance threshold and least at 8 Å.

Jack-knife trials were performed to determine the effect of variation in the benchmark set on the ROC curves (Methods). The AUC of the jack-knifed ROC curves exhibited narrow distributions, with the lowest standard deviation (0.002) achieved by the inter-domain SS potential at 8 Å, and the highest (0.02) achieved by the intra-domain MM potential at 4 Å. This suggests the ROC analysis is robust to variations in the benchmark set.

Here, the potentials were assessed using a benchmark set of native interface structures. In the predictive setting, the absolute performance of each potential will likely be diminished due to errors in the comparative models. However, the relative performance of the different formulations, as captured by the ROC curves, remains a valid guide for selection of the potential to use in the predictions.

3.3.2 Predictions

The best statistical potential, as determined above, was then used to assess candidate interactions between *S. cerevisiae* proteins. 12,867 binary interactions that scored at or below a Z-score threshold of -1.7 were predicted between 1,390 *S. cerevisiae* proteins (Fig. 3.3(a)). Next, the co-function and co-localization filters were separately applied, reducing the original 12,867 interactions to 6,808 and 4,606, respectively. The combined co-localization and co-function filter resulted in 3,387 predictions. 12,702 higher-order complexes were also predicted at a Z-score threshold of -1.7 between 589 proteins. Similar to the binary predictions, the orthogonal filters reduced this number to 1,234 complexes between 195 proteins.

The predictions spanned the entire spectrum of target-template sequence similarity (Fig. 3.3(b)). This distribution reflects both the comparative modeling procedure used to build models of the individual proteins and the procedure used to identify potential interaction templates. The mean target-template sequence identity of the reliable models built for *S. cerevisiae* proteins is 31%. Domains from different families within the same superfamily, the SCOP level used to identify potential interaction templates, often share less than 30% sequence identity. Both of these factors influence the distribution of target-template identities observed for the predicted interactions.

The fractions of predicted binary interactions that passed the co-function (53%), co-localization (36%), and both co-function and co-localization (26%) filters were similar to the fractions for BIND interactions (39%, 34%, and 22%, respectively). The Cellzome set more readily passed these filters (65%, 60%, and 46%, respectively).

3.3.3 Validation

The predictions were then compared with known experimental interactions, as deposited in the BIND database. 270 of the 3,387 predicted binary interactions that passed the combined co-localization and co-function filter overlapped with known binary interactions. 8 of the 1,234 predicted higher-order complexes were also found as subcomplexes of experimental complexes.

The enrichment of the unfiltered predictions with known binary interactions begins to plateau at 0.2 around a Z-score threshold of -3.5 , with an enrichment value of 0.03 at the Z-score of -1.7 (Fig. 3.4(a)). The predictions that passed the separate localization

and function filters both reached a peak of 0.28 at a Z-score of -3.6 . Both filters produced enrichment values of 0.06 at the Z-score threshold of -1.7 . The enrichment of the predictions that passed the combined co-localization and co-function filter exhibited a higher peak of 0.36 at the Z-score of -3.6 . At the Z-score threshold of -1.7 , the combined filter produced an enrichment of 0.08, a more than two-fold increase compared to the unfiltered predictions.

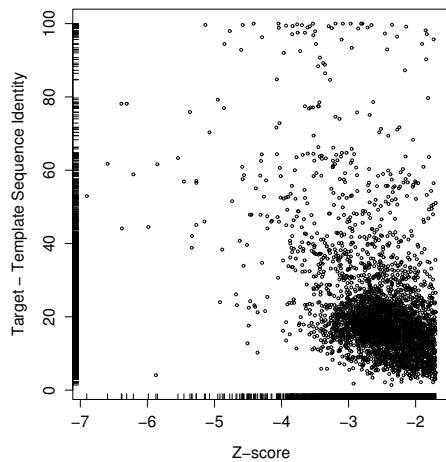
3.3.4 Comparison to other computational methods

The performance of the method in predicting binary interactions is comparable to similar structure-based methods that have been previously applied to *S. cerevisiae* on a genomic scale. Here, an overlap of 270 binary interactions is observed between the set of 3,387 (8%) predictions and 19,424 (1.4%) experimentally observed binary interactions. 374 of 7,321 (5 %) interactions predicted by threading occurred in a set of 78,930 (0.4%) experimentally determined yeast interactions [126]. An overlap of 59 predicted interactions with an experimental set of 2,590 (2.3%) interactions was obtained by interface model assessment [9].

To compare it directly to a method that does not use structural assessment, PSI-BLAST [195] was used to predict binary interactions by detecting similarities between *S. cerevisiae* proteins and the template complexes. An overlap of 929 binary interactions was observed between the set of 36,790 (2.5%) predictions and the 19,424 (4.8%) experimentally observed binary interactions.

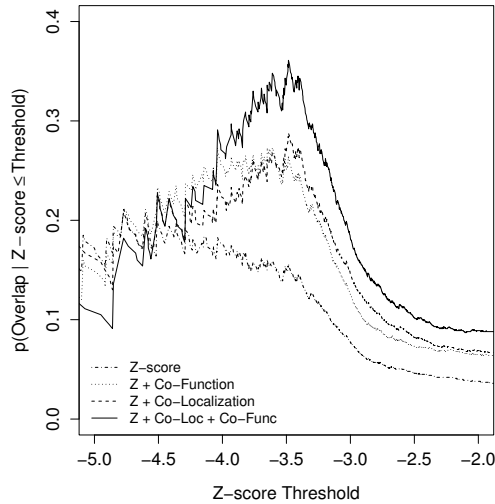
	Protein Interactions	Proteins	Domain Interfaces	Domains
<i>Input</i>				
MODBASE models	-	3,440	-	5,219
Template Complexes	-	-	5,275	9,314
<i>Binary Interactions</i>				
Z-score ≤ -1.7	12,867 (5.1%)	1,390	13,773	1,727
Z + Co-Function	6,808 (9.6%)	1,152	7,364	1,389
Z + Co-Localization	4,606 (14.1%)	1,021	5,030	1,255
Z + Co-Loc + Co-Func	3,387 (19.2%)	924	3,738	1,112
<i>Higher-Order Complexes</i>				
Z-score ≤ -1.7	12,702	589		
Z + Co-Function	3,544	332		
Z + Co-Localization	2,189	280		
Z + Co-Loc + Co-Func	1,234	195		

(a)



(b)

Figure 3.3: *S. cerevisiae* predictions. (a) Predictions of binary and higher-order complexes filtered by sub-cellular localization and annotated function. The homomeric fraction of interactions is listed in parenthesis. (b) Average sequence identity of predicted interaction partners to template interacting domains *vs* Z-score. The predictions shown were scored with Z-score ≤ -1.7 , and passed the combined co-localization and co-function filter.

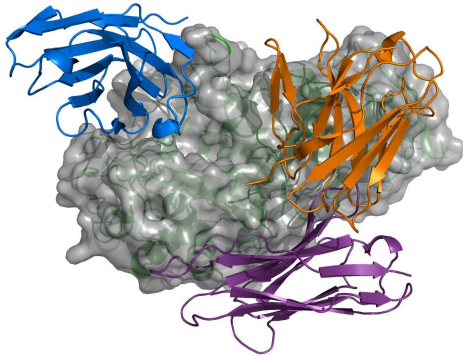


(a)

	Predicted	Experimental Overlap		
		All	BIND	Cellzome
<i>Binary Interactions</i>				
experimental		19,424	13,191	6,942
Z-score ≤ -1.7	12,867	409	324	151
Z + Co-Function	6,808	390	311	145
Z + Co-Localization	4,606	278	220	102
Z + Co-Loc + Co-Func	3,387	270	217	97
<i>Higher-Order complexes</i>				
experimental		783	296	491
Z-score	12,702	66	54	35
Z + Co-Function	3,544	51	45	28
Z + Co-Localization	2,189	14	7	10
Z + Co-Loc + Co-Func	1,234	8	4	7

(b)

Figure 3.4: Experimental overlap of *S. cerevisiae* predictions. (a) The probability of finding an experimentally observed interaction in the predicted set, as a function of the statistical potential Z-score. The unfiltered predictions are represented by dotted-dashed, the co-function filtered by dotted, the co-localization by dashed, and the combined co-localization and co-function filtered set by solid lines. The curves are only shown to a Z-score threshold of -5.0 , because of the sparseness of predictions below this level. (b) Experimental overlap of the binary and higher-order predictions filtered by sub-cellular localization and annotated function.



	AMB7 mode	AMD10 mode	AMD9 mode	K_d [nM]
AMB7	-3.27 (-2.27)	-1.19 (14.02)	-2.65 (5.00)	235
AMD10	-1.39 (12.61)	-3.40 (-4.84)	-2.36 (6.73)	25
AMD9	-2.13 (4.94)	-0.97 (15.78)	-3.60 (-9.75)	3.5

Figure 3.5: Selection among alternate binding modes. Camelid VHH domains AMB7 (orange), AMD10 (magenta), and AMD9 (blue) bind to porcine pancreatic α -amylase (PPA, grey surface) through three distinct binding modes (PDB codes 1kxt, 1kxv, and 1kxq, respectively). All three modes were evaluated for each VHH-PPA complex using the interface statistical potential. The Z-scores are presented along with the raw score in parenthesis. Dissociation constants measured by total internal reflectance (IASys) were obtained from literature [120]. Image created by PyMOL (Delano Scientific, 2002).

3.3.5 Alternate Binding Modes

The ability of the algorithm to correctly select the native binding mode when alternate templates are available was tested. The native binding mode was correctly selected for all three VHH domains interacting with porcine pancreatic α -amylase (Fig. 3.5). In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank-order as the affinity measured experimentally by total internal reflectance [120].

	Superfamily pairs	BIND or Cellzome	BIND	Cellzome
BIND or Cellzome	13,586	13,586	3,997	11,594
PDB direct	671	181	131	159
PDB co-complexed	1,555	420	143	393
Predicted co-complexed	100	43	24	35

Figure 3.6: Co-complexed domain superfamilies. The pairs of co-complexed superfamilies observed in the BIND and Cellzome complexes are compared to the direct interactions in the PDB, co-complexed pairs in the PDB, and the predicted co-complexed pairs resulting from the complex extension procedure.

3.3.6 Co-complexed domains

An extension process merged predicted complexes containing different domains of a single target protein (Fig. 3.1(c)). This process predicted 279 pairs of co-complexed SCOP domain families that were not present in the structures of template complexes. The comparison to experimental complexes was done at the superfamily level, as many of the domains in the experimental complexes were assigned domains that were classified only to this level in the SCOP hierarchy (Fig. 3.6).

3.4 Discussion

We presented a method to predict protein complex compositions by generating comparative models of candidate complexes based on sequence similarity to structurally known complexes followed by model assessment (Fig. 3.1). We applied the method to the *S. cerevisiae* proteome (Fig. 3.3) and compared the predicted complexes with experimental data (Fig. 3.4, Fig. 3.6). We further tested the method by distinguishing between multiple template binding modes (Fig. 3.5). We now discuss the observed performance and describe

the limitations of the algorithm. We close by discussing the information gained in the present study and its applications to increasing structural description of protein interactions.

3.4.1 Accuracy

Because a large set of true negative interactions is not available, only the positives, or predicted interactions, can be compared between experiment and predictions. This limitation restricts the validation of the predictions because if the Z -score threshold is loosened, maximal overlap can be achieved at the expense of the false positive rate. However, the false positive rate can not be counted with certainty, as false positives can not be distinguished from false negatives in the experimental data sets, which can be quite high [235]. Similar validation problems are encountered when testing protein ligand docking algorithms. Here, a measure related to the enrichment factor used in protein ligand docking was applied (Fig. 3.4(a)).

The overlap observed between the predicted and experimentally observed complexes is comparable to that between different experimental procedures [235]. 270 of the 3,387 predicted binary interactions and 8 of the 1,234 predicted higher-order complexes were present in the BIND or Cellzome datasets (Fig. 3.4).

This overlap is a result of several factors. First, by construction our method is restricted to protein interactions for which structural templates exist. For this reason, our method is also biased towards complexes that are stable enough to be amenable to structure determination, whereas the yeast-two-hybrid method that populates most of the high-throughput entries in BIND, is biased towards transient interactions [235]. Secondly,

many PDB entries do not contain complete domains for both partners (*eg.*, SH3 domain - peptide complexes) and were thus not considered as templates in the current prediction protocol. Finally, the challenge faced in predicting binary interactions increases combinatorially for higher-order complexes.

Errors in the predicted interactions are also a result of errors that may arise in each stage of the comparative modeling procedure, including fold assignment, alignment, and structure modeling. Comparative modeling errors vary in type and magnitude with the sequence identity between the template and target proteins[128]. At very low sequence identities, the fold type of the target sequence may be assigned erroneously. When the proper fold has been assigned, misalignments may still occur due to gaps and insertions in the target sequence. Given the correct alignment, main chain distortions may still occur due to differences in the target and template backbone structure. At the finest resolution, side chains may suffer from errors in packing. These comparative modeling errors contribute to both false positives and false negatives in the predicted interactions.

The use of sub-cellular localization data and functional annotation as filters for the predictions increased their overlap with experimental complexes, as compared to the unfiltered predictions. This finding is in agreement with previous observations that combining multiple sources of information improves the accuracy of function annotation as well as interaction prediction [96, 121, 127]. Our method easily allows for the use of additional biological filters when other types of data are available, such as synthetic gene lethality [219], co-expression [217], *etc.* This incremental addition of orthogonal information is also necessary to more accurately represent the conditions in the cellular milieu, where the propensity

of two protein structures to interact is not limited only by the physical chemistry of the interaction, but also by higher levels of biological regulation, including compartmentalization, expression, degradation, abundance, *etc.* Depending on the application, the user may decide to apply different biological filters.

3.4.2 Importance of Structure

The majority (98.6%) of the filtered binary interactions as well as the subset that overlapped with experimentally observed interactions (86.9%) were based on templates sharing less than 80% sequence identity, a threshold previously established for reliable transfer of a known interaction to a putative interaction between homologous proteins (Fig. 3.3(b)) [247]. This distribution highlights the advantage garnered by the use of structure and the importance of a structure-based assessment.

One such example is the experimentally observed interaction between LSM2 and LSM7 that was predicted here based on structural similarity to the 14-mer complex of SmAP3, an Sm-like protein from the archae *Pyrobaculum aerophilum* (PDB 1m5q). The sequence identities of LSM2 and LSM7 to SmAP3 are 23% and 2.4%, respectively. While interface templates with higher sequence identities were available (highest identities of 20.7% for LSM2 and 32.1% for LSM7 to chains G and A of PDB 1jbm, respectively), the 1m5q-based model was scored most favorably by the statistical potential. Another example of a known interaction predicted using a distantly related template interaction is that between the delta (GCD2) and beta (GCD7) subunits of the translation initiation factor eIF2B, predicted based on similarity to the structure of Ypr118w, a methylthioribose-1-phosphate

isomerase related to regulatory eIF2B subunits. The prediction was made based on sequence similarities of 16% and 15%, respectively.

For comparison, a naïve search for putative interaction partners was performed by using PSI-BLAST to detect similarities between yeast proteins and the template complexes. As expected, this approach, which is equivalent to the current method performed without the structural assessment, predicted more binary interactions that have been observed previously by experiment (929) than the structure-based method (270). However, the naïve approach likely suffers from a higher false positive rate, as can be observed in the lower enrichment of its predictions with experimentally observed interactions (2.5%) than the structure-based method (8%) (Methods).

3.4.3 Alternative Binding Modes

The ability of the algorithm to choose the correct binding mode when multiple templates are available was illustrated by evaluation of three alternative binding modes that have been structurally characterized between porcine pancreatic α -amylase and camelid VHH domains (Fig. 3.5). The algorithm successfully chose the native binding mode for all three VHH domains. In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank-order as the affinity of the interactions measured by total internal reflectance [120].

However, this example is also cautionary in that each VHH domain had one non-native mode that scored below the optimal Z-score threshold, though only the native modes produced negative raw scores (Results). In a large-scale predictive setting, if the native

binding mode was not available as a template, the VHH domain would have been predicted to interact with PPA, but through an incorrect binding mode. This example illustrates a connection between the observed performance and the underlying scoring scheme. However, a systematic analysis of alternative binding modes in protein interactions, and the ability of our method to distinguish them, remains a useful goal for the future.

3.4.4 Network Specificities

A more difficult test of the method is the prediction of specificities within interaction networks between homologous proteins. To address this problem, the method was applied to predict the specificities within the Epidermal Growth Factor Receptor (EGFR) and Tumor Necrosis Factor β (TNF β) networks of ligand receptor interactions (data not shown). In both networks the method failed to recapitulate known binding preferences. Specifically, the rank order of the Z-scores for the assessed pairs did not correlate with known binding preferences.

This error was not surprising. The randomization scheme employed in the Z-score assessment of the raw statistical potential score simulated alternative binding modes. In contrast, it was not designed or tested to determine specificities. This task is difficult as large training data sets of this type are not available.

Rather than predicting specificities, the method presented here is applicable as a first pass for genome-wide predictions of protein complexes. The resulting predictions are then suitable for a follow up with more accurate computational methods, which on their own are not feasible on a large-scale.

3.4.5 Extension of Known Co-complexed Domain Superfamilies

Large protein complexes present unique challenges to structural characterization. Direct physical interactions have been experimentally observed between domains from 671 pairs of different SCOP superfamilies (excluding homo-family interactions). Domains from 1,555 pairs of different superfamilies have been observed to co-complex in the same PDB entry. 420 of these pairs have also been observed in biochemical complexes. Through an extension process that merged predicted complexes containing different domains of a single target protein, an additional 100 pairs of superfamilies were predicted to be co-complexed (Fig. 3.1(c), Fig 3.6). 43 of these newly predicted pairs were also found in the experimental complexes. This extension procedure will be especially informative when applied to proteins from higher organisms with greater domain architecture complexity than *S. cerevisiae* [25].

3.4.6 Future Directions

We presented a tool for the prediction and assessment of the composition and structure of protein complexes. The results suggest that the algorithm may in practice be useful in conjunction with additional biological data, such as protein localization and functional annotation. Through its integration with MODBASE, the method is applicable, in an automated fashion, to all genomes with sequences that are amenable to comparative protein structure modeling. The method will be especially informative for proteomes of species that have not been characterized to the extent of *S. cerevisiae*, either because the genomes have only recently been sequenced or because the organisms are difficult to analyze experimentally.

In addition to proposing new protein complexes that have not previously been observed, the present study also enables a more rigorous, structure-based, analysis of experimental protein interaction data. For instance, the system could be used to distinguish complexes from temporally distinct interactions by assessing whether the interactions are sterically compatible or exclusive [79]. The predictions may also prove useful in guiding experiments that aim to probe the interactions, such as various site-directed mutagenesis and interaction design studies.

Comparative protein structure modeling is increasingly used to help bridge the resolution gap between electron cryo-microscopy (cryo-EM) density maps and atomic protein structures [221]. Fitting of protein and protein domain models into density maps of large assemblies is already common, but depending on the resolution, the information encoded in the map is often insufficient for an unambiguous determination of the positions and orientations of the individual proteins [55]. Models of the complexes predicted here may provide additional restraints for a more accurate fitting of proteins into large complexes studied by cryo-EM and electron cryo-tomography [12, 190].

As the number and size of experimentally determined structures of protein complexes increase, the number of complexes that can be predicted and modeled using these structures as templates increases correspondingly, expanding the structural coverage of protein interaction space [10]. In combination with other computational methods, the presented method will allow biologists to harness interaction information that has been experimentally determined for similar systems to inform their hypotheses or experiments.

Acknowledgments

We would like to thank members of the Sali laboratory for valuable comments and suggestions, in particular Maya Topf, Eswar Narayanan, and Frank Alber. We acknowledge funding by the Sandler Family Supporting Foundation, NSF (EIA-0325004), NIH (AI035707), as well as computer hardware gifts from Sun, Intel, and IBM. FPD acknowledges support from a Howard Hughes Medical Institute predoctoral fellowship.

Chapter 4

Host–Pathogen Protein Interactions Predicted by Structure

This chapter includes text that will be submitted for publication.

Abstract

Pathogens have evolved numerous strategies to invade their hosts, acquire nutrients, and evade host immune defenses, while hosts have evolved immune responses and other defenses to these foreign challenges. The vast majority of these interactions involve protein–protein recognition. Here, we present and apply a computational whole-genome protocol that generates testable predictions of potential host–pathogen protein interac-

tions. The protocol first scans the host and pathogen genomes for proteins with similarity to known protein complexes, then assesses these putative interactions, using structure if available, and finally filters the remaining interactions using biological context, such as the stage-specific expression of pathogen proteins and tissue expression of host proteins. The technique was applied to ten pathogens, including species of mycobacterium, apicomplexa, and kinetoplastida, responsible for “neglected” human diseases. The method was assessed by (i) comparison to a set of known host–pathogen interactions, (ii) comparison to gene expression and essentiality data describing host and pathogen genes involved in infection, and (iii) analysis of the functional properties of the human proteins predicted to interact with pathogen proteins. The final set of 1,501 potential host–pathogen protein interactions predicted for *Plasmodium falciparum*, one of the ten pathogens studied, is approximately 5 orders of magnitude smaller than the initial set of all possible pairs, with an estimated 2 order of magnitude enrichment in true interactions. The predictions include interactions from previously characterized mechanisms, such as cytoadhesion and protease inhibition, as well as suspected interactions in hypothesized networks, such as apoptotic pathways. We present several specific predictions which warrant experimental follow-up. Our computational method provides a means to mine whole-genome data and is complementary to experimental efforts in elucidating networks of host–pathogen protein interactions. It can ultimately aid in the identification of potential targets for immunization and chemotherapy strategies.

4.1 Introduction

Genome sequencing has changed the scale and diversity of biomedical problems amenable to investigation [77]. Complete genome sequences are now readily available for many species, including human and a number of biomedically relevant microbes. Functional insights into the proteins encoded by these genomes are emerging from technical advances such as three-dimensional structure determination and the detection of genetic and physical interactions [18, 19, 22, 240, 244]. For example, within one year of the SARS outbreak, the causative agent was classified as a new coronavirus using DNA microarray technology, its genome was sequenced, structures of key proteins were proposed by comparative modeling, and potential inhibitors were identified through molecular docking studies [14, 117, 184, 234, 245]. However, in general, the wealth of genomic information available for both human host and pathogens remains unmined due to the lack of whole-genome protocols that can predict host–parasite interactions.

Pathogens have evolved numerous strategies to successfully invade their hosts, acquire nutrients, and evade their immune defenses [147]. These strategies often involve direct interactions between host and pathogen molecules, including the formation of protein complexes. Molecular mimicry of host cellular components has been observed as a bacterial and viral infection strategy in which the pathogen is able to circumvent the host immune system as well as harness the host signaling pathways to benefit its own survival [7, 210]. For example, the *Salmonella* virulence protein SptP activates the human Rho-GTPase Rac1 in a structurally similar mode as human Cdc42 GAP activates the human Rho-GTPase Cdc42 [209]. Host factors also determine the course of infection [105]. Productive HIV in-

fection requires several host factors, including ATM kinase, inhibition of which was recently demonstrated to suppress HIV infection [71, 119].

The host and pathogen genes and proteins involved in the infection process are currently studied using traditional small-scale biochemical and genetic experiments, which focus on one protein or pathway at a time, or a few genome-scale studies that survey the response of many genes and proteins simultaneously. The genome-scale studies utilize technologies such as RNA interference, transposon mutagenesis, and genomic microarrays to identify and characterize the genes involved in infection and the time-course of their involvement [28, 89, 166].

Although the exact number of host–pathogen protein interactions involved in infection is difficult to estimate, it is likely that the currently known interactions for specific host–pathogen pairs, which vary in number from none to approximately ten, do not represent the complete interaction networks. Recently, a computational method predicted 20 putative interactions between 20 human and 8 Kaposi’s sarcoma-associated herpesvirus (KSHV) proteins. 13 of these interactions were verified by co-immunoprecipitation, all of which were previously unidentified [226]. Thus, even for pathogens with small genomes, such as the ~ 80 open reading frames in KSHV, much is left to be learned about the network of interactions between host and pathogen proteins.

While experimental efforts to characterize these networks can enhance our knowledge of the infection process, large-scale experimental methods, such as tandem affinity purification and yeast-two-hybrid experiments, also exhibit significant false negative as well as false positive error rates. For example, yeast-two-hybrid experiments have an estimated false

positive rate of over 50% and a false negative rate approaching 90% [81]. Computational methods have demonstrated utility in improving the coverage and accuracy of identifying protein–protein interactions in combination with experimental data sets [96, 121]. Computational methods may similarly complement large-scale experimental efforts to characterize host–pathogen interaction networks. These methods incur essentially no material costs, present no safety concerns to laboratory workers, and are unaffected by the experimental challenges of culturing pathogens, and expressing and purifying their proteins.

One useful role of computation is to reduce the total number of potential host–pathogen protein interactions to an experimentally tractable number of interactions, while improving the enrichment of true interactions. Here, we develop and apply a protocol that sequentially reduces the number of potential protein interactions between a host and pathogen by combining experimental genomic, proteomic, and structural data together with comparative protein structure modeling to predict pairs of pathogen and host proteins that potentially interact physically. The protocol begins by identifying pairs of host and pathogen proteins that each have similarity to components of a known interaction. Next, the putative interface is assessed using structure, if available. Finally, independent information that captures the biological context of potential interactions, such as the stage-specific expression of pathogen proteins and tissue expression of host proteins, is used to filter the initial set of pairs. The result of the protocol is an enriched candidate set that is suitable for subsequent experimental study. Here, we have applied the protocol to ten human pathogens, including species of mycobacteria, kinetoplastida, and apicomplexa, that are responsible for “neglected” human diseases. These pathogens infect over one billion people and incur over

one million annual deaths [242].

We first describe the protocol, detailing the data sources and the computations used. We then present the predictions made for the ten pathogens. We assess the protocol by (i) comparison to a set of known host–pathogen protein interactions, (ii) comparison to gene expression and essentiality data describing host and pathogen genes involved in infection, and (iii) analysis of the functional properties of the human proteins predicted to potentially interact with pathogen proteins. We present several specific predictions whose support in the literature warrants experimental follow-up. We discuss the observed performance of the method and future improvements. Finally, we conclude by discussing the implications of these results for understanding the molecular mechanisms of pathogenesis and designing immunization and chemotherapy strategies.

4.2 Results

The protocol begins with the target set of host and pathogen protein sequences (Fig 4.1, Materials and Methods). These sequences are first modeled by an automated comparative protein structure modeling pipeline that detects similarities to all known protein sequences and structures. Pairs of host and pathogen proteins that each have detectable similarity to components of a known interaction are identified. If an experimentally determined three dimensional structure of the known interaction is available, structural models are built for the host and pathogen target sequences, and their putative interface is assessed using a statistical potential score. In the absence of structure, the significance of the similarity between the target proteins and their corresponding interaction templates is

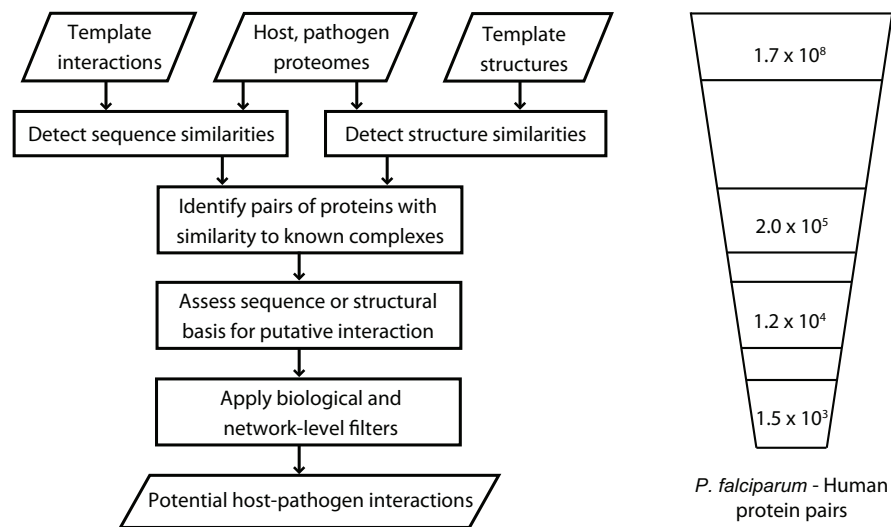


Figure 4.1: Prediction protocol. The protocol begins with the set of host and pathogen proteins. Sequence matching procedures are then used to identify similarities between the host or pathogen proteins and proteins with known structure or known interaction partners. A structure-based statistical potential assessment, or a sequence similarity score in the absence of structure, is then used to predict potential interacting partners. Finally, this set of potential interactions is filtered using the biological contexts of the host and pathogen proteins and a network-level filter. The protocol reduces the number of potential *P. falciparum* – human protein interactions by approximately 5 orders of magnitude (Table 4.2).

assessed using a sequence similarity score. These potential interactions are then filtered by the biological context of their component proteins, such as life-cycle stage and tissue expression, and by the network-level information, including the template usage frequencies.

4.2.1 Detecting sequence and structure similarities

The interaction template coverage of the pathogen proteomes, or fractions of the pathogen proteomes that had detectable similarity to a protein that in turn interacted with a protein to which a human protein had detectable similarity, ranged from 16% of *T. cruzi* sequences to 25% of *C. parvum* sequences (Table 4.1). 34% of the human proteome had

Pathogen	Protein Sequences	With Interaction Templates		With Biological Data	
<i>M. leprae</i>	1,601	359	22%	1,023	64%
<i>M. tuberculosis</i>	3,954	729	18%	2,551	65%
<i>L. major</i>	8,009	1,908	24%	3,749	47%
<i>T. brucei</i>	8,965	1,817	20%	4,040	45%
<i>T. cruzi</i>	19,245	3,147	16%	8,604	45%
<i>C. hominis</i>	3,886	780	20%	1,591	41%
<i>C. parvum</i>	3,806	958	25%	1,828	48%
<i>P. falciparum</i>	5,342	1,126	21%	4,691	88%
<i>P. vivax</i>	5,334	1,131	21%	413	8%
<i>T. gondii</i>	7,787	1,311	17%	3,627	47%
<i>H. sapiens</i>	32,010	10,993	34%	26,595	83%

Table 4.1: Interaction template and biological data coverage of the genomes analyzed. Our automated comparative protein modeling pipeline MODPIPE was used to detect sequence and structure similarities to proteins in known complexes. Biological coverage refers to those proteins for which at least one type of annotation was available (Table 4.6).

detectable similarity to a protein that in turn interacted with a protein to which a pathogen protein had detectable similarity.

4.2.2 Identifying pairs of proteins with similarity to known complexes

The number of pairs of host and pathogen proteins that had detectable similarity to known interacting proteins varied widely among the pathogen species, with the bacterial pathogens having far fewer pairs than the eukaryotic pathogens (Table 4.2 column 2). For example, 43,528 host–pathogen protein pairs were identified for the *M. tuberculosis* proteome (3,954 sequences, 18% interaction template coverage), while 160,952 pairs were identified for the *C. hominis* proteome with approximately the same proteome size and interaction template coverage (3,886 sequences, 20% interaction template coverage). Among the eukaryotic pathogens, the number of host–pathogen protein pairs varied approximately

in proportion to the size of the pathogen proteome (Tables 4.1, 4.2).

4.2.3 Assessing the sequence or structural basis of the potential interactions

The assessment procedure, using the statistical potential score when structure was available, or the sequence identity threshold in the absence of structure, identified approximately 5% of the host–pathogen pairs identified in the previous step as possible interacting partners (Table 4.2). The reduction in the number of pairs by this assessment was greatest for the *T. gondii*–human pairs, of which only 3.4% passed the scoring thresholds.

As expected from the number of host–pathogen protein pairs with interaction templates, the *Mycobacterium* species have far fewer predicted potential interactions than the eukaryotic pathogens. The number of potential interactions predicted between *M. tuberculosis* and human proteins was approximately 3-fold smaller than that for the similarly sized *C. hominis* proteome.

4.2.4 Applying biological and network-level filters

The fractions of the pathogen proteomes with biological annotations varied widely, from the well characterized *P. falciparum* proteome (88%) to the minimally characterized *P. vivax* proteome (8%) (Tables 4.1, 4.6). The low biological coverage of *P. vivax*, whose genome sequencing is still in progress, results from minimal gene ontology function annotation as well as poor experimental coverage relative to other *Plasmodium* species. Potential interactions between host and pathogen proteins that each met at least one biological cri-

terion were considered to pass the biological context filter (Materials and Methods).

Next, we implemented a network-level filter that flagged those predictions based on templates that were used for more than 1% of the total predictions. This filtering was done because these predictions exhibited a low level of observed interaction specificity. For example, many pairs of G-protein subunits α and β were predicted to potentially interact with each other based on the crystal structure of the G-protein GI heterotrimer (PDB 1GG2).

Application of the biological and network-level filters resulted in a wide range of reductions in predicted complexes (Table 4.2). This variability in reduction is due to the different levels of biological annotation used for the genomes. For example, *P. falciparum* had the highest biological annotation coverage (88%) and, as expected, the highest fraction of potential interactions that passed the biological and network-level filters (13%). This final set of potential *P. falciparum*–human interactions is a 5 order of magnitude smaller than the initial set of all possible protein pairs. The low coverage of biological annotation for other pathogens was also evident, as filtering the predictions for two pathogens, *T. brucei* and *T. gondii*, resulted in removal of all potential interactions.

4.2.5 Assessment

Next, the predictions were assessed to characterize the coverage and accuracy of the method. Coverage refers to the fraction of interactions that are accessible by the method and accuracy refers to the fraction of the covered interactions that were correctly identified. We assessed the quality of the protocol in three ways.

Pathogen	Pairs with templates		Potential interactions		Filtered interactions	
<i>M. leprae</i>	26,234	(6,200/359)	1,351	(706/101)	13	(13/1)
<i>M. tuberculosis</i>	43,528	(6,549/729)	2,474	(992/240)	45	(41/13)
<i>L. major</i>	411,468	(9,978/1,908)	22,243	(2,680/656)	289	(186/29)
<i>T. brucei</i>	427,884	(9,935/1,817)	20,797	(2,546/661)	0	(0/0)
<i>T. cruzi</i>	750,419	(10,078/3,147)	33,869	(2,601/1,028)	914	(356/138)
<i>C. hominis</i>	160,952	(9,118/780)	7,237	(1,854/257)	79	(59/8)
<i>C. parvum</i>	203,570	(9,242/958)	10,987	(2,108/335)	211	(156/13)
<i>P. falciparum</i>	200,428	(9,554/1,126)	11,655	(2,291/434)	1,501	(826/216)
<i>P. vivax</i>	211,185	(9,546/1,131)	12,159	(2,305/399)	34	(26/4)
<i>T. gondii</i>	216,187	(9,683/1,311)	7,282	(2,024/261)	0	(0/0)

Table 4.2: Potential interaction set reduction by assessment and filtering. The potential interactions meet the structural assessment or sequence alignment significance criteria. These interactions are then filtered so that they meet at least one pathogen biological criterion, one host biological criterion, and are based on a template that is used for less than 1% of the total number of predictions in a given host–pathogen network.

4.2.6 Assessment I: Comparison of predicted and known host–pathogen protein interactions

The predicted potential interactions were first compared to a set of known host–pathogen protein interactions (Table 4.3). This set of known interactions, which are the result of studies that have focused on single proteins or interactions, is too small to rigorously assess the method. However, this set still allows insight into the performance of the method, highlighting what is missed and what is captured. Our protocol recovered 4 of the previously identified host–pathogen protein interactions published in the literature for the ten pathogen species. Other known interactions were not identified because of the lack of available template interactions in the interaction databases. None of these latter cases was due to incorrect assessment by our method (Fig 4.1 step 3). As expected, this result suggests that currently, a limitation of the protocol is the restriction of the coverage

to interactions with an appropriate template.

	Pathogen	Pathogen Protein	Human Protein	Evidence	Predicted
1	<i>M. leprae</i>	histone-like protein	laminin-2	[41]	No template
2	<i>M. leprae</i>	fibronectin-attachment protein	fibronectin	[216]	No template
1	<i>M. tuberculosis</i>	Rv3763	TLR2	[124]	No template
2	<i>M. tuberculosis</i>	Rv1411c	TLR2	[67]	No template
3	<i>M. tuberculosis</i>	glycoprotein Apa	pulmonary surfactant protein A	[176]	No template
4	<i>M. tuberculosis</i>	heparin-binding hemagglutinin	complement C3	[146]	No template
5	<i>M. tuberculosis</i>	fibronectin-attachment protein	fibronectin	[2, 146]	No template
	<i>L. major</i>	none			
1	<i>T. brucei</i>	ornithine decarboxylase	ornithine decarboxylase	[158]	Yes
2	<i>T. brucei</i>	serum resistance associated protein	apolipoprotein L-I	[231]	No template
3	<i>T. brucei</i>	trypanopain-Tb	cystatins	[223]	Yes
1	<i>T. cruzi</i>	Tc85-11 (trans-sialidase)	cytokeratin 18	[129]	No template
2	<i>T. cruzi</i>	Tc85-11 (trans-sialidase)	laminin	[69, 134]	No template
3	<i>T. cruzi</i>	calreticulin	complement component 1 q	[4]	No template
4	<i>T. cruzi</i>	cruzipain	alpha-2-macroglobulin	[177, 178]	No template

Continued on next page

Table 4.3 – continued from previous page

	Pathogen	Pathogen Protein	Human Protein	Evidence	Predicted
5	<i>T. cruzi</i>	cruzipain	cystatins	[212]	Yes
6	<i>T. cruzi</i>	cruzipain	pregnancy zone protein	[178]	No template
7	<i>T. cruzi</i>	gp82 (trans-sialidase)	mucin	[151]	No template
8	<i>T. cruzi</i>	SA85-1.1	mannose receptor	[102]	No template
9	<i>T. cruzi</i>	SA85-1.1	mannose-binding protein	[102]	No template
10	<i>T. cruzi</i>	Tc13 (trans-sialidase)	beta-1-adrenergic receptor	[63]	No template
11	<i>T. cruzi</i>	trans-sialidase	sialomucin cd43	[218]	No template
12	<i>T. cruzi</i>	trans-sialidase	cruzin	[171, 172]	No template
13	<i>T. cruzi</i>	gp72	complement component C3	[97]	No template
	<i>C. hominis</i>	none			
	<i>C. parvum</i>	none			
1	<i>P. falciiparum</i>	falcipain-2	(<i>G. gallus</i>) cystatin	[238]	Yes
2	<i>P. falciiparum</i>	MESA	protein 4.1	[237]	No template
3	<i>P. falciiparum</i>	PfEMP1	CD36	[236]	No template
4	<i>P. falciiparum</i>	PfEMP1	ICAM-1	[236]	No template
Continued on next page					

Table 4.3 – continued from previous page

	Pathogen	Pathogen Protein	Human Protein	Evidence	Predicted
5	<i>P. falciparum</i>	PfHRP1	ankyrin	[130]	No template
6	<i>P. falciparum</i>	MSP1	band 3	[70]	No template
7	<i>P. falciparum</i>	EBA-181	erythrocyte protein 4.1	[118]	No template
8	<i>P. falciparum</i>	EBA-175	glycophorin A	[156, 202]	No template
9	<i>P. falciparum</i>	EBA140	glycophorin C	[131]	No template
10	<i>P. falciparum</i>	PfEMP1	complement receptor 1	[116]	No template
11	<i>P. falciparum</i>	Circumsporozoite	LDLR-related protein	[197]	No template
1	<i>P. vivax</i>	Duffy-binding protein	Duffy antigen	[80]	No template
1	<i>T. gondii</i>	microneme protein 2	ICAM-1	[20]	No template

Table 4.3: Comparison to known host-pathogen protein interactions. The predicted potential interactions were compared to known host-pathogen interactions to identify those that were found and those that were not.

4.2.7 Assessment II: Comparison to gene expression and essentiality data

Next, we compared our pre-filtered predictions to genome-scale datasets describing pathogen genes involved in *M. tuberculosis* infection and human genes involved in *L. major*, *M. tuberculosis*, and *T. gondii* infections. These comparisons were performed because genomic studies are so far the only techniques that have produced large-scale datasets describing host–pathogen interactions, even though previous studies have observed only weak correlation between physical protein interactions and expression data [95, 145].

Previous studies have identified 194 *M. tuberculosis* genes that are essential for *in vivo* infection [194], as well as 286 genes that are up-regulated in granuloma, pericavity, or distal lung infection sites compared to *in vitro* conditions [174]. We compared these two sets of genes to the set of *M. tuberculosis* proteins that we predicted to potentially interact with human proteins. The overlap between these three data sets is minimal (Table 4.4). In fact, only one gene occurs in both experimental data sets as well as our computational predictions: Rv3910 (15611046), a probable conserved trans-membrane protein. The overlap of our predictions with the set of genes upregulated during infection (23 genes) is greater than the overlap between the two experimental sets of upregulated genes and essential genes (18 genes).

Previous studies have identified human genes that are differentially regulated in response to a variety of protozoal infections, in particular the macrophage and dendritic cells that are involved in the immune response [33]. The human proteins that we predicted to potentially interact with *L. major*, *M. tuberculosis*, and *T. gondii* include respectively 231, 78, and 169 proteins that are encoded by genes observed to be differentially expressed

Pathogen	Data Set 1 (size)	Data Set 2 (size)	Overlap
(a) Pathogen proteins			
<i>M. tuberculosis</i>	Rachman (286)	Predictions (240)	23
<i>M. tuberculosis</i>	Sassetti (194)	Predictions (240)	8
<i>M. tuberculosis</i>	Rachman (286)	Sassetti (194)	18
(b) Host proteins			
<i>L. major</i>	Chaussabel (3060)	Predictions (2680)	231
<i>M. tuberculosis</i>	Chaussabel (2893)	Predictions (992)	78
<i>T. gondii</i>	Chaussabel (2475)	Predictions (2024)	169

Table 4.4: Comparison of predictions to experimental observations of proteins involved in infection. (a) *M. tuberculosis* proteins predicted (pre-filtered) to potentially interact with host proteins are compared to genes observed to be essential for *in vivo* growth (Sassetti [194]) and those up-regulated in granuloma, pericavity, or distal lung infection sites (Rachman [174]). (b) *H. sapiens* proteins predicted (pre-filtered) to potentially interact with pathogens are compared to genes that are differentially regulated in macrophages or dendritic cells upon infection by *L. major*, *M. tuberculosis*, and *T. gondii* (Chaussabel [33]).

in macrophages and dendritic cells upon infection by these pathogens (Table 4.4(b)) [33].

Again, the minimal observed overlap is a function of the weak correlation between physical protein interactions and gene expression data, as previously observed [95, 145].

4.2.8 Assessment III - Functional overview of predicted potential interactions

Finally, we analyzed the functional annotations of the human proteins predicted to potentially interact with pathogen proteins in order to identify functions that were significantly enriched compared to the whole human proteome. This functional analysis is useful both to quickly summarize the predictions and to evaluate the relevance of the predicted interactions to the infection process. The human proteins predicted to potentially interact with pathogen proteins were significantly enriched in several gene ontology function terms, even before application of the biological filters (Table 4.5). For example, the hu-

man proteins predicted to potentially interact with *M. tuberculosis* are enriched in cellular component terms that make sense in light of known mechanisms of tuberculosis infection including immunological synapse (7.7 fold enrichment, p-value = 10^{-3}), T-cell receptor complex (8.5 fold enrichment, p-value = $1.6 \cdot 10^{-2}$), and autophagic vacuole (17.1 fold enrichment, p-value = $3 \cdot 10^{-4}$). These terms all reflect the known immunobiology of this pathogen, which is known to elicit a T-cell response and was recently found to be eliminated through autophagy [36, 42, 76, 204, 233]. Similarly, the human proteins predicted to potentially interact with *P. falciparum* proteins are enriched in terms such as extrinsic to plasma membrane (5.2 fold enrichment, p-value = $9.2 \cdot 10^{-15}$) and homophilic cell adhesion (4.2 fold enrichment, p-value = $2.8 \cdot 10^{-21}$).

The enriched functional terms that have not been previously implicated in the infection process represent either novel biological insight or false positives. Identification of these terms as mechanisms that are relevant *in vivo* or false positive requires experiments beyond the scope of this paper. However, some of the enriched terms suggest that false positives could be identified and discarded if they arise from conservation of core cellular components. For example, the conservation of core translation machinery across all divisions of life [215] could result in an erroneously predicted potential interactions causing the enrichment in the human-*P. falciparum* network for eukaryotic translation elongation factor (7.4 fold, p-value = $8.4 \cdot 10^{-4}$). Similarly, terms such as pyruvate dehydrogenase activity (25.6 fold, p-value = $2.2 \cdot 10^{-2}$) and aspartate-tRNA ligase activity (24.4 fold, $5.3 \cdot 10^{-5}$) that are enriched in the human proteins predicted to interact with *M. tuberculosis*, may also be false positives caused by the conservation of core cellular components, and should

be filtered.

Rank	GO ID	Function	Number	Enrichment	P-value
(a) Cellular component of all human proteins predicted to potentially interact with <i>M. tuberculosis</i>					
1	GO:0005776	autophagic vacuole	5	17.1	$3.0 \cdot 10^{-4}$
2	GO:0005853	eukaryotic translation elongation factor 1 complex	5	12.2	$2.2 \cdot 10^{-3}$
3	GO:0042101	T cell receptor complex	5	8.5	$1.6 \cdot 10^{-2}$
4	GO:0001772	immunological synapse	7	7.7	$1.0 \cdot 10^{-3}$
5	GO:0005884	actin filament	8	5.3	$4.3 \cdot 10^{-3}$
6	GO:0005746	mitochondrial electron transport chain	8	4.9	$7.6 \cdot 10^{-3}$
7	GO:0044455	mitochondrial membrane part	12	3.8	$1.2 \cdot 10^{-3}$
8	GO:0042995	cell projection	23	2.2	$1.2 \cdot 10^{-3}$
9	GO:0015629	actin cytoskeleton	25	2.2	$5.1 \cdot 10^{-4}$
10	GO:0031410	cytoplasmic vesicle	22	1.9	$1.5 \cdot 10^{-2}$
(b) Biological process of all human proteins predicted to potentially interact with <i>M. tuberculosis</i>					
1	GO:0006021	myo-inositol biosynthetic process	3	34.1	$1.4 \cdot 10^{-2}$
2	GO:0019642	anaerobic glycolysis	5	34.1	$6.5 \cdot 10^{-6}$
3	GO:0006422	aspartyl-tRNA aminoacylation	5	24.4	$1.3 \cdot 10^{-4}$
Continued on next page					

Table 4.5 – continued from previous page

Rank	GO ID	Function	Number	Enrichment	P-value
4	GO:0032011	ARF protein signal transduction	7	23.9	$3.4 \cdot 10^{-7}$
5	GO:0032012	regulation of ARF protein signal transduction	7	23.9	$3.4 \cdot 10^{-7}$
6	GO:0046847	filopodium formation	6	17.1	$1.1 \cdot 10^{-4}$
7	GO:0051014	actin filament severing	4	17.1	$1.9 \cdot 10^{-2}$
8	GO:0043088	regulation of Cdc42 GTPase activity	5	14.2	$4.5 \cdot 10^{-3}$
9	GO:0032489	regulation of Cdc42 protein signal transduction	5	14.2	$4.5 \cdot 10^{-3}$
10	GO:0032318	regulation of Ras GTPase activity	5	14.2	$4.5 \cdot 10^{-3}$
(c) Molecular function of all human proteins predicted to potentially interact with <i>M. tuberculosis</i>					
1	GO:0016872	intramolecular lyase activity	3	34.1	$5.6 \cdot 10^{-3}$
2	GO:0004512	inositol-3-phosphate synthase activity	3	34.1	$5.6 \cdot 10^{-3}$
3	GO:0019967	interleukin-1, Type I, activating binding	4	27.3	$5.9 \cdot 10^{-4}$
4	GO:0004909	interleukin-1, Type I, activating receptor activity	4	27.3	$5.9 \cdot 10^{-4}$
5	GO:0004739	pyruvate dehydrogenase (acetyl-transferring) activity	3	25.6	$2.2 \cdot 10^{-2}$
6	GO:0005094	Rho GDP-dissociation inhibitor activity	3	25.6	$2.2 \cdot 10^{-2}$
7	GO:0004738	pyruvate dehydrogenase activity	3	25.6	$2.2 \cdot 10^{-2}$
Continued on next page					

Table 4.5 – continued from previous page

Rank	GO ID	Function	Number	Enrichment	P-value
8	GO:0004591	oxoglutarate dehydrogenase (succinyl-transferring) activity	3	25.6	$2.2 \cdot 10^{-2}$
9	GO:0004815	aspartate-tRNA ligase activity	5	24.4	$5.3 \cdot 10^{-5}$
10	GO:0004459	L-lactate dehydrogenase activity	7	23.9	$1.4 \cdot 10^{-7}$

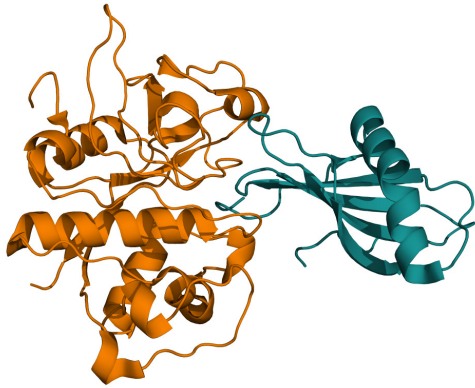
Table 4.5: Functional annotation of human proteins predicted to interact with *M. tuberculosis*. The ten (a) cellular component, (b) biological process, and (c) molecular function annotation terms that are most enriched in the set of human proteins predicted to potentially interact with *M. tuberculosis* proteins, compared to the background, are listed. The enriched terms were identified and their significance were computed by GO::TermFinder using a Bonferroni correction[26].

4.3 Discussion

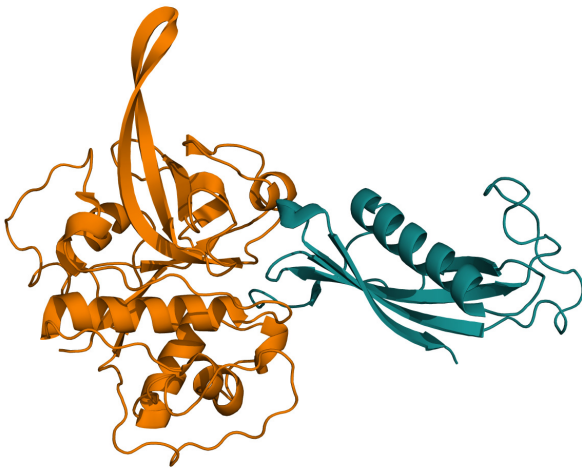
We will now discuss several of the predicted potential interactions and their support in the literature. We then describe the observed performance of the method, including its advantages and limitations, and future developments to improve the performance. We close by discussing possible applications of the method to aid in understanding host-pathogen interactions as well as other types of inter-species interactions.

4.3.1 Specific examples of potential interactions

(1) *Enzyme dimerization.* We predicted several potential inter-species enzyme dimerizations, such as *T. brucei* ornithine decarboxylase (ODC) binding to human ODC. Functional dimerization of parasitic and host enzyme subunits has been previously observed, such as in *T. brucei* and mouse ODC [158]. The number of instances and relevance of this kind of dimerization to infection need to be experimentally validated. In the case of ODC, both host and pathogen ODCs have been implicated in viral and protozoal infections. For example, inhibition of host ODC reduces the ability of host macrophages to take up parasites, such as *T. cruzi* [109], and inhibition of parasite ODC has been demonstrated to reduce parasite growth in *P. falciparum* and *Leishmania donovani* [75, 203]. However, these phenotypes are not necessarily due to interaction of ODC subunits across the species, but rather the polyamines synthesized by the enzyme. Because the *in vivo* relevance of these homodimer-like complexes is not clear, we generally removed predictions based on homodimer sequence templates or template structures of subunits classified in the same domain family (Materials and Methods). This restriction also facilitates visualization and

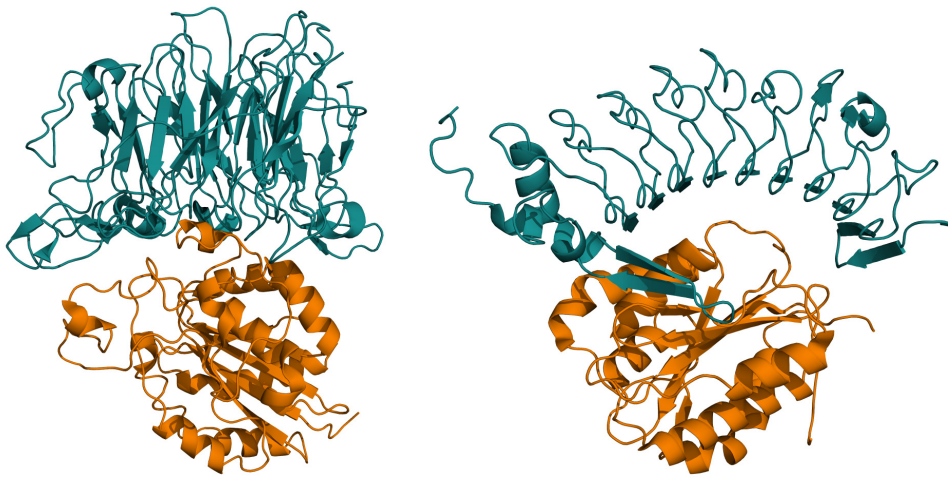


(a) Template structure: cathepsin-H –
cystatin-A



(b) Experimental structure: falcipain-2 – *Gallus gallus*
cystatin

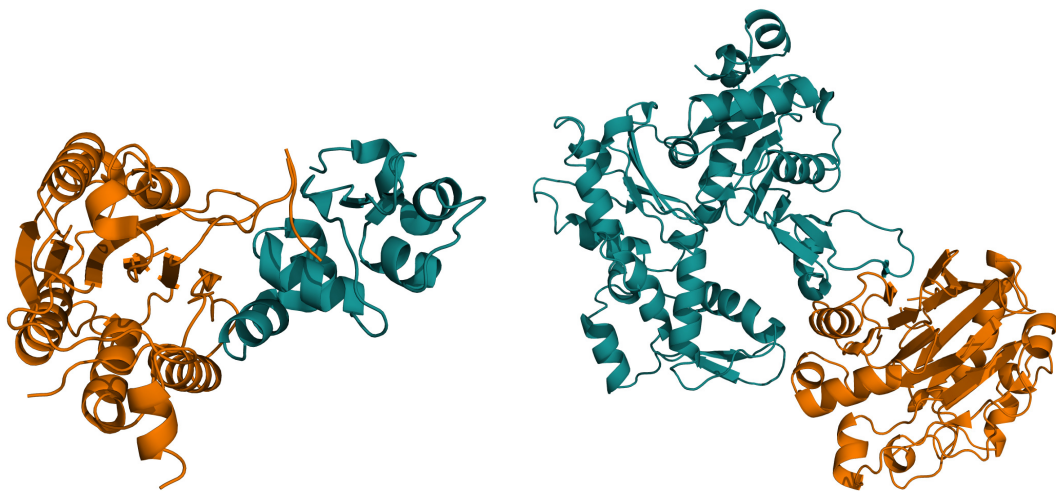
Figure 4.2: Example of a validated prediction: falcipain-2 – cystatin-A. A potential interaction was predicted between falcipain-2 and cystatin-A based on (a) a template structure of cathepsin-H (orange) bound to cystatin-A (teal) (PDB 1nb3). (b) The structure of falcipain 2 bound to chicken cystatin was recently experimentally determined (PDB 1yvb). Although the interaction is experimentally verified, the question remains whether it would occur *in vivo*. Figures were generated by PyMOL (<http://www.pymol.org>).



(a) *P. falciparum* PFI0595c – CD51

(b) *P. falciparum* TRAP – TLR4

Figure 4.3: Examples of predicted potential interactions. (a) *P. falciparum* PFI0595c was predicted to potentially interact with human CD51 (Integrin alpha-V; ENSP0000261023) based on a structure of platelet membrane glycoprotein IIIA bound to CD51, respectively (PDB 1JV2). (b) *P. falciparum* Thrombospondin-related anonymous protein (TRAP) was predicted to potentially interact with Toll-like receptor 4 (TLR4) based on a template structure of Glycoprotein IB alpha (orange) bound to Von Willenbrand Factor (blue), respectively (PDB 1M10).



(c) *P. falciparum* PF13.0289 – XIAP

(d) *M. tuberculosis* probable exported protein

Rv0888 – Actin

Figure 4.3: Examples of predicted potential interactions. (cont) (c) *P. falciparum* Hypothetical metacaspase PF13.0289 was predicted to potentially interact with X-linked inhibitor of apoptosis protein (XIAP) based on a template structure of Caspase-9 (orange) bound to XIAP (teal), respectively (PDB 1NW9). (d) *M. tuberculosis* probable exported protein Rv0888 was predicted to interact with actin, based on a structure of DNase bound to actin (PDB 1ATN). Figures were generated by PyMOL (<http://www.pymol.org>).

analysis of the networks, although some true positive predictions may be lost.

(2) *Cell adhesion.* *P. falciparum* is known to adhere to host cell cytoadhesion molecules. Several of the potential interactions predicted by our method involved human cell adhesion proteins such as integrins and fibronectin containing proteins. For example, we predicted that PFI0595c, a hypothetical *P. falciparum* protein, may potentially interact with human CD51 (Integrin alpha-V; ENSP0000261023) based on a structure of platelet membrane glycoprotein IIIA bound to CD51 (PDB 1JV2 [246]) (Fig 4.3). Previous studies have observed that *P. falciparum* adherence to human microvascular endothelial cells involves CD51 [199]. In addition, no potential interactions were predicted between integrins, such as CD51, and *P. vivax* proteins. This prediction is in agreement with previous experimental evidence that indicates *P. vivax* does not engage in microvascular sequestration [201].

(3) *Actin cytoskeleton rearrangement.* We predicted that *M. tuberculosis* probable exported protein Rv0888 (GI 15608028) may potentially interact with several human α -actins (ENSP00000295137) based on the template structure of DNase-I bound to actin (PDB 1ATN) [101] (Fig 4.3). Rv0888 is thought to be exported by *M. tuberculosis*, while the human protein is essentially ubiquitously expressed, including in macrophages where the pathogen lives. The interaction between DNase and actin is known to be strong enough to depolymerize actin [101], and could potentially explain the observed *M. tuberculosis* rearrangement of host actin [74]. In addition to phagocytic cells, such as macrophages, *M. tuberculosis* has been observed to be internalized by non-phagocytic cells, such as pneumocytes [64]. Actin filament rearrangement has also been observed in these types of cells,

and is hypothesized to be triggered by a secreted pathogen factor [64]. The potential interaction we predicted suggests a possible molecular mechanism for the changes in the actin cytoskeleton that are observed upon infection.

(4) *Innate Immunity: Toll-like receptor signaling.* We predicted that *P. falciparum* thrombospondin-related adhesive protein (TRAP, SSP2, PF13.0201) may potentially interact with human Toll-like receptor 4 (TLR4, ENSP00000346893), based on a template structure of Glycoprotein IB alpha bound to Von Willenbrand Factor (PDB 1M10 [91]) (Fig 4.3). TLR4, a “pattern recognition module” involved in the innate immune response, is known to play a role in malaria through recognition of the glycosylphosphatidylinositol (GPI) membrane anchors of *P. falciparum* proteins [114]. Single nucleotide polymorphisms have been observed in the TLR4 gene that are associated with an increased severity of malaria, but they fall outside of the modeled region [143]. TRAP is known to elicit an immune response and so has been used as a component of several vaccine candidates [21, 85, 144, 170]. Analysis of TRAP sequence data from a Gambian *P. falciparum* population indicates that the gene is under strong selection for variation in the sequence, with peaks in this variation occurring in the A-domain, that we predicted could potentially mediate the interaction with TLR4 [239]. The possible encounter of these two proteins is also supported by the biological context, TRAP is expressed on the surface of the parasite during the sporozoite stage of the plasmodium life-cycle and TLR4 is known to express in the liver. While alternative explanations are possible, the biological evidence and the structural predictions made here suggest that a TRAP – TLR4 interaction may play a role *in vivo* in the infection process.

(5) *Proteases*. We predicted several potential interactions between proteases and protease inhibitors, such as the *P. falciparum* falcipain-2 protease and the human cystatin-A inhibitor suggested based on a template structure of human cathepsin-H bound to cystatin-A (PDB 1NB3, Fig 4.2(a)). This prediction was recently experimentally validated, with chicken cystatin [238] (PDB 1YVB, Fig 4.2(b)). This crystal structure was not present in our template set, because it has not yet been classified by the SCOP domain annotation database (Materials and Methods, [148]). Thus, the predicted complex was a true blind prediction.

Although the experimentally determined structure provides direct validation of our prediction, it does not imply relevance to the infection process. However, cysteine proteases are known to be involved in malaria pathogenesis [159]. The known involvement of cysteine proteases and experimentally established cross-talk between host and pathogen protease and inhibitors suggests that the predicted interaction may play a role during infection.

This case is an example where structure is important both in making the prediction and in highlighting its relevance as a potential pharmacologic target. A sequence-based method with the established threshold of 80% joint sequence identity [247] would have missed this interaction, because falcipain-2 and cathepsin-H share only 34% sequence identity. However, comparison of the experimental falcipain-2–cystatin structure with the template cathepsin-H–cystatin-A structure reveals a C- α RMSD of 0.43 Å at the interface. In addition, this structure can be used to search for small-molecules that may disrupt or mimic the target interaction.

(6) *Apoptosis*. We predicted that the hypothetical *P. falciparum* meta-caspase

PF13_0289 may potentially interact with several human apoptosis inhibitors, including X-linked inhibitor of apoptosis protein (XIAP), melanoma inhibitor of apoptosis protein (ML-IAP), and neuronal apoptosis inhibitory protein, based on a template structure of caspase-9 interacting with baculovirus inhibitor of apoptosis protein (PDB 1NW9) (Fig 4.3). *P. falciparum* does not contain a true caspase, such as those found in animals, but instead contains metacaspases, which are also found in plants, fungi, and other protozoa [227]. Nevertheless, animal inhibitors of apoptosis, such as XIAP, have been shown to affect cell death programs in plants, which also lack true caspases [45, 123]. In addition, apoptosis-like cell death has recently been observed in the mid-gut of mosquitoes as a result of malaria infection [92]. These observations suggest that apoptotic machinery cross-talk, such as the potential interactions predicted here, may be relevant to *in vivo* infection.

4.3.2 Enrichment of potential interactions with actual interactions

We initially suggested that one role of computation is to reduce the total number of possible host–pathogen protein interactions to a more experimentally tractable number of interactions, while improving the enrichment of true interactions. The protocol we presented here sequentially reduced the total number of potential interactions using a series of assessments: (i) identifying template interactions, (ii) assessing the putative interaction, using structure if available, and finally (iii) filtering using biological context and network-level information (Fig 4.1). For example, the procedure resulted in a 5 order of magnitude reduction in the number of possible human–*P. falciparum* protein interactions (Table 4.2).

We expect that this reduced set of potential interactions predicted by the protocol

contains a higher fraction of true interactions compared to the initial set of all possible host–pathogen protein pairs. This expected enrichment is difficult to assess directly, because there are no large sets of known host–pathogen protein interactions. However, the enrichment can be estimated using the intra-proteomic predictions previously made for *S. cerevisiae*. Of the 3,387 interactions predicted by the structure-based method, 270 occurred in experimental datasets (8%) [40]. In total, 19,424 interactions have been observed out of the possible 21,776,700 pairs of yeast proteins (0.09%) (Jan 2006) [40]. Thus, the number of potential yeast protein pairs was reduced by approximately 4 orders of magnitude, while the enrichment was increased by approximately ~ 2 orders of magnitude. These estimates reflect intra-proteomic predictions and do not reflect the errors that uniquely affect host–pathogen predictions, such as the conservation of core components discussed above.

4.3.3 Limitations in coverage

The performance of any method, computational or experimental, can be characterized by two factors: coverage, describing the fraction of all interactions covered by the method, and accuracy, describing the fraction of the covered interactions that were correctly identified. The method presented is affected by various sources of error that affect both the coverage and accuracy.

The main factor that limits the coverage of our method is that, like all comparative approaches, it depends on previous experimental observations of similar interactions. The effect of this limitation is reflected in the low number of currently known host–pathogen protein interactions that were recovered by the method. Despite the limited coverage, the

availability of structure enables a more rigorous assessment of the interaction than that allowed by sequence alone (Fig 4.2) [40]. As experimental efforts identify more interactions and further characterize the biology of host and pathogen proteins, the increased number of templates and expanded biological context data will increase the coverage and accuracy of our method, respectively.

Another factor that limits the coverage of our method is that the template identification procedure is primarily restricted to domain-mediated interactions, although peptide-mediated interactions are also known to contribute to protein interaction networks [149, 163]. Peptide motifs that mediate protein interactions are being identified through a combination of computational and experimental methods [150, 220]. Applying these motif-based methods will likely expand the coverage of host–pathogen protein interactions.

4.3.4 Errors in accuracy

The accuracy of interaction prediction methods is most rigorously assessed when large sets of true positive interactions as well as true negative interactions are available. Large negative sets are not yet available for any protein interaction system; however, large positive sets are available for intra-proteomic interactions, particularly for *S. cerevisiae* [66, 115]. In the case of host–pathogen interactions, neither large positive, nor large negative sets of interactions are available. This lack of a true host–pathogen protein interaction set makes direct assessment of our predictions difficult. However, we attempted three different methods to gauge the accuracy of the predictions: (1) comparison to known host–pathogen protein interactions (Table 4.3), (2) comparison to gene expression and essentiality data

describing human and pathogen genes involved in infection (Table 4.4), and (3) analysis of the functional annotations of the host proteins predicted to potentially interact with pathogen proteins (Table 4.5). All three of these indirect methods of assessment suffer from problems: (1) very few host–pathogen protein interactions have been described - in some pathogens none is known; (2) gene expression is, at best, only weakly correlated with physical protein-protein interactions [95, 145]; and (3) automated function annotation methods, responsible for a significant portion of annotations, exhibit various types of error [229].

Although direct assessment of the method is not feasible, it is possible to estimate its accuracy in predicting intra-proteome protein interactions, for which at least a large true positive set is available. The interaction map for *S. cerevisiae* is the most well studied experimentally and so is typically used to benchmark interaction prediction methods. Previous benchmarking of the sequence-based method in *S. cerevisiae* demonstrated that the interactions predicted, using an 80% joint sequence identity threshold, were all correct [247]. We previously benchmarked the structure-based prediction module for protein interactions within *S. cerevisiae* and found that 270 of the 3,387 interactions (8%) we predicted were previously observed experimentally [40]. Although the sequence-based threshold of 80% sequence identity demonstrated a 100% true positive rate, the structure-based method is able to predict below this high threshold. In fact, approximately 90% of the true interactions predicted using the structure-based method fell below the 80% sequence identity threshold. These performance characteristics observed for *S. cerevisiae* represent a baseline for performance in the current application to host–pathogen interactions.

Several factors affect the accuracy of the method. First, errors in each stage of the comparative protein structure modeling procedure can affect the accuracy of the predicted interactions. The magnitude of these errors vary depending on the similarity between the target protein sequence and the template structure. The errors range from local errors in side chain packing and minor shifts in backbone conformation, to more severe errors in loop conformations, target-template alignment, and fold assignment [136]. These errors in the comparative models propagate into the interactions that are predicted [40].

Second, the structural assessment was done using a coarse-grained statistical potential that aims to capture frequencies of residue types in contacts that occur across interfaces [40]. Although the coarse-grained statistical potential allows greater coverage of interaction space compared to sequence-based methods alone, it also suffers from false positives and false negatives [40]. More accurate interaction scoring functions have been developed; however, these require explicit atomic modeling and refinement of candidate protein complexes, which is computationally expensive [72]. Thus, these more accurate methods can serve as computational follow-ups to the genome-scale predictions made here with the faster coarse-grained potential.

Third, the method evaluated interactions as mediated by independent domains. It is possible that even though the interaction between a given pair of domains is favorable, that the protein–protein interaction can not occur due to unfavorable interactions between the rest of the two proteins, which may not be amenable to comparative structure modeling. In a similar vein, we do not attempt to assess the impact of post-translational modifications, which are known to play a significant role in the intricate specificities of biological

networks [162, 196]. Although attempts can be made to take these factors into account, they require explicit modeling which increases the computational expense beyond what is currently feasible for genome-scale prediction.

These first three sources of error affect both intra-proteomic and host–pathogen protein interactions. However, there is an additional source of error that uniquely affects inter-species protein interactions. As the pathogen and host species are both eukaryotic for eight of the ten pathogens studied, many of the predicted potential interactions are between core cellular components, such as translation machinery, metabolic enzymes, and ubiquitin-signaling components (Table 4.5). Addressing the relevance to infection requires *in vivo* testing in an appropriate infection model. Although these interactions could potentially occur if the host and pathogen proteins were to encounter one another, their availability for such an encounter is not guaranteed. It is difficult to capture this “accessibility” information. We have used biological data, such as known exported pathogen proteins and known host tissue targets, but the precise spatial and temporal locations of these proteins are generally difficult to characterize. We expect this last source of false positives to be diminished when the evolutionary distance between pathogen and host is greater, such as between bacterial or viral pathogens and their human hosts.

4.3.5 Other computational methods

Numerous computational methods have been developed to infer physical protein interactions within a species [9, 38, 50, 96, 132, 164, 165, 208, 230, 247]. Several of these methods rely on information such as genomic proximity, gene fission/fusion, phylogenetic

tree similarity, gene co-occurrence, co-localization, co-expression and other features that only make sense or are currently feasible in the context of a single genome. However, comparative approaches that infer interactions based on previously observed interactions remain applicable to host–pathogen protein interactions, including the sequence and structure-based methods we have used here [40, 247]. Other methods that are applicable include those that identify peptide motifs that mediate specific protein interactions [149] or identify pairs of sequence signatures that are found to mediate interactions [208].

One possible extension of the presented method that may improve the accuracy of the predictions is an analysis of the selective pressure on the proposed interacting proteins. If the host and pathogen genes both appear to be under selective pressure, for example human TLR4 and *P. falciparum* TRAP (Fig 4.3), then there may be greater reason to believe that evolution has driven the proteins to interact with one another. In addition to potentially improving the accuracy of the predictions, it may also help highlight those potential interactions that are most relevant to the infection process.

4.3.6 Potential impact

We have developed a computational whole-genome method to study potential host–pathogen protein interactions. Knowledge of host–pathogen interactions is useful in the development of strategies to treat and prevent infectious diseases. These interactions may serve as pharmacologic targets, both for traditional drug discovery efforts aimed at disrupting individual pathogen proteins and for small molecule or antibody inhibitors of protein–protein interactions. The proposed interactions also highlight pathogen proteins

that may be potential immunization targets.

We have also applied our method to ten pathogens involved in human infectious diseases. The predictions are available on the internet at <http://salilab.org/hostpathogen> and can be viewed and filtered according to criteria of interest to an investigator, such as particular host tissues or pathogen life-cycle stages. We hope that the predictions serve the larger biomedical research community in moving towards the ultimate goal of treating infectious diseases, in the “open source” model of the Tropical Disease Initiative [140]. The Tropical Disease Initiative is a decentralized, web-based, community-wide effort where scientists from laboratories, universities, institutes, and corporations can work together for a common cause (<http://www.tropicaldisease.org>).

In closing, as illustrated in Discussion, we expect our method to help provide insight into the basic biology of host–pathogen systems, as well as other inter-species relationships that fall elsewhere on the mutualism – parasitism continuum.

4.4 Materials and Methods

4.4.1 Detecting sequence and structure similarities

The input to the protocol is the host and pathogen protein sequences (CryptoDB [82], GeneDB [84], OrthoMCL-DB [34], PlasmoDB [211], ToxoDB [110], TubercuList <http://genolist.pasteur.fr/TubercuList/>) (Table 4.1). First, protein structure models were calculated for all sequences using MODPIPE, our automated software pipeline for large-scale protein structure modeling [52]. MODPIPE relies on MODELLER [188] for its functionality and calculates comparative models for a large number of sequences using dif-

ferent template structures and sequence-structure alignments. Sequence-structure matches are established using a variety of fold-assignment methods including sequence-sequence [205], profile-sequence [13, 53], and profile-profile alignments [54, 138]. Increased sensitivity of the search for known template structures is achieved by using an E-value threshold of 1.0. Ten models are calculated for each of the sequence-structure matches to achieve a reasonable degree of conformational sampling [188]. The best scoring model for each alignment is then chosen using a statistical potential [198]. Finally, all models generated for a given input sequence are evaluated for the correctness of the fold using a composite model quality criterion that includes the coverage of the model, sequence identity of the sequence-structure alignment, the fraction of gaps in the alignment, the compactness of the model, and statistical potential Z-scores [51, 141, 198]. Only models that are assessed to have the correct fold were included in the final datasets. An important feature of the pipeline is that the validity of sequence-structure relationships is not pre-judged at the fold assignment stage, but is assessed after the construction of the model and its evaluation. This approach enables a thorough exploration of fold assignments, sequence-structure alignments, and conformations, with the aim of finding the model with the best evaluation score. The models have been deposited in our database of comparative models, MODBASE [168] (<http://salilab.org/modbase>) as publicly accessible datasets.

4.4.2 Identifying pairs of proteins with similarity to known interactions and assessing the sequence or structural basis of the potential interactions

Next, the detected structural similarities were used to assign structural domain boundaries to the modeled sequences, according to the SCOP classification system [148]. Pairs of proteins that contained domains classified in the same superfamily as those previously observed to interact (PIBASE [39]) were assessed by alignment of their comparative structure models onto the corresponding domains of the template complexes, and subsequent assessment of the putative interface by a statistical potential [40]. A statistical potential Z-score threshold of -1.7 was used, as previously benchmarked [40]. Interactions predicted based on template complexes formed by protein domains from the same SCOP family were omitted from the analysis, because these predictions primarily consisted of multimeric enzyme complexes formed by both host and pathogen proteins as well as core cellular components, such as ribosome subunits and proteasome subunits.

Sequence profiles, built by MODPIPE, were searched for proteins that participate in binary protein interactions (IntAct; [107]). Host and pathogen sequences were predicted to potentially interact when each aligned to at least 50% of the sequence of members of a template complex with a joint sequence identity of $\sqrt{\text{sequence identity}_1 * \text{sequence identity}_2} \geq 80\%$ [247]. Interactions predicted based on homodimer templates were omitted from the analysis, because the predictions primarily consisted of complexes formed between corresponding core cellular components of host and pathogens (*eg*, histones).

4.4.3 Applying biological and network-level filters

The predicted potential interactions were filtered using biological context and network-level information. The biological context filter was imposed at two levels, individual proteins and their potential interactions (Table 4.6). The host proteins were filtered by expression in tissues known to be targeted by the pathogen (GNF Tissue Atlas [214], Harrison's Principles of Internal Medicine [104]), known expression on cell surface, and known immune system involvement (ENSEMBL [90], Gene Ontology Annotation (GOA) [30], IRIS [1]). The pathogen proteins were filtered by known or predicted secretion, known expression on cell surface, infective life-cycle stage, and functional annotation to defense response mechanisms (PlasmoDB [211], ToxoDB [110], CryptoDB [82], GeneDB [84], references in Table 4.6). The GO terms for human protein involvement in immune system were: GO:0051707, GO:0002376, GO:0006955. The GO terms for pathogen protein involvement in host-pathogen interactions were: GO:00044419 (involved in defense response), GO:0043657 (cellular component: host cell), and GO:0009405 (pathogenesis). Potential interactions between human and pathogen proteins that each met at least 1 biological criteria were considered to pass the biological filter.

The second level of biological filters was applied simultaneously to both human and pathogen proteins, as follows. *M. tuberculosis*: pairs of human proteins expressed in lung tissue or bronchial epithelial cells and pathogen proteins upregulated in granuloma, pericavity, or distal infection sites [174]. *L. major*: pairs of human proteins expressed in skin and pathogen proteins expressed in the promastigote or metacyclic life-cycle stage, human proteins expressed in blood and pathogen proteins expressed in amastigote life-cycle stage.

T. brucei: pairs of human proteins expressed in blood and pathogen proteins expressed in the bloodstream life-cycle stage. *P. falciparum*: pairs of human proteins expressed in erythrocytes and pathogen proteins expressed in the merozoite life-cycle stage, known or predicted to be secreted, and found on the surface of infected erythrocytes; human proteins expressed in liver and pathogen proteins expressed in the sporozoite life-cycle stage. *P. vivax*: pairs of human proteins expressed in erythrocyte and pathogen proteins predicted to be secreted.

The network-level filter removed predictions based on templates that were used for more than 1% of the total number of predictions in each host–pathogen network. This filter was imposed due to the lack of specificity in the predictions based on these highly used templates. On average, 15 interaction templates were removed from each run.

4.4.4 Assessment: Functional overview of predicted complexes

The human proteins predicted to potentially interact with pathogen proteins were analyzed for significant enrichment of gene ontology function terms using GO::TermFinder [26]. The enrichment for a given GO term was computed as the ratio of the fraction of proteins in the predicted set annotated with the GO term to the fraction in the entire human genome. The significance of this enrichment was computed as a p-value with Bonferroni correction for multiple hypothesis testing [206].

4.4.5 Assessment: Comparison to gene expression and essentiality data

Human genes previously observed to be differentially regulated (two-tailed t-test, p-value < 0.05) in macrophages and dendritic cells during infection by *L. major*, *M. tuberculosis*, and *T. gondii* were retrieved from GEO Omnibus (GDS2600) [33, 47]. Lists of *M. tuberculosis* genes that were found to be essential for *in vivo* infection [194], and genes that are upregulated in granuloma, pericavity, or distal lung infection sites compared to *in vitro* conditions [174], were obtained from literature. Differential regulation and essentiality do not imply direct physical interactions with pathogen proteins, but rather some involvement in the infection process. Nevertheless, these studies are useful because they provide large-scale data that is not yet available for direct host–pathogen protein interactions.

Acknowledgments.

We thank J. Cox (UCSF), E. Brown (UCSF), and K. Kim (AECOM) for their help in building the set of known host–pathogen protein interactions. We thank D. Shanmugam (U. Penn) for collating the protein sequences analyzed and U. Pieper (UCSF) for assistance with MODBASE. We thank Tanja Kortemme (UCSF) and members of the Sali laboratory for valuable comments and suggestions. We are also grateful for the support of the US National Institutes of Health grant P01-A135707, US National Science Foundation grant EIA-0325004, Human Frontier Science Program, The Sandler Family Supporting Foundation, Hewlett-Packard, NetApps, IBM, and Intel. FPD acknowledges support from a Howard Hughes Medical Institute predoctoral fellowship.

Pathogen	Pathogen Information	Host Tissues
<i>M. leprae</i>	Transcribed during infection [243]	Skin, lymph node, lung
<i>M. tuberculosis</i>	Differential transcription at granuloma, pericavity, or distal lung infection sites <i>vs in vitro</i> [173, 174]	Lung, bronchial epithelial cells, lymph node
<i>L. major</i>	Metacyclic, procyclic, amastigote stage-specific expression [8, 122]	Skin, whole blood, monocyte[1]
<i>T. brucei</i>	Procyclic, bloodstream stage-specific expression [27]	Erythrocyte[161], whole blood, lymph node, brain, endothelial
<i>T. cruzi</i>	Metacyclic, amastigote, trypomastigotes, epimastigote stage-specific expression [16]	Erythrocyte[161], whole blood, lymph node, skeletal muscle, smooth muscle, cardiac myocytes, endothelial
<i>C. hominis</i>	Sporozoite stage expression [200]	Colorectal adenocarcinoma
<i>C. parvum</i>	Sporozoite stage expression [200]	Colorectal adenocarcinoma
<i>P. falciparum</i>	Ring, trophozoite, schizont, merozoite, gametocyte, sporozoite stage expression [59, 182, 183], expression in infected erythrocyte plasma membrane [60], predicted secreted [86, 135, 193]	Erythrocyte[161], liver, brain, whole blood, endothelial
<i>P. vivax</i>	Predicted secreted [193]	Erythrocyte[161], liver, whole blood
<i>T. gondii</i>	Bradyzoite, tachyzoite, encystation stage expression [175]	Lymph node, skeletal muscle, cardiac myocytes, placenta, brain, lung

Table 4.6: Biological data characterizing host and pathogen proteins. Host tissue expression data was obtained from the GNF Tissue Atlas [214] unless as noted.

Author contributions.

FPD and JHM conceived and designed the experiments. FPD, DB, and NE performed the experiments. FPD, JHM, and AS analyzed the data. FPD and AS wrote the paper.

Chapter 5

Conclusion

I presented a series of computational tools, primarily based on three-dimensional structures, that aim to enrich the functional characterization of protein–protein interactions, particularly those formed between host and pathogen proteins. I will now describe how these tools have been used and how they can be improved. I will close with a general discussion of the role of computation and structure in the investigation of biological networks.

5.1 Summary

First, I described a database, PIBASE, that extracts binary domain interfaces from the protein structure databases, and characterizes them in a number of ways. In addition to the subsequent two studies described in this dissertation, this database is used by a number of other projects as a source of protein interaction information. Analysis of the binding sites in the database led to the development of a method to model the structure of protein complexes that combines docking and comparative modeling [111, 112]. The

binding site information has also been used to analyze the potential functional effects of single nucleotide polymorphisms [103, 106]. Most recently, it has been integrated into an automated protein function annotation pipeline to identify potential binding sites on the structures of unannotated proteins, such as many of those generated by structural genomics consortia [139]. In addition to these large-scale applications, the database tools have been used to compare the geometric and physicochemical properties of newly determined protein complex structures to previously observed complexes (*eg* [73]).

Next, I described a comparative modeling method that uses experimentally determined structures of protein complexes as templates to predict physical protein interactions. In addition to host–pathogen protein interactions, the protocol has been applied to data sets generated by fluid and tissue proteomics consortia.

Lastly, I presented a computational protocol, employing these tools, to predict potential interactions between human and pathogen proteins. The predictions are difficult to assess, since not many host–pathogen interactions have been previously identified. However, analysis of the functional annotation of the predicted complexes suggests that the procedure proposes interactions that are in concordance with what is known about the infection process. In addition to recovering previously observed interactions and invasion mechanisms, the procedure generates testable hypothesis of host–pathogen protein interactions that warrant experimental follow-up.

5.2 Future Directions

The studies I have presented suggest that the use of structure, in combination with genomic and proteomic data, is a valid approach to investigating protein interaction networks. There are a number of ways in which the methods I presented can be improved, both in terms of coverage and accuracy.

5.2.1 Improvements in coverage

Expansion of the template interaction set is essential for improving the coverage of the comparative methods presented here. As more protein structures are determined experimentally, both alone and in complex, and more interactions are identified experimentally, the coverage of the method will subsequently increase. However, specific modifications can be made now that will likely improve the coverage of the methods.

The structure-based method currently requires that domain boundaries and classifications have been assigned to the template protein structures. This causes the current template set to be restricted to those protein structures solved more than 1 - 1.5 years ago, due to the lag time in defining structural domains in new protein structures. The use of automatic domain boundary assignment and classification tools will immediately increase the coverage of the method.

The templates that are currently used, in both the structure- and sequence-based methods, are restricted to domain mediated protein interactions. However, it is known that domain-peptide interactions contribute significantly to protein interaction networks. The expansion of PIBASE to include the structures of domain-peptide interactions will

likely increase the coverage of the method. This expansion will likely require changes to the statistical potential. The potential may have to be rebuilt specifically for these types of interactions, or at the very least, the potential must be benchmarked on known peptide-mediated interactions. Similarly, the use of previously identified sequence motifs, such as the proline-rich PXXP motif recognized by the SH3 domain, will also increase the coverage of the template set in the absence of experimentally determined structures.

Finally, the currently available set of protein complexes are heavily biased towards interactions that occur within species. Although the number of physical host–pathogen protein interactions that have been observed is currently fairly low, they represent a template set that are especially applicable to the prediction of host–pathogen protein interactions. Manual curation, and perhaps automated literature parsing tools, are necessary to convert the microbiology and immunology literature into a format that is easily computable. This expanded template set will likely improve coverage of host–pathogen protein interactions.

As the accuracy of protein structure modeling and docking methods improves, and computational power increases, these methods will become increasingly applicable on a genome-wide scale. The use of *de novo* methods will ultimately overcome the coverage barrier inherent to comparative methods. However, the current accuracy and efficiency of these methods is not suitable for genome-wide predictions of protein interactions.

5.2.2 Improvements in accuracy

Improvements in the assessment protocol can increase the accuracy of the structure-based prediction method. First, the explicit modeling of candidate complexes will likely

improve the accuracy of the predictions. The current protocol uses structural alignments of individual comparative models onto the template complex to transfer pairwise residue interactions from the template onto the candidate complex. Although this procedure is faster than explicit modeling, it can only assess target residue interactions for which corresponding residues interact in the template. Explicit structural models of the candidate complex will allow a more accurate assessment by the current statistical potential.

Second, the use of more fine-grained statistical potentials, such as the DOPE [198] potential, will likely improve the accuracy of the predictions. Although this modification necessitates explicit modeling of candidate complexes, thereby increasing the computational expense, it may provide greater performance, specifically in predicting the specificities of protein–protein interaction networks. These more accurate potentials will also prove useful in optimizing the structural models of candidate complexes.

Third, the use of negative or graded interaction information will likely improve the prediction of interaction specificities. The balance between interactions that occur and those that *do not* occur is what defines the intricate specificities observed in biological networks. Large negative sets of protein interactions known not to occur would be a great help to the prediction of interaction specificities. One specific direction that I believe could improve the accuracy of the predictions is the development of statistical potentials that are explicitly trained on both negative and positive sets of protein interactions. Of course, protein interactions do not occur in a strictly binary fashion, but rather exhibit a continuum of binding affinities. Therefore, rather than truly negative interactions, the use of experimentally determined binding affinities, for example within a family of homologous ligands

and receptors, such as the epidermal growth factor ligand–receptor network, will likely be an important step towards improving the specificity of structure-based predictions. Experimental affinities have proven difficult to recapitulate using only structural information, but recent efforts to parameterize structure-based scoring functions using thermodynamic data have shown promise in assessing protein interface structures [113]. Both the binary and graded binding affinity approaches will require manual curation efforts to extract the appropriate parameters from the literature. In collating a set of true negative interactions it is important to distinguish interactions that do not physically occur, from those that do not occur *in vivo* due to higher levels of regulation, such as sub-cellular localization.

The distinction between interactions that are physically possible and those that occur *in vivo* is an important consideration for interaction prediction efforts. While structure-based methods, like the statistical potential developed here, aim to predict whether a prediction is physically feasible, the ultimate goal is to predict interactions that are relevant *in vivo*. Here, I attempted to bridge this physical–*in vivo* gap using biological context, such as sub-cellular localization and function annotation of *S. cerevisiae* proteins. In the host–pathogen networks, I used biological context such as host protein tissue expression and pathogen life-cycle stage-specific expression. However, these properties represent just a few of the ways in which genes, and the proteins they encode, are regulated. Further developments in high-throughput methods to study the properties of genes, and the proteins they encode, will enable more precise descriptions of the spatio-temporal environments of cellular components. These more precise descriptions will further reduce the gap between predictions of physical interactions and those that are relevant *in vivo*.

5.3 Role of computation and structure in the investigation of inter-specific biomolecular networks

The methods that I have developed here aim to study just one kind of inter-specific biomolecular interaction. Inter-species interactions have been observed between all kinds of biomolecules including small molecules, carbohydrates, nucleic acids, and lipids. Although computational structure analysis tools are most well developed for proteins and nucleic acids, improvements are being made that allow them to address post-translation modifications, such as phosphorylation, and other kinds of biomolecules, such as lipids and carbohydrates. As knowledge of the mechanistic basis of protein interactions continues to grow, and structure becomes integrated with energetics, computational tools will become a more accurate tool for investigating biological networks, including inter-specific networks. In particular, inter-specific biomolecular interactions is an exciting area that has received relatively little attention compared to interactions within species. Characterizing these inter-specific interactions will improve our understanding of medically relevant interactions as well as the mechanisms that underly co-evolutionary processes.

Much of computational biology, and especially structural bioinformatics, is devoted to benchmarking and incremental improvements in basic methods. However, there are also clear areas where these methods can be used to investigate the basic biology that underly human diseases. The results I have presented here will likely have little direct effect on human health, however the approach I have presented is a generalized technique that seems to show promise in identifying inter-specific interactions. The approach I have pre-

sented is only an initial computational attempt at investigating protein–protein interactions that may form in infectious diseases, and as further experimental and computational development enable more accurate understanding of protein interactions, the applicability will increase. As advances in medicinal chemistry increase our ability to target protein–protein interactions with small molecules, these host–pathogen protein interactions will represent a valuable new class of targets for antimicrobial agents.

The goal of reducing the global infectious disease burden will require public health initiatives, such as adequate sewage treatment systems and clean water supplies, and reprise from the political upheaval that threatens much of the world. However, there is no doubt that basic and applied biology is a significant component in improving global health. For example, vaccination strategies led to a practical eradication of smallpox. As biology has moved into an era of parallel data acquisition, computation has come to play an essential role in interpreting these observations. The understanding gained from these new kinds of basic biology data have the potential to contribute towards improvements in human health.

Computation has proven to be a valuable tool for biology. However, without properly posed questions, developing these tools for their own sake is a less than optimal endeavor. In addition, computation can only operate on available experimental data and so is limited in the types of questions that it can address, when used alone. As experimental data is being generated in larger quantities, the need for computational tools to interpret has grown. However, these large datasets have not always produced great insight into meaningful biological questions. Nevertheless, there are still basic biological and biomedical questions that have received relatively little attention, but can be addressed us-

ing experimental measurements and computational methods that are currently feasible. For this reason, I believe computation will continue to be an exciting area than can contribute to our understanding of biology and medicine, most significantly when used as a component of an overall approach that includes experimental observations.

Bibliography

- [1] A. R. Abbas, D. Baldwin, Y. Ma, W. Ouyang, A. Gurney, F. Martin, S. Fong, M. van Lookeren Campagne, P. Godowski, P. M. Williams, A. C. Chan, and H. F. Clark. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun*, 6(4):319–331, Jun 2005.
- [2] C. Abou-Zeid, T. Garbe, R. Lathigra, H. G. Wiker, M. Harboe, G. A. Rook, and D. B. Young. Genetic and immunological analysis of mycobacterium tuberculosis fibronectin-binding proteins. *Infect Immun*, 59(8):2712–2718, Aug 1991.
- [3] A. T. Adai, S. V. Date, S. Wieland, and E. M. Marcotte. Lgl: creating a map of protein function with an algorithm for visualizing very large biological networks. *J Mol Biol*, 340(1):179–190, Jun 2004.
- [4] L. Aguilar, G. Ramirez, C. Valck, M. C. Molina, A. Rojas, W. Schwaeble, V. Ferreira, and A. Ferreira. F(ab')₂ antibody fragments against trypanosoma cruzi calreticulin inhibit its interaction with the first component of human complement. *Biol Res*, 38(2-3):187–195, 2005.

- [5] F. Alber, M. F. Kim, and A. Sali. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure*, 13(3):435–445, Mar 2005.
- [6] B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–294, Feb 1998.
- [7] A. Alcami. Viral mimicry of cytokines, chemokines and their receptors. *Nat Rev Immunol*, 3(1):36–50, Jan 2003.
- [8] R. Almeida, B. J. Gilmartin, S. H. McCann, A. Norrish, A. C. Ivens, D. Lawson, M. P. Levick, D. F. Smith, S. D. Dyall, D. Vetrie, T. C. Freeman, R. M. Coulson, I. Sampaio, H. Schneider, and J. M. Blackwell. Expression profiling of the *Leishmania* life cycle: cDNA arrays identify developmentally regulated genes present but not annotated in the genome. *Mol Biochem Parasitol*, 136(1):87–100, Jul 2004.
- [9] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–5901, Apr 2002.
- [10] P. Aloy and R. B. Russell. Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, 22(10):1317–1321, Oct 2004.
- [11] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332(5):989–998, Oct 2003.
- [12] P. Aloy, B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. B. Russell. Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):2026–2029, Mar 2004.

- [13] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.
- [14] K. Anand, J. Ziebuhr, P. Wadhvani, J. R. Mesters, and R. Hilgenfeld. Coronavirus main proteinase (3clpro) structure: basis for design of anti-sars drugs. *Science*, 300(5626):1763–1767, Jun 2003.
- [15] P. Argos. An investigation of protein subunit and domain interfaces. *Protein Eng*, 2(2):101–113, Jul 1988.
- [16] J. A. Atwood, III, D. B. Weatherly, T. A. Minning, B. Bundy, C. Cavola, F. R. Operdoes, R. Orlando, and R. L. Tarleton. The trypanosoma cruzi proteome. *Science*, 309(5733):473–476, Jul 2005.
- [17] G. D. Bader and C. W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, Jan 2003.
- [18] G. D. Bader, D. Betel, and C. W. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, 31(1):248–250, Jan 2003.
- [19] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O’Donovan, N. Redaschi, and L. S. Yeh. The universal protein resource (uniprot). *Nucleic Acids Res*, 33(Database issue):D154–D159, Jan 2005.
- [20] A. Barragan, F. Brossier, and L. D. Sibley. Transepithelial migration of toxoplasma

gondii involves an interaction of intercellular adhesion molecule 1 (icam-1) with the parasite adhesin mic2. *Cell Microbiol*, 7(4):561–568, Apr 2005.

- [21] P. Bejon, J. Mwacharo, O. Kai, T. Mwangi, P. Milligan, S. Todryk, S. Keating, T. Lang, B. Lowe, C. Gikonyo, C. Molyneux, G. Fegan, S. C. Gilbert, N. Peshu, K. Marsh, and A. V. Hill. A Phase 2b Randomised Trial of the Candidate Malaria Vaccines FP9 ME-TRAP and MVA ME-TRAP among Children in Kenya. *PLoS Clin Trials*, 1(6):e29, Oct 2006. (ENG).
- [22] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 33(Database issue):D34–D38, Jan 2005.
- [23] M. D. Berg, M. V. Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer Verlag, Berlin, 2nd edition, 1998.
- [24] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.
- [25] E. Bornberg-Bauer, F. Beaussart, S. K. Kummerfeld, S. A. Teichmann, and . r. d. Weiner J. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, 62(4):435–445, Feb 2005.
- [26] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec 2004.

- [27] S. Brems, D. L. Guilbride, D. Gundlesdodjir-Planck, C. Busold, V. D. Luu, M. Schanne, J. Hoheisel, and C. Clayton. The transcriptomes of *Trypanosoma brucei* Lister 427 and TREU927 bloodstream and procyclic trypomastigotes. *Mol Biochem Parasitol*, 139(2):163–172, Feb 2005.
- [28] L. S. Burrack and D. E. Higgins. Genomic approaches to understanding bacterial virulence. *Curr Opin Microbiol*, Dec 2006. (ENG).
- [29] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, Jan 2004.
- [30] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue): D262–D266, Jan 2004.
- [31] O. Carugo and P. Argos. Protein-protein crystal-packing contacts. *Protein Sci*, 6(10): 2261–2263, Oct 1997.
- [32] J. M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL Compendium in 2004. *Nucleic Acids Res*, 32(Database issue): D189–D192, Jan 2004.
- [33] D. Chaussabel, R. T. Semnani, M. A. McDowell, D. Sacks, A. Sher, and T. B. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 102(2):672–681, Jul 2003.

- [34] F. Chen, A. J. Mackey, C. Stoeckert, Jr., and D. S. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34 (Database issue):D363–D368, Jan 2006.
- [35] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826, Apr 1986.
- [36] M. I. Colombo, M. G. Gutierrez, and P. S. Romano. The two faces of autophagy: Coxiella and mycobacterium. *Autophagy*, 2(3):162–164, Jul 2006.
- [37] L. L. Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, Feb 1999.
- [38] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324–328, Sep 1998.
- [39] F. P. Davis and A. Sali. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–1907, May 2005.
- [40] F. P. Davis, H. Braberg, M. Y. Shen, U. Pieper, A. Sali, and M. S. Madhusudhan. Protein complex compositions predicted by structural similarity. *Nucleic Acids Res*, 34(10):2943–2952, 2006.
- [41] C. S. de Lima, L. Zulianello, M. A. Marques, H. Kim, M. I. Portugal, S. L. Antunes, F. D. Menozzi, T. H. Ottenhoff, P. J. Brennan, and M. C. Pessolani. Mapping the

laminin-binding and adhesive domain of the cell surface-associated hlp/lbp protein from mycobacterium leprae. *Microbes Infect*, 7(9-10):1097–1109, Jul 2005.

- [42] V. Deretic. Autophagy as an immune defense mechanism. *Curr Opin Immunol*, 18(4):375–382, Aug 2006.
- [43] D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B. T. Chait, A. Sali, and M. P. Rout. Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol*, 2(12):e380, Dec 2004.
- [44] D. Devos, S. Dokudovskaya, R. Williams, F. Alber, N. Eswar, B. T. Chait, M. P. Rout, and A. Sali. Simple fold composition and modular architecture of the nuclear pore complex. *Proc Natl Acad Sci U S A*, 103(7):2172–2177, Feb 2006.
- [45] M. B. Dickman, Y. K. Park, T. Oltersdorf, W. Li, T. Clemente, and R. French. Abrogation of disease development in plants expressing animal antiapoptotic genes. *Proc Natl Acad Sci U S A*, 98(12):6957–6962, Jun 2001.
- [46] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry. Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Res*, 30(1):69–72, Jan 2002.
- [47] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.

- [48] Editorial. Psi-phase 1 and beyond. *Nat Struct Mol Biol*, 11(3):201, Mar 2004.
- [49] A. H. Elcock and J. A. McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A*, 98(6):2990–2994, Mar 2001.
- [50] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 1999.
- [51] D. Eramian, M. Y. Shen, D. Devos, F. Melo, A. Sali, and M. A. Marti-Renom. A composite score for predicting errors in protein structure models. *Protein Sci*, 15(7):1653–1666, Jul 2006.
- [52] N. Eswar, B. John, N. Mirkovic, A. Fiser, V. A. Ilyin, U. Pieper, A. C. Stuart, M. A. Marti-Renom, M. S. Madhusudhan, B. Yerkovich, and A. Sali. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res*, 31(13):3375–3380, Jul 2003.
- [53] N. Eswar, M. S. Madhusudhan, M. A. Marti-Renom, and A. Sali. BUILD_PROFILE: A module for calculating sequence profiles in MODELLER. 2005. URL <http://www.salilab.org/modeller>.
- [54] N. Eswar, M. S. Madhusudhan, M. A. Marti-Renom, and A. Sali. PROFILE_SCAN: A module for fold-assignment using profile-profile scanning in MODELLER. 2005. URL <http://www.salilab.org/modeller>.

- [55] F. Fabiola and M. S. Chapman. Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure*, 13(3):389–400, Mar 2005.
- [56] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, HP Labs, Palo Alto, CA, USA, Jan 2003. URL www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.
- [57] J. Fiaux, E. B. Bertelsen, A. L. Horwich, and K. Wuthrich. Nmr analysis of a 900k groel groes complex. *Nature*, 418(6894):207–211, Jul 2002.
- [58] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, Jul 1989.
- [59] L. Florens, M. P. Washburn, J. D. Raine, R. M. Anthony, M. Grainger, J. D. Haynes, J. K. Moch, N. Muster, J. B. Sacci, D. L. Tabb, A. A. Witney, D. Wolters, Y. Wu, M. J. Gardner, A. A. Holder, R. E. Sinden, J. R. Yates, and D. J. Carucci. A proteomic view of the plasmodium falciparum life cycle. *Nature*, 419(6906):520–526, Oct 2002.
- [60] L. Florens, X. Liu, Y. Wang, S. Yang, O. Schwartz, M. Peglar, D. J. Carucci, . r. d. Yates JR, and Y. Wub. Proteomics approach reveals novel proteins on the surface of malaria-infected erythrocytes. *Mol Biochem Parasitol*, 135(1):1–11, May 2004.
- [61] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans Math Software*, 3:209–226, 1977.
- [62] H. Fu. *Protein-Protein Interactions: methods and applications*. Humana Press, 2004.

- [63] G. A. Garcia, L. G. Joensen, J. Bua, N. Ainciart, S. J. Perry, and A. M. Ruiz. Trypanosoma cruzi: molecular identification and characterization of new members of the tc13 family. description of the interaction between the tc13 antigen from tulahuen strain and the second extracellular loop of the beta(1)-adrenergic receptor. *Exp Parasitol*, 103(3-4):112–119, Mar 2003.
- [64] B. E. Garcia-Perez, R. Mondragon-Flores, and J. Luna-Herrera. Internalization of mycobacterium tuberculosis by macropinocytosis in non-phagocytic cells. *Microb Pathog*, 35(2):49–55, Aug 2003.
- [65] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, Jan 2002.
- [66] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dumpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey

- reveals modularity of the yeast cell machinery. *Nature*, Jan 2006. (ENG).
- [67] A. J. Gehring, K. M. Dobos, J. T. Belisle, C. V. Harding, and W. H. Boom. Mycobacterium tuberculosis lprg (rv1411c): a novel tlr-2 ligand that inhibits human macrophage class ii mhc antigen processing. *J Immunol*, 173(4):2660–2668, Aug 2004.
- [68] S. Ghaemmaghami, W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O’Shea, and J. S. Weissman. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–741, Oct 2003.
- [69] R. Giordano, R. Chammas, S. S. Veiga, W. Colli, and M. J. Alves. An acidic component of the heterogeneous tc-85 protein family from the surface of trypanosoma cruzi is a laminin binding glycoprotein. *Mol Biochem Parasitol*, 65(1):85–94, May 1994.
- [70] V. K. Goel, X. Li, H. Chen, S. C. Liu, A. H. Chishti, and S. S. Oh. Band 3 is a host receptor binding merozoite surface protein 1 during the plasmodium falciparum invasion of erythrocytes. *Proc Natl Acad Sci U S A*, 100(9):5164–5169, Apr 2003.
- [71] S. P. Goff. Genetic control of retrovirus susceptibility in mammalian cells. *Annu Rev Genet*, 38:61–85, 2004.
- [72] J. J. Gray. High-resolution protein-protein docking. *Curr Opin Struct Biol*, 16(2):183–193, Apr 2006.
- [73] F. Gruswitz, . r. d. O’Connell J, and R. M. Stroud. Inhibitory complex of the transmembrane ammonia channel, amtb, and the cytosolic regulatory protein, glnk, at 1.96 a. *Proc Natl Acad Sci U S A*, 104(1):42–47, Jan 2007.

- [74] I. Guerin and C. de Chastellier. Pathogenic mycobacteria disrupt the macrophage actin filament network. *Infect Immun*, 68(5):2655–2662, May 2000.
- [75] R. D. Gupta, T. Krause-Ihle, B. Bergmann, I. B. Muller, A. R. Khomutov, S. Muller, R. D. Walter, and K. Luersen. 3-aminooxy-1-aminopropane and derivatives have an antiproliferative effect on cultured plasmodium falciparum by decreasing intracellular polyamine concentrations. *Antimicrob Agents Chemother*, 49(7):2857–2864, Jul 2005.
- [76] M. G. Gutierrez, S. S. Master, S. B. Singh, G. A. Taylor, M. I. Colombo, and V. Deretic. Autophagy is a defense mechanism inhibiting bcg and mycobacterium tuberculosis survival in infected macrophages. *Cell*, 119(6):753–766, Dec 2004.
- [77] A. E. Guttmacher and F. S. Collins. Realizing the promise of genomics in biomedical research. *JAMA*, 294(11):1399–1402, Sep 2005.
- [78] U. Hamann. Merkmalsbestand und verwandtschaftsbeziehungen der farinosae. ein beitrag zum system der monokotyledonen. *Willdenowia*, 2:639–768, 1961.
- [79] J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, Jul 2004.
- [80] D. Hans, P. Pattnaik, A. Bhattacharyya, A. R. Shakri, S. S. Yazdani, M. Sharma, H. Choe, M. Farzan, and C. E. Chitnis. Mapping binding residues in the plasmodium vivax domain that binds duffy antigen during red cell invasion. *Mol Microbiol*, 55(5):1423–1434, Mar 2005.

- [81] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7(11):120, 2006.
- [82] M. Heiges, H. Wang, E. Robinson, C. Aurrecochea, X. Gao, N. Kaluskar, P. Rhodes, S. Wang, C. Z. He, Y. Su, J. Miller, E. Kraemer, and J. C. Kissinger. CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res*, 34(Database issue):D419–D422, Jan 2006.
- [83] K. Henrick and J. M. Thornton. PQS: a protein quaternary structure file server. *Trends Biochem Sci*, 23(9):358–361, Sep 1998.
- [84] C. Hertz-Fowler, C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A. C. Ivens, M. A. Rajandream, and B. Barrell. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*, 32(Database issue):D339–D343, Jan 2004.
- [85] A. V. Hill. Pre-erythrocytic malaria vaccines: towards greater efficacy. *Nat Rev Immunol*, 6(1):21–32, Jan 2006.
- [86] N. L. Hiller, S. Bhattacharjee, C. van Ooij, K. Liolios, T. Harrison, C. Lopez-Estrano, and K. Haldar. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science*, 306(5703):1934–1937, Dec 2004.
- [87] B. Hitz and B. Honig. Spin-pp: Surface properties of interfaces - protein protein interfaces. 1999. URL <http://trantor.bioc.columbia.edu/cgi-bin/SPIN/>.
- [88] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar,

- P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, Jan 2002.
- [89] H. Hossain, S. Tchatalbachev, and T. Chakraborty. Host gene expression profiling in pathogen-host interactions. *Curr Opin Immunol*, 18(4):422–429, Aug 2006.
- [90] T. J. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–D617, Jan 2007.
- [91] E. G. Huizinga, S. Tsuji, R. A. Romijn, M. E. Schiphorst, P. G. de Groot, J. J. Sixma,

- and P. Gros. Structures of glycoprotein α 1 and its complex with von willebrand factor α 1 domain. *Science*, 297(5584):1176–1179, Aug 2002.
- [92] H. Hurd, K. M. Grant, and S. C. Arambage. Apoptosis-like death as a feature of malaria infection in mosquitoes. *Parasitology*, 132 Suppl:S33–S47, 2006.
- [93] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, Apr 2001.
- [94] J. Janin, S. Miller, and C. Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol*, 204(1):155–164, Nov 1988.
- [95] R. Jansen, N. Lan, J. Qian, and M. Gerstein. Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*, 2(2):71–81, 2002.
- [96] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, Oct 2003.
- [97] K. Joiner, S. Hieny, L. V. Kirchhoff, and A. Sher. gp72, the 72 kilodalton glycoprotein, is the membrane acceptor site for c3 on trypanosoma cruzi epimastigotes. *J Exp Med*, 161(5):1196–1212, May 1985.
- [98] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, Feb 1996.

- [99] S. Jones, A. Marin, and J. M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng*, 13(2):77–82, Feb 2000.
- [100] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983.
- [101] W. Kabsch, H. G. Mannherz, D. Suck, E. F. Pai, and K. C. Holmes. Atomic structure of the actin:dnase i complex. *Nature*, 347(6288):37–44, Sep 1990.
- [102] S. Kahn, M. Wleklinski, A. Aruffo, A. Farr, D. Coder, and M. Kahn. Trypanosoma cruzi amastigote adhesion to macrophages is facilitated by the mannose receptor. *J Exp Med*, 182(5):1243–1258, Nov 1995.
- [103] R. Karchin, L. Kelly, and A. Sali. Improving functional annotation of non-synonomous snps with information theory. *Pac Symp Biocomput*, pages 397–408, 2005.
- [104] D. L. Kasper, E. Braunwald, A. Fauci, S. Hauser, D. Longo, and J. L. Jameson. *Harrison's Principles of Internal Medicine 16th Edition*. McGraw-Hill Professional, 2004. ISBN 0071402357.
- [105] P. Kellam. Attacking pathogens through their hosts. *Genome Biol*, 7(1):201, 2006.
- [106] L. Kelly, R. Karchin, and A. Sali. Protein interactions and disease phenotypes in the abc transporter superfamily. *Pac Symp Biocomput*, 12:51–63, 2007.

- [107] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct—open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–D565, Jan 2007.
- [108] O. Keskin, C. J. Tsai, H. Wolfson, and R. Nussinov. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, 13(4):1043–1055, Apr 2004.
- [109] F. Kierszenbaum, J. J. Wirth, P. P. McCann, and A. Sjoerdsma. Impairment of macrophage function by inhibitors of ornithine decarboxylase activity. *Infect Immun*, 55(10):2461–2464, Oct 1987.
- [110] J. C. Kissinger, B. Gajria, L. Li, I. T. Paulsen, and D. S. Roos. ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res*, 31(1):234–236, Jan 2003.
- [111] D. Korkin, F. P. Davis, and A. Sali. Localization of protein-binding sites within families of proteins. *Protein Sci*, 14(9):2350–2360, Sep 2005.
- [112] D. Korkin, F. P. Davis, F. Alber, T. Luong, M. Y. Shen, V. Lucic, M. B. Kennedy, and A. Sali. Structural modeling of protein interactions by analogy: application to psd-95. *PLoS Comput Biol*, 2(11):e153, Nov 2006.
- [113] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A*, 99(22):14116–14121, Oct 2002.

- [114] G. Krishnegowda, A. M. Hajjar, J. Zhu, E. J. Douglass, S. Uematsu, S. Akira, A. S. Woods, and D. C. Gowda. Induction of proinflammatory responses in macrophages by the glycosylphosphatidylinositols of plasmodium falciparum: cell signaling receptors, glycosylphosphatidylinositol (gpi) structural requirement, and regulation of gpi activity. *J Biol Chem*, 280(9):8606–8616, Mar 2005.
- [115] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O’Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, Mar 2006.
- [116] M. Krych-Goldberg, J. M. Moulds, and J. P. Atkinson. Human complement receptor type 1 (cr1) binds to a major malarial adhesin. *Trends Mol Med*, 8(11):531–537, Nov 2002.
- [117] T. G. Ksiazek, D. Erdman, C. S. Goldsmith, S. R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J. A. Comer, W. Lim, P. E. Rollin, S. F. Dowell, A. E. Ling, C. D. Humphrey, W. J. Shieh, J. Guarner, C. D. Paddock, P. Rota, B. Fields, J. DeRisi,

- J. Y. Yang, N. Cox, J. M. Hughes, J. W. LeDuc, W. J. Bellini, and L. J. Anderson. A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med*, 348(20):1953–1966, May 2003.
- [118] R. Lanzillotti and T. L. Coetzer. The 10 kda domain of human erythrocyte protein 4.1 binds the plasmodium falciparum eba-181 protein. *Malar J*, 5:100, 2006.
- [119] A. Lau, K. M. Swinbank, P. S. Ahmed, D. L. Taylor, S. P. Jackson, G. C. Smith, and M. J. O’Connor. Suppression of HIV-1 infection by a small molecule inhibitor of the ATM kinase. *Nat Cell Biol*, 7(5):493–500, May 2005.
- [120] M. Lauwereys, M. A. Ghahroudi, A. Desmyter, J. Kinne, W. Holzer, E. D. Genst, L. Wyns, and S. Muyldermans. Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J*, 17(13):3512–3520, Jul 1998.
- [121] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, Nov 2004.
- [122] K. Leifso, G. Cohen-Freue, N. Dogra, A. Murray, and W. R. McMaster. Genomic and proteomic expression analysis of leishmania promastigote and amastigote life stages: The leishmania genome is constitutively expressed. *Mol Biochem Parasitol*, 152(1):35–46, Mar 2007. (ENG).
- [123] J. E. Lincoln, C. Richael, B. Overduin, K. Smith, R. Bostock, and D. G. Gilchrist. Expression of the antiapoptotic baculovirus p35 gene in tomato blocks programmed cell death and provides broad-spectrum resistance to disease. *Proc Natl Acad Sci U S A*, 99(23):15217–15221, Nov 2002.

- [124] M. Lopez, L. M. Sly, Y. Luu, D. Young, H. Cooper, and N. E. Reiner. The 19-kda mycobacterium tuberculosis protein induces macrophage apoptosis through toll-like receptor-2. *J Immunol*, 170(5):2409–2416, Mar 2003.
- [125] L. Lu, H. Lu, and J. Skolnick. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–364, Nov 2002.
- [126] L. Lu, A. K. Arakaki, H. Lu, and J. Skolnick. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the saccharomyces cerevisiae proteome. *Genome Res*, 13(6A):1146–1154, Jun 2003.
- [127] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 15(7):945–953, Jul 2005.
- [128] M. S. Madhusudhan, M. A. Marti-Renom, N. Eswar, B. John, U. Pieper, R. Karchin, M. Y. Shen, and A. Sali. *The Proteomics Protocols Handbook. Edited by Walker JM.*, chapter Comparative Protein Structure Modeling, pages 831–860. Humana Press Inc, 2005.
- [129] M. H. Magdesian, R. Giordano, H. Ulrich, M. A. Juliano, L. Juliano, R. I. Schumacher, W. Colli, and M. J. Alves. Infection by trypanosoma cruzi. identification of a parasite ligand and its host cell receptor. *J Biol Chem*, 276(22):19382–19389, Jun 2001.
- [130] C. Magowan, W. Nunomura, K. L. Waller, J. Yeung, J. Liang, H. V. Dort, P. S. Low, R. L. Coppel, and N. Mohandas. Plasmodium falciparum histidine-rich protein 1 as-

sociates with the band 3 binding domain of ankyrin in the infected red cell membrane.

Biochim Biophys Acta, 1502(3):461–470, Nov 2000.

- [131] A. G. Maier, M. T. Duraisingh, J. C. Reeder, S. S. Patel, J. W. Kazura, P. A. Zimmerman, and A. F. Cowman. Plasmodium falciparum erythrocyte invasion through glycophorin c and selection for gerbich negativity in human populations. *Nat Med*, 9(1):87–92, Jan 2003.
- [132] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, Jul 1999.
- [133] N. J. Marianayagam, M. Sunde, and J. M. Matthews. The power of two: protein dimerization in biology. *Trends Biochem Sci*, 29(11):618–625, Nov 2004.
- [134] M. Marroquin-Quelopana, J. r. Oyama S, T. A. Pertinhez, A. Spisni, M. A. Juliano, L. Juliano, W. Colli, and M. J. Alves. Modeling the trypanosoma cruzi tc85-11 protein and mapping the laminin-binding site. *Biochem Biophys Res Commun*, 325(2):612–618, Dec 2004.
- [135] M. Marti, R. T. Good, M. Rug, E. Knuepfer, and A. F. Cowman. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science*, 306(5703):1930–1933, Dec 2004.
- [136] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

- [137] M. A. Marti-Renom, V. A. Ilyin, and A. Sali. Dbali: a database of protein structure alignments. *Bioinformatics*, 17(8):746–747, Aug 2001.
- [138] M. A. Marti-Renom, M. S. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Sci*, 13(4):1071–1087, Apr 2004.
- [139] M. A. Marti-Renom, A. Rossi, F. Al-Shahrour, F. P. Davis, U. Pieper, J. Dopazo, and A. Sali. The annolite and annolyze programs for comparative annotation of protein structures. *BMC Bioinformatics*, 2007 (in press).
- [140] S. M. Maurer, A. Rai, and A. Sali. Finding cures for tropical diseases: is open source an answer? *PLoS Med*, 1(3):e56, Dec 2004.
- [141] F. Melo, R. Sanchez, and A. Sali. Statistical potentials for fold assessment. *Protein Sci*, 11(2):430–448, Feb 2002.
- [142] K. Mizuguchi, C. M. Deane, T. L. Blundell, M. S. Johnson, and J. P. Overington. Joy: protein sequence-structure representation and analysis. *Bioinformatics*, 14(7):617–623, 1998.
- [143] F. P. Mockenhaupt, J. P. Cramer, L. Hamann, M. S. Stegemann, J. Eckert, N. R. Oh, R. N. Otchwemah, E. Dietz, S. Ehrhardt, N. W. Schroder, U. Bienzle, and R. R. Schumann. Toll-like receptor (TLR) polymorphisms in African children: Common TLR-4 variants predispose to severe malaria. *Proc Natl Acad Sci U S A*, 103(1):177–182, Jan 2006.
- [144] V. S. Moorthy, E. B. Imoukhuede, P. Milligan, K. Bojang, S. Keating, P. Kaye,

- M. Pinder, S. C. Gilbert, G. Walraven, B. M. Greenwood, and A. S. Hill. A randomised, double-blind, controlled vaccine efficacy trial of dna/mva me-trap against malaria infection in gambian adults. *PLoS Med*, 1(2):e33, Nov 2004.
- [145] R. Mrowka, A. Patzak, and H. Herzel. Is there a bias in proteome research? *Genome Res*, 11(12):1971–1973, Dec 2001.
- [146] S. L. Mueller-Ortiz, A. R. Wanger, and S. J. Norris. Mycobacterial protein hbha binds human complement component c3. *Infect Immun*, 69(12):7501–7511, Dec 2001.
- [147] S. Munter, M. Way, and F. Frischknecht. Signaling during pathogen infection. *Sci STKE*, 2006(335):re5, May 2006.
- [148] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.
- [149] V. Neduva and R. B. Russell. Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol*, 17(5):465–471, Oct 2006.
- [150] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12):e405, Dec 2005.
- [151] I. Neira, F. A. Silva, M. Cortez, and N. Yoshida. Involvement of trypanosoma cruzi metacyclic trypomastigote surface molecule gp82 in adhesion to gastric mucin and invasion of epithelial cells. *Infect Immun*, 71(1):557–561, Jan 2003.

- [152] R. Nussinov and H. J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A*, 88(23):10495–10499, Dec 1991.
- [153] T. M. Nye, C. Berzuini, W. R. Gilks, M. M. Babu, and S. A. Teichmann. Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001, Apr 2005.
- [154] Y. Ofran and B. Rost. Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–387, Jan 2003.
- [155] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, Aug 1997.
- [156] P. A. Orlandi, F. W. Klotz, and J. D. Haynes. A malaria invasion receptor, the 175-kilodalton erythrocyte binding antigen of plasmodium falciparum recognizes the terminal neu5ac(alpha 2-3)gal- sequences of glycophorin a. *J Cell Biol*, 116(4):901–909, Feb 1992.
- [157] A. R. Ortiz, C. E. Strauss, and O. Olmea. Mammoth (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, 11(11):2606–2621, Nov 2002.
- [158] A. Osterman, N. V. Grishin, L. N. Kinch, and M. A. Phillips. Formation of functional cross-species heterodimers of ornithine decarboxylase. *Biochemistry*, 33(46):13662–13667, Nov 1994.

- [159] K. C. Pandey, N. Singh, S. Arastu-Kapur, M. Bogyo, and P. J. Rosenthal. Falstatin, a cysteine protease inhibitor of plasmodium falciparum, facilitates erythrocyte invasion. *PLoS Pathog*, 2(11):e117, Nov 2006.
- [160] J. Park, M. Lappe, and S. A. Teichmann. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the pdb and yeast. *J Mol Biol*, 307(3):929–938, Mar 2001.
- [161] E. M. Pasini, M. Kirkegaard, P. Mortensen, H. U. Lutz, A. W. Thomas, and M. Mann. In-depth analysis of the membrane and cytosolic proteome of red blood cells. *Blood*, 108(3):791–801, Aug 2006.
- [162] T. Pawson. Specificity in signal transduction: from phosphotyrosine-sh2 domain interactions to complex cellular systems. *Cell*, 116(2):191–203, Jan 2004.
- [163] T. Pawson and P. Nash. Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452, Apr 2003.
- [164] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–614, Sep 2001.
- [165] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, Apr 1999.
- [166] N. Perrimon and B. Mathey-Prevot. Applications of high-throughput rna interference

- screens to problems in cell and developmental biology. *Genetics*, 175(1):7–16, Jan 2007.
- [167] U. Pieper, N. Eswar, H. Braberg, M. S. Madhusudhan, F. P. Davis, A. C. Stuart, N. Mirkovic, A. Rossi, M. A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C. C. Huang, T. E. Ferrin, and A. Sali. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 32 Database issue:D217–D222, Jan 2004.
- [168] U. Pieper, N. Eswar, F. P. Davis, H. Braberg, M. S. Madhusudhan, A. Rossi, M. Marti-Renom, R. Karchin, B. M. Webb, D. Eramian, M. Y. Shen, L. Kelly, F. Melo, and A. Sali. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*, 34(Database issue):D291–D295, Jan 2006.
- [169] C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71, 2002.
- [170] E. Prieur, S. C. Gilbert, J. Schneider, A. C. Moore, E. G. Sheu, N. Goonetilleke, K. J. Robson, and A. V. Hill. A plasmodium falciparum candidate vaccine based on a six-antigen polyprotein encoded by recombinant poxviruses. *Proc Natl Acad Sci U S A*, 101(1):290–295, Jan 2004.
- [171] R. P. Prioli, I. Rosenberg, and M. E. Pereira. Specific inhibition of trypanosoma cruzi neuraminidase by the human plasma glycoprotein 'cruzin'. *Proc Natl Acad Sci U S A*, 84(10):3097–3101, May 1987.

- [172] R. P. Prioli, I. Rosenberg, S. Shivakumar, and M. E. Pereira. Specific binding of human plasma high density lipoprotein (cruzin) to trypanosoma cruzi. *Mol Biochem Parasitol*, 28(3):257–263, Apr 1988.
- [173] H. Rachman, M. Strong, U. Schaible, J. Schuchhardt, K. Hagens, H. Mollenkopf, D. Eisenberg, and S. H. Kaufmann. Mycobacterium tuberculosis gene expression profiling within the context of protein networks. *Microbes Infect*, 8(3):747–757, Mar 2006.
- [174] H. Rachman, M. Strong, T. Ulrichs, L. Grode, J. Schuchhardt, H. Mollenkopf, G. A. Kosmiadi, D. Eisenberg, and S. H. Kaufmann. Unique transcriptome signature of mycobacterium tuberculosis in pulmonary tuberculosis. *Infect Immun*, 74(2):1233–1242, Feb 2006.
- [175] J. R. Radke, M. S. Behnke, A. J. Mackey, J. B. Radke, D. S. Roos, and M. W. White. The transcriptome of toxoplasma gondii. *BMC Biol*, 3:26, 2005.
- [176] A. Ragas, L. Roussel, G. Puzo, and M. Riviere. The mycobacterium tuberculosis cell-surface glycoprotein apa as a potential adhesin to colonize target cells via the innate immune system pulmonary c-type lectin surfactant protein a. *J Biol Chem*, 282(8):5133–5142, Feb 2007.
- [177] A. Ramos, M. S. Remedi, C. Sanchez, G. Bonacci, M. A. Vides, and G. Chiabrand. Inhibitory effects of human alpha 2-macroglobulin on trypanosoma cruzi epimastigote proteinases. *Acta Trop*, 68(3):327–337, Dec 1997.

- [178] A. M. Ramos, V. G. Duschak, N. M. G. de Burgos, M. Barboza, M. S. Remedi, M. A. Vides, and G. A. Chiabrando. Trypanosoma cruzi: cruzipain and membrane-bound cysteine proteinase isoform(s) interacts with human alpha(2)-macroglobulin and pregnancy zone protein. *Exp Parasitol*, 100(2):121–130, Feb 2002.
- [179] T. J. Richmond and F. M. Richards. Packing of alpha-helices: geometrical constraints and contact areas. *J Mol Biol*, 119(4):537–555, Mar 1978.
- [180] C. H. Robert and P. S. Ho. Significance of bound water to local chain conformations in protein crystals. *Proc Natl Acad Sci U S A*, 92(16):7600–7604, Aug 1995.
- [181] C. H. Robert and J. Janin. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J Mol Biol*, 283(5):1037–1047, Nov 1998.
- [182] K. G. L. Roch, Y. Zhou, P. L. Blair, M. Grainger, J. K. Moch, J. D. Haynes, P. D. L. Vega, A. A. Holder, S. Batalov, D. J. Carucci, and E. A. Winzeler. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639):1503–1508, Sep 2003.
- [183] K. G. L. Roch, J. R. Johnson, L. Florens, Y. Zhou, A. Santosyan, M. Grainger, S. F. Yan, K. C. Williamson, A. A. Holder, D. J. Carucci, . r. d. Yates JR, and E. A. Winzeler. Global analysis of transcript and protein levels across the plasmodium falciparum life cycle. *Genome Res*, 14(11):2308–2318, Nov 2004.
- [184] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, M. H. Chen, S. Tong, A. Tamin, L. Lowe,

- M. Frace, J. L. DeRisi, Q. Chen, D. Wang, D. D. Erdman, T. C. Peret, C. Burns, T. G. Ksiazek, P. E. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen-Rasmussen, R. Fouchier, S. Gunther, A. D. Osterhaus, C. Drosten, M. A. Pallansch, L. J. Anderson, and W. J. Bellini. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300(5624):1394–1399, May 2003.
- [185] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.
- [186] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–324, Jun 2004.
- [187] A. Sali. Nih workshop on structural proteomics of biological complexes. *Structure (Camb)*, 11(9):1043–1047, Sep 2003.
- [188] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, Dec 1993.

- [189] A. Sali and J. P. Overington. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci*, 3(9):1582–1596, Sep 1994.
- [190] A. Sali, R. Glaeser, T. Earnest, and W. Baumeister. From words to literature in structural proteomics. *Nature*, 422(6928):216–225, Mar 2003.
- [191] L. Salwinski and D. Eisenberg. Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, 13(3):377–382, Jun 2003.
- [192] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004.
- [193] T. J. Sargeant, M. Marti, E. Caler, J. M. Carlton, K. Simpson, T. P. Speed, and A. F. Cowman. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol*, 7(2):R12, 2006.
- [194] C. M. Sassetti and E. J. Rubin. Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A*, 100(22):12989–12994, Oct 2003.
- [195] A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*, 29(14):2994–3005, Jul 2001.

- [196] B. T. Seet, I. Dikic, M. M. Zhou, and T. Pawson. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol*, 7(7):473–483, Jul 2006.
- [197] M. Shakibaei and U. Frevert. Dual interaction of the malaria circumsporozoite protein with the low density lipoprotein receptor-related protein (lrp) and heparan sulfate proteoglycans. *J Exp Med*, 184(5):1699–1711, Nov 1996.
- [198] M. Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11):2507–2524, Nov 2006.
- [199] J. P. Siano, K. K. Grady, P. Millet, and T. M. Wick. Short report: Plasmodium falciparum: cytoadherence to alpha(v)beta3 on human microvascular endothelial cells. *Am J Trop Med Hyg*, 59(1):77–79, Jul 1998.
- [200] A. Z. Siddiki and J. M. Wastling. Proteome analysis of cryptosporidium. In *2005 Bioscience meeting abstracts*, page 0408, 2005.
- [201] K. Silamut, N. H. Phu, C. Whitty, G. D. Turner, K. Louwrier, N. T. Mai, J. A. Simpson, T. T. Hien, and N. J. White. A quantitative analysis of the microvascular sequestration of malaria parasites in the human brain. *Am J Pathol*, 155(2):395–410, Aug 1999.
- [202] B. K. Sim, C. E. Chitnis, K. Wasniowska, T. J. Hadley, and L. H. Miller. Receptor and ligand domains for invasion of erythrocytes by plasmodium falciparum. *Science*, 264(5167):1941–1944, Jun 1994.
- [203] S. Singh, A. Mukherjee, A. R. Khomutov, L. Persson, O. Heby, M. Chatterjee,

- and R. Madhubala. Antileishmanial effect of 3-aminooxy-1-aminopropane is due to polyamine depletion. *Antimicrob Agents Chemother*, 51(2):528–534, Feb 2007.
- [204] S. B. Singh, A. S. Davis, G. A. Taylor, and V. Deretic. Human irgm induces autophagy to eliminate intracellular mycobacteria. *Science*, 313(5792):1438–1441, Sep 2006.
- [205] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.
- [206] R. R. Sokal and F. J. Rohlf. *Biometry*. W. H. Freeman, third edition, 1995. ISBN 0716724111.
- [207] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–12128, Oct 2003.
- [208] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, Aug 2001.
- [209] C. E. Stebbins and J. E. Galan. Modulation of host signaling by a bacterial mimic: structure of the salmonella effector sptp bound to rac1. *Mol Cell*, 6(6):1449–1460, Dec 2000.
- [210] C. E. Stebbins and J. E. Galan. Structural mimicry in bacterial virulence. *Nature*, 412(6848):701–705, Aug 2001.
- [211] C. Stoeckert, Jr., S. Fischer, J. C. Kissinger, M. Heiges, C. Aurrecochea, B. Gajria, and D. S. Roos. PlasmoDB v5: new looks, new genomes. *Trends Parasitol*, 22(12): 543–546, Dec 2006. (ENG).

- [212] V. Stoka, M. Nycander, B. Lenarcic, C. Labriola, J. J. Cazzulo, I. Bjork, and V. Turk. Inhibition of cruzipain, the major cysteine proteinase of the protozoan parasite, *trypanosoma cruzi*, by proteinase inhibitors of the cystatin superfamily. *FEBS Lett*, 370 (1-2):101–104, Aug 1995.
- [213] A. C. Stuart, V. A. Ilyin, and A. Sali. Ligbase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics*, 18 (1):200–201, Jan 2002.
- [214] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, Apr 2004.
- [215] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.
- [216] J. E. Thole, R. Schoningh, A. A. Janson, T. Garbe, Y. E. Cornelisse, J. E. Clark-Curtiss, A. H. Kolk, T. H. Ottenhoff, R. R. D. Vries, and C. Abou-Zeid. Molecular and immunological analysis of a fibronectin-binding protein antigen secreted by *mycobacterium leprae*. *Mol Microbiol*, 6(2):153–163, Jan 1992.
- [217] I. Tirosh and N. Barkai. Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics*, 6(1):40, 2005.
- [218] A. R. Todeschini, M. F. Girard, J. M. Wieruszkeski, M. P. Nunes, G. A. DosReis, L. Mendonca-Previato, and J. O. Previato. trans-sialidase from *trypanosoma cruzi*

- binds host t-lymphocytes in a lectin manner. *J Biol Chem*, 277(48):45962–45968, Nov 2002.
- [219] A. H. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, Dec 2001.
- [220] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324, Jan 2002.
- [221] M. Topf and A. Sali. Combining electron microscopy and comparative protein structure modeling. *Curr Opin Struct Biol*, 15(5):578–585, Oct 2005.
- [222] M. Topf, M. L. Baker, B. John, W. Chiu, and A. Sali. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J Struct Biol*, 149(2):191–203, Feb 2005.
- [223] L. Troeberg, R. N. Pike, R. E. Morty, R. K. Berry, T. H. Coetzer, and J. D. Lonsdale-Eccles. Proteases from trypanosoma brucei brucei. purification, characterisation and interactions with host regulatory molecules. *Eur J Biochem*, 238(3):728–736, Jun 1996.

- [224] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 260(4):604–620, Jul 1996.
- [225] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, Feb 2000.
- [226] P. Uetz, Y. A. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, and J. Haas. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–242, Jan 2006.
- [227] A. G. Uren, K. O’Rourke, L. A. Aravind, M. T. Pisabarro, S. Seshagiri, E. V. Koonin, and V. M. Dixit. Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in malt lymphoma. *Mol Cell*, 6(4):961–967, Oct 2000.
- [228] W. S. Valdar and J. M. Thornton. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol*, 313(2):399–416, Oct 2001.
- [229] A. Valencia. Automatic annotation of protein function. *Curr Opin Struct Biol*, 15(3):267–274, Jun 2005.
- [230] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–373, Jun 2002.

- [231] L. Vanhamme, F. Paturiaux-Hanocq, P. Poelvoorde, D. P. Nolan, L. Lins, J. V. D. Abbeele, A. Pays, P. Tebabi, H. V. Xong, A. Jacquet, N. Moguelevsky, M. Dieu, J. P. Kane, P. D. Baetselier, R. Brasseur, and E. Pays. Apolipoprotein l-i is the trypanosome lytic factor of human serum. *Nature*, 422(6927):83–87, Mar 2003.
- [232] S. Veretnik, P. E. Bourne, N. N. Alexandrov, and I. N. Shindyalov. Toward consistent assignment of structural domains in proteins. *J Mol Biol*, 339(3):647–678, Jul 2004.
- [233] I. Vergne, S. Singh, E. Roberts, G. Kyei, S. Master, J. Harris, S. de Haro, J. Naylor, A. Davis, M. Delgado, and V. Deretic. Autophagy in immune defense against mycobacterium tuberculosis. *Autophagy*, 2(3):175–178, Jul 2006.
- [234] M. von Grotthuss, L. S. Wyrwicz, and L. Rychlewski. mrna cap-1 methyltransferase in the sars genome. *Cell*, 113(6):701–702, Jun 2003.
- [235] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.
- [236] K. L. Waller, W. Nunomura, B. M. Cooke, N. Mohandas, and R. L. Coppel. Mapping the domains of the cytoadherence ligand plasmodium falciparum erythrocyte membrane protein 1 (pfemp1) that bind to the knob-associated histidine-rich protein (kahrp). *Mol Biochem Parasitol*, 119(1):125–129, Jan 2002.
- [237] K. L. Waller, W. Nunomura, X. An, B. M. Cooke, N. Mohandas, and R. L. Coppel. Mature parasite-infected erythrocyte surface antigen (mesa) of plasmodium fal-

- ciparum binds to the 30-kda domain of protein 4.1 in malaria-infected red blood cells. *Blood*, 102(5):1911–1914, Sep 2003.
- [238] S. X. Wang, K. C. Pandey, J. R. Somoza, P. S. Sijwali, T. Kortemme, L. S. Brinen, R. J. Fletterick, P. J. Rosenthal, and J. H. McKerrow. Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease. *Proc Natl Acad Sci U S A*, 103(31):11503–11508, Aug 2006.
- [239] G. D. Weedall, B. M. Preston, A. W. Thomas, C. J. Sutherland, and D. J. Conway. Differential evidence of natural selection on two leading sporozoite stage malaria vaccine candidate antigens. *Int J Parasitol*, 37(1):77–85, Jan 2007.
- [240] J. Westbrook, Z. Feng, S. Jain, T. N. Bhat, N. Thanki, V. Ravichandran, G. L. Gilliland, W. Bluhm, H. Weissig, D. S. Greer, P. E. Bourne, and H. M. Berman. The protein data bank: unifying the archive. *Nucleic Acids Res*, 30(1):245–248, Jan 2002.
- [241] D. R. Westhead, T. W. Slidel, T. P. Flores, and J. M. Thornton. Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci*, 8(4): 897–904, Apr 1999.
- [242] WHO. *The world health report 2003: shaping the future*. World Health Organization, 2003.
- [243] D. L. Williams, M. Torrero, P. R. Wheeler, R. W. Truman, M. Yoder, N. Morrison, W. R. Bishai, and T. P. Gillis. Biological implications of mycobacterium leprae gene expression during infection. *J Mol Microbiol Biotechnol*, 8(1):58–72, 2004.

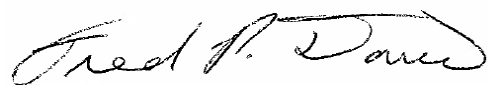
- [244] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305, Jan 2002.
- [245] B. Xiong, C. S. Gui, X. Y. Xu, C. Luo, J. Chen, H. B. Luo, L. L. Chen, G. W. Li, T. Sun, C. Y. Yu, L. D. Yue, W. H. Duan, J. K. Shen, L. Qin, T. L. Shi, Y. X. Li, K. X. Chen, X. M. Luo, X. Shen, J. H. Shen, and H. L. Jiang. A 3d model of sars_cov_3cl proteinase and its inhibitors design by virtual screening. *Acta Pharmacol Sin*, 24(6):497–504, Jun 2003.
- [246] J. P. Xiong, T. Stehle, B. Diefenbach, R. Zhang, R. Dunker, D. L. Scott, A. Joachimiak, S. L. Goodman, and M. A. Arnaout. Crystal structure of the extracellular segment of integrin alpha vbeta3. *Science*, 294(5541):339–345, Oct 2001.
- [247] H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J. D. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6):1107–1118, Jun 2004.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

4/16/2007

Author Signature

Date