Computational Identification of Protein-Peptide Interaction Specificity

David Barkan

Dissertation

University of California, San Francisco

Dedicated to my parents, Barbara Tennent and Steven Barkan, for all of their love, support, and encouragement

Acknowledgements

Grad school is serious business, and wouldn't be possible without an extensive support network. The following is not a comprehensive list of everyone who helped me both academically and emotionally, but it's a start.

Andrej Sali, my Ph.D. advisor, taught me how to do good science. I came into his lab having no idea what I was doing, but left telling other people how it should be done (Andrej has commented on this a number of times). I am now able to decompose a problem into methods, benchmark, and application; write an R01 and evaluate others (for my own benefit, and admittedly Andrej's); be paranoid about bugs in my code; determine whether errors are due to scoring *vs* sampling; find where the synergy exists between me and other lab members; leave ego out of collaborations; insist on making software open source; never use the word "this" without an antecedent; decompose all problems into points and restraints; and most importantly, love what I do because it is so cool. All the work I do for the rest of my career will have Andrej's imprint on it, which generally leads to success.

There are many other professors at UCSF who contributed to my development. Jim Wells and Patsy Babbitt were both on my thesis committee. Patsy was a huge help from the very beginning, assisting me in navigating the treacherous first year of grad school, and helping lead the UCSF Bioinformatics graduate program to being one of the best in the country. Jim was also a great collaborator, and the analysis he suggested doing on our proteomic datasets was always the type that seemed to have the greatest impact. Along with Patsy and Jim, Tanja Kortemme and Brian Shoichet were on my qualifying exam committee, and despite inducing a huge amount of anxiety, were ultimately fair and accommodating.

Charly Craik and Andrej demonstrated the model interaction between computational and experimental researchers. Charly also helped me land my current position at Protagonist Therapeutics, and I am excited to continue working with him in this capacity. Jim McKerrow, Al Burlingame, Phil Rosenthal were other experimental collaborators who were patient enough to listen to how the computational component worked (and its limitations). Tom Ferrin's passion for the Bioinformatics program was contagious, even when he was being interviewed by an idiot with a yellow jump suit who asked him why only one student joined the previous year. Hao Li and Chao Tang were nice enough to let me rotate in their labs, David Agard taught me how to give a scientific presentation, and Ajay Jain talked me off the ledge when things weren't going very well in my first year. I think I made the right choice there. Additionally, Julia Molla and Rebecca Brown were superstar graduate program administrators. We had sub-par admins in the past, and it was bad. Julia and Rebecca keep things running and cannot get enough accolades for their work.

Some of my favorite times at UCSF were the interactions with the scientists running the experimental components of my research. Dan Hostetter patiently explained to me how a Western blot worked, helped me study the basics in preparation for orals, and together we knocked the granzyme B project out of the park. Sami Mahrus kindly included me in the caspase study, Jon Trinidad asked a lot of great questions that led to some results in O-GlcNAcylation that could shake up the field, and Kailash Pandey asked questions that necessitated me becoming an expert comparative modeler.

Mike Cary, Leonard Apeltsin, Michelle Dimon, and Colin Smith were the other four people in my entering Bioinformatics class and had front seats for the ride. It is crazy how far we've come since that first breakfast.

The best part about the Sali Lab is all of the people who work there. The ups and downs of science require people with whom you want to be in the trenches. For some reason, our lab continues to attract those people. I thank the old guard for showing me how it's done: Michael F. Kim, Libusha Kelly, Dave Eramian, Fred Davis, Min-Yi Shen, Dmitry Korkin, Karin Asensio, Sebnem Essiz, Eswar Narayanan, and Madhu.

It is hard leaving my friends who are in the lab now, so hard that I was able to set it up to work there a couple days a week. In particular, thanks to J-Boss, Vadim, Robert, Conchita, Sir Charles, Farmer, Dr. P, J-Crest, Hao, GQ, Max, Riccardo, Natalia, SJ, and my "protégés" Bart, Brittany, and Javona. TTT. I would especially like to acknowledge the tireless efforts of Elina, Ben, and Ursula, our systems administrators, who put up with my noob questions and changed the code when I broke it because I wanted to do something weird, and Daniel Russel, the IMP caretaker who basically did all of my work for me. Finally, thanks to Hilary Mahon for keeping the lab running without going crazy, and being incredibly good natured during the process.

There are too many people outside of the lab to thank individually, but they all had a hand in my success and helped me blow off steam during the rare down times; in particular, the Haverford crew of Stavis, Lovi, and Watson; Palmer, Phil, Kim, Raquel, Kathy, Roger, and Team Sunrise; Amy, Jay, and Shines; Cindy and Adrienne; the 'Bill Yim from upstairs' 1543 8th avenue crew of Laura, Liz, Hengameh, and Jenn, who put up with me when I didn't have time to clean (although honestly I returned the favor with interest).

Special acknowledgement goes to Peter Skewes-Cox and Patrick Hullett. I interviewed Pete when he applied to UCSF and it turned out we had met at a concert a few years before. We have since gone to many more. I can't believe he finished school before I did, but I take solace in my victory over him in the 2009 UCSF Rising championship match. Patrick moved into my house the day after I did, and we have since lived together for six years (after being kicked out of that first place). I hope that in exchange for putting up with him, I get a good seat at his Nobel Laureate ceremony.

Even more special acknowledgement goes to my girlfriend Kelly Strickland. She entered the picture relatively recently, but I couldn't imagine being with anyone else. She has been a trooper during this past year, showing an immense amount patience and understanding when time gets allocated to finishing school. She helps keep me sane and in the moment, which is all I need to keep going. I am very much looking forward to the future with her.

Finally, I could never have come anywhere close to UCSF without the support of my family. My extended family is as close as any you'll find, united by the dinner table, the Lake, and the emailed Christmas lists. My brother Joe has it figured out in San Diego, interspersed by six week summer vacations to Mexico (he calls it a "job"). And most importantly, I would like to thank my parents, who have done so much for me, not only during my time at UCSF, but every day leading up to it. Their unconditional encouragement and support is impossible to pay back.

Abstract

Evolution is the uniting concept of biology and life. At its fundamental level, it operates by sampling amino acid residues in proteins to optimize stability and function. One definition of the function of a protein is through its interactions with other molecules, especially other proteins. Growing evidence suggests a widespread phenomenon involving the domain of one protein interacting with a short, linearly extended peptide region on another protein, accounting for up to 40% of all protein interactions in the cell. As such, these interactions are manipulated by invasive organisms and human diseases to cause pathogenesis, and are targets for new classes of drugs. Identifying specific interactions is therefore critical for human health. Experimental techniques have characterized thousands of peptide binding events, but conducting experiments can be costly and time-consuming, and their results can be prone to false positives. To both help guide experiments and to analyze their output, there is a need to develop accurate computational methods for predicting protein-peptide interaction specificity. This dissertation addresses this challenge, describing four complementary approaches: (i) a machine-learning algorithm to predict proteolytic cleavage in substrates of pro-apoptotic proteases; (ii) a peptide docking method that models the conformation of peptides in complex with protein binding sites; (iii) statistical analysis of peptide datasets derived from high throughput proteomic experiments to characterize factors mediating binding specificity; and (iv) prediction of peptide interactions contributing to pathogenic invasion.

List of publications associated with this dissertation

1. **D.T. Barkan**, B. Lenselink, E. Tijoe, D. Russel, K. Lasker, A. Sali. “A divide-and-conquer approach to protein-peptide docking” (manuscript in preparation).

2. **D.T. Barkan**, D.R. Hostetter, S. Mahrus, U. Pieper, J.A. Wells, C.S. Craik, A. Sali. “Prediction of protease substrates using sequence and structure features.” *Bioinformatics* 26, 1714-22, 2010.

3. J.C Trinidad, **D.T. Barkan**, B.F. Gulledge, A. Thalhammer, A. Sali, R. Schoepfer, A.L. Burlingame. “Global Identification and Characterization of Both O-GlcNAcylation and Phosphorylation at the Murine Synapse” (manuscript submitted)

4. S. Mahrus, J.C. Trinidad, **D.T. Barkan**, A. Sali, A.L. Burlingame, J.A. Wells. “Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini.” *Cell* 134, 866-876, 2008.

5. K.C. Pandey, **D.T. Barkan**, A. Sali, P.J. Rosenthal. “Regulatory elements within the prodomain of falcipain-2, a cysteine protease of the malaria parasite *Plasmodium falciparum*.” *PLoS One* 4, e5694, 2009.

6. F.P. Davis, **D.T. Barkan**, N. Eswar, J.H. McKerrow, A. Sali. "Host-pathogen protein interactions predicted by comparative modeling." *Protein Science* 16, 2585-2596, 2007.

7. U. Pieper, B.M. Webb, **D.T. Barkan**, *et. al.* “MODBASE, a database of annotated comparative protein structure models, and associated resources.” *Nucleic Acids Research* 39, 465-474, 2011.

Table of Contents

Table of Figures

Table of Tables

Chapter 1. Introduction

1.1. Peptide-mediated interactions are prominent across life

The interactions a protein makes with other proteins and molecules have been optimized by evolution. These interactions allow proteins to carry out a wide range of biological functions that are crucial for all forms of life. Typically, protein-protein interactions are thought of as being between two domains where each domain contributes multiple non-contiguous segments of amino acid residues that form a binding patch, the size of which averages 1,000 $Å^2$ per monomer[1]. These interactions generally occur at high binding affinity, resulting in the formation of a stable protein complex. In contrast, recent evidence has demonstrated the widespread phenomenon of another type of interaction that occurs between a globular domain from one protein and a short, linearly extended peptide region of another protein[2, 3]. These so-called linear motifs (LMs) frequently occur in disordered regions of a protein, or on ordered loops or unfolded segments[3, 4] (Figure 1.1a). These interactions are usually highly specific, with the specificity largely determined by the amino acid sequence of the LM[5]. The nature and extent of peptide-mediated interactions (PMIs) is just beginning to be understood due to a long-standing focus of experimental techniques on stable, folded proteins[6].

A subset of PMIs involves interactions where a globular domain interacts with a small peptide that is not part of a larger protein (*e.g.* MHC Class I peptides; Figure 1.1b). In many respects, two types of interactions are identical, as LMs are as flexible as small peptides due to being found primarily in disordered regions, and both types involve

many of the same biophysical principles mediating interaction specificity[7]. As such, PMIs and protein-small peptide interactions will be treated as interchangeable for the purposes of this dissertation, except where noted.

Statistical analysis of large datasets of PMIs have given insight into their prevalence, functions, and properties. PMIs have been estimated to account for up to 40% of all protein-protein interactions[2]. They occur at low (typically μ-molar) affinity and are often transient, making them ideal mediators of protein signaling cascades and other cellular contexts where a rapid response is required. Indeed, PMIs account for 60% of all interactions in the Signal-Transduction Knowledge Environment[2, 8]. 85% of LMs are in disordered regions of proteins[9]; and given that disordered proteins are more likely to be hubs in large protein interaction networks[10], LMs allow certain proteins to participate in multiple cellular processes. This is the case with the tumor suppressor p53, which acts in several contexts to regulate the cell cycle, largely controlled through PMIs with different kinases and other proteins[11]. PMIs can also form stable complexes, as is observed with nuclear export and localization signals (NES and NLS, respectively). These peptides bind to karyopherins, which transport proteins across the nuclear membrane. Dissociation of the peptide from the karyopherin occurs only through various associations with G proteins and GTPase activating proteins[12].

PMIs and their annotation are stored in a number of databases, the most prominent being ELM[9] and Domino[13]. Two recent datasets of particular interest focused only on protein-peptide interactions with solved structures[7, 14]; here, the peptide was not part of a larger protein chain, although the authors note that they were probably derived from LMs on proteins in the majority of cases.

Researchers are only beginning to create a full picture of the structure and function of PMIs. In particular, understanding of protein-peptide interaction specificity, defined as identification of which peptides bind to a given protein, is an open problem. Determining this specificity is important for elucidating many critical cellular processes, and for developing medicine to treat human disease and combat pathogenic invasion. Following is a discussion of general properties of PMIs, their role in human health, and experimental and computational methods for characterizing new interactions.

1.2. The majority of linear motifs can be grouped into a few classes

Highlighted are several examples of protein-peptide interactions. In addition to being prevalent in the cell, these examples are well-characterized model systems for studying general principles of PMIs, and are frequently used as benchmark sets when developing new experimental and computational methods for detecting these interactions.

Src Homology 2 and 3 (SH2 and SH3) are 7 kDa modular signaling domains found in many kinases and other protein families, binding specific LMs on kinase substrates to promote specificity between the two proteins. One of the most abundant domain types, they are present in more than 300 human proteins[9]. SH3 domains recognize proline-rich peptides possessing the residue sequence Pro-X_{aa}-X_{aa}-Pro, where X_{aa} can be any amino acid, while SH2 domains recognize various sequences containing a phosphorylated Tyr residue. These two domains are often found on the same protein (including their namesake Src) and as such can modulate signaling pathways by addition and recognition of phosphate groups, and can also cooperate to auto-inhibit their proteins as necessary[15].

PDZ domains are 90-residue β-rich domains that bind the C-terminal ends of proteins. Upon association, PDZ substrates add a strand to the existing PDZ β-sheet, mediated by recognition of peptide side chains by the domain[16]. This recognition can be highly specific; one study showed that there are at least 16 different PDZ classes, with each recognizing a different peptide sequence motif[17]. In signaling contexts, PDZ domains are prevalent in processes directing protein localization and frequently act as scaffolds for higher order signaling complex assembly[18].

Major histocompatibility complex (MHC) proteins are important players in the immune response. These proteins, having been secreted through the golgi apparatus to the cell surface, bind peptides derived from pathogenic proteins and display them on the surfaces of different cell types, where they are recognized by lymphocytes to stimulate an adaptive immune response against the pathogen. MHC proteins are grouped into two classes. MHC Class I are present on a variety of human cell types. The peptide binds into a cleft created by a β-sheet floor bounded by two helices. The second and third residue from the peptide C-terminus act as anchors, contributing most of the binding energy, while the rest of the peptide fits into as many as six binding pockets on the protein floor. Peptides binding to MHC Class I domains are usually 9 residues in length[19]. MHC Class II domains are found on “professional” antigen-presenting cells such as macrophages. They consist of two membrane-spanning chains that each comprise half of the peptide binding region. One end of the domain is open, accommodating peptides of up to 22 residues in length, and the binding pocket is shallower than that of MHC Class I proteins and has broader specificity. Even so, both MHC Class I and II domains are specific for between 10^3 to 10^4 peptides, encoded by

the highly polymorphic HLA genes to allow for a complicated map of MHC proteins to the peptides they bind. This system has important implications for human health, as T cells need to differentiate between host and pathogen peptides to avoid an autoimmune response; additionally, these binding events are the basis for epitope-based vaccine development. (Lafuente 2009)

Other examples of domains that bind to linear motifs include WW, 14-3-3, and many different protein kinases. Together with the described examples, these six domain types account for more than 75% of all PMIs in the Domino database[13]. The rate of PMI discovery is increasing, and these particular domains may be overrepresented in LM databases due to selection bias of certain systems for research [2].

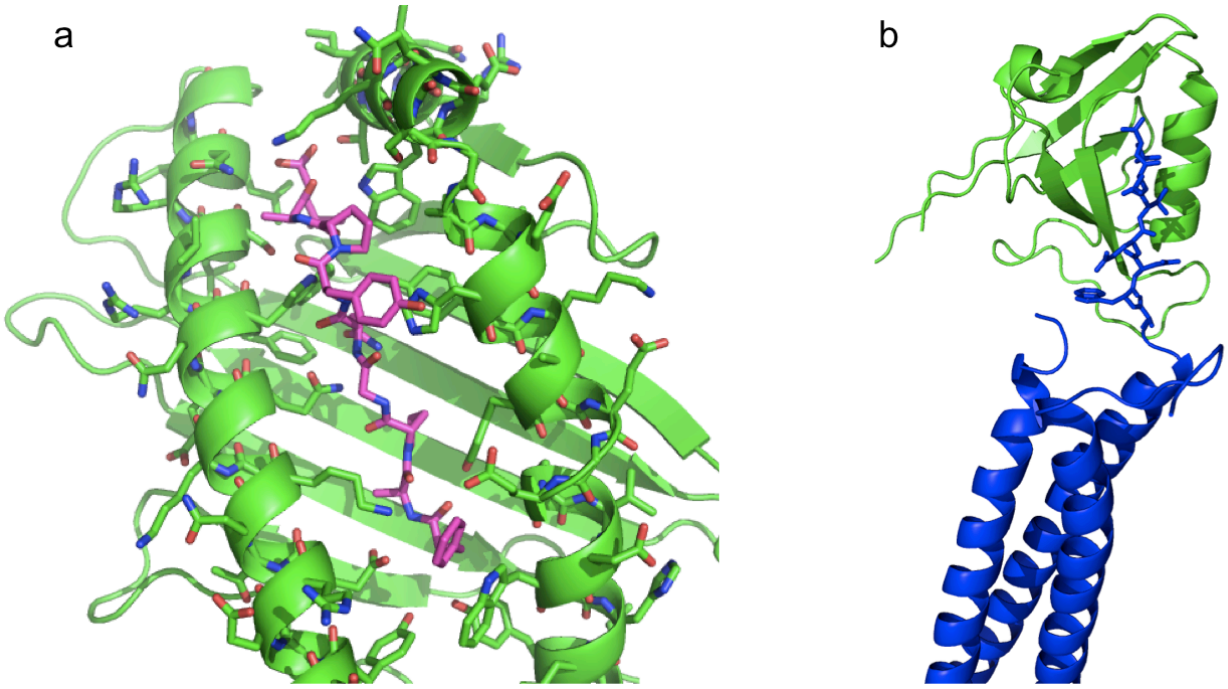


Figure 1.1 Examples of protein-peptide interactions.
(a) Major histocompatibility class I domain (green) interacting with a ten-residue peptide (magenta).
(b) PDZ scaffolding domain (green) interacting with a linear motif located on the C-terminal end of its substrate protein (blue).

1.3. Protein-peptide complexes have unique characteristics

Protein-peptide interfaces possess a number of structural properties distinguishing them from those in larger protein-protein interactions. A recent study compiled 103 high resolution complexes (the “PeptiDB” dataset) to analyze these properties in detail[7]. The average complex buries 500 $Å^2$ of total accessible surface between the two monomers, which is 1500 $Å^2$ less than in interactions between two globular domains. The secondary structure of the peptide upon binding is distributed evenly between α-helix, loop, and β-sheet, the latter involving the peptide associating with an existing strand in the protein[9]. About a third of residues at the protein-peptide interface are polar, which is similar to the fraction observed in protein-protein residues. However, residues mediating protein-peptide interactions pack at a higher density than do those in protein-protein interactions and exhibit a greater frequency of hydrogen bonds across the interface, mainly derived from main-chain atoms. Interestingly, protein binding sites are generally inflexible; in the PeptiDB dataset, the conformation of 87% of proteins shifted at an average of only 1.48 Å RMSD between the peptide-bound and -unbound states (considering complexes where both states were available as solved structures). While the remaining 13% of proteins could change conformation substantially upon peptide binding, this result suggests that in general, the peptide adapts its conformation to accommodate the protein binding site[7]. Finally, it was observed that most pairs of interacting protein-peptide segments across interfaces could be reconstructed by motifs found in monomeric protein folds, indicating that the wealth of data derived from protein structures can be used to model protein-peptide complexes where fewer solved structures are available[20].

A separate study distinguished LMs from their larger context in disordered regions[5]. The amino acid composition of LMs was found to be more hydrophobic and similar to those residue types found in globular protein cores than were the surrounding disordered residues. Moreover, these hydrophobic residues (typically Trp, Leu, Phe, and Tyr) were thought to be the primary contributors to recognition of the LM by the binding site. This observation is supported by analysis on the PeptiDB dataset demonstrating that certain residues act as “hot spots” in binding, where the majority of the binding energy comes from a subset of residues participating in the interaction[7]. Additionally, it was shown that in disordered regions, the residues surrounding the LM contribute an average of 20% to the total binding energy of the interaction. This contribution represents an important difference in PMIs *versus* protein-small peptide interactions and should be accounted for in computational modeling of the latter[5].

1.4. Peptide-mediated interactions are important in human health

PMIs play a prominent role in human health and disease. Analysis of functional annotation suggested that disordered proteins are associated with more diseases than ordered proteins (Uverski, 2008). There are a number of reasons for this observation. The propensity for disordered proteins in large signaling networks, which are responsible for cell growth regulation among other processes, allows for mutations in these proteins to lead to unregulated cell growth and tumor formation. Also, pathogens will mimic interactions in signaling networks to modulate host gene expression, increasing the concentration of proteins necessary for pathogen survival and suppressing proteins involved in the host immune response[21]. Additionally, longer linear motifs in disordered proteins are prone to aggregation, which is responsible for a

number of human neurodegenerative diseases[22]. A few examples follow, demonstrating the scope of the roles of PMIs in human health.

The nuclear pore complex is an 120 mDA assembly consisting of ~450 subunits to form a pore in the nuclear membrane, regulating transport of molecules to and from the nucleus[23]. The NLS and NES peptide motifs mediate protein import into the nucleus and export into the cytosol, respectively, through their specific association with karyopherins. One study demonstrated the role of this process in Huntington's disease (HD) (Xia, 2003. The protein Huntington contains both an NLS and NES, suggesting it plays a role in multiple cellular compartments. However, mutant Huntington, which is characterized by a polyglutamate stretch exceeding a length threshold, is associated with cleavage of the NES by endogenous enzymes, leading to accumulation of Huntington in the nucleus and neural toxicity. Additionally, viral proteins have acquired NLSs to exploit the karyopherin mechanism to gain entry into the nucleus. An example is bornavirus, which is fatal in many animals and has been implicated in human psychiatric disorders. Bornavirus RNA polymerase, which acts to transcribe viral DNA in the host nucleus, was found to have an NLS that is necessary for viral replication (Walker, 2002).

Many gram-negative pathogenic bacteria use the type-three secretion system (TTSS) to deliver virulent effector proteins into the host cell through an 80 nm syringe-like appendage protruding from the bacterial surface. Effector proteins specifically bind homodimeric globular chaperone proteins to induce localization to the base of the TTSS complex and mediate active transport through the syringe apparatus. Effector proteins bind through their disordered N-terminal domain that wraps around the chaperone

dimer, adding additional β-strands to an existing sheet on the sides of each monomer and forming specific interactions across the top in a linearly extended conformation. While bacterial genomes encode several paralagous chaperones, each virulence protein binds specifically to its own chaperone. It is likely that this specificity is an additional temporal regulator of delivery of virulence proteins into the host cell (Stebbins 2003).

Due to the role of protein-peptide interactions in human health, it is critical to develop methods to characterize their specificity on a proteomic scale. Knowledge of binding and conformation in endogenous PMIs has lead to successful rational design of peptide and small-molecule drugs to inhibit harmful interactions. For example, the nutlin small molecule class binds to MDM2, which has been implicated in cancer through over-negative regulation of p53; nutlin mimics a p53 peptide to bind MDM2 and disrupt this interaction (Tovar 2006). Indeed, there has been a recent focus on searching for peptide mimics among approved small molecule drugs, as peptide binding sites may prove to be better targets for small molecules due to their small size and easily defined interaction surface (Pasarathi 2008, Eichler 2008). Alternatively, peptides themselves are well-suited as therapeutics, as they have been evolutionarily optimized to bind with protein domains. This concept is discussed in detail in the next section.

1.5. Peptides have potential for therapeutic use in different contexts

Traditional drug discovery has searched for small molecules that bind to enzyme active sites. As drug discovery pipeline output has declined in recent years, other biological contexts have been examined as targets. As discussed, one approach has been to disrupt protein-protein interactions to inhibit normal protein function (Eichler 2008).

Protein interfaces can involve many atomic contacts over a relatively flat topology, making small molecules unsuitable for binding in many cases (Rubinstein 2009). Alternatively, in addition to their use in PMI inhibition, peptides are well-suited for this task, as they can be designed to mimic the binding interface and their larger size can disrupt more atomic contacts in the native surface (Eichler 2008). In all cases, a primary challenge in using peptides to inhibit protein-protein interactions is to determine of a peptide sequence that will bind specifically to one of the protein surfaces at high affinity. A number of experimental and computational approaches have been developed to address this task, discussed in later sections.

Several examples of peptides acting as potential therapeutics are highlighted. The first relates to the protein endostatin, which is an endogeonous suppressor of angiogenesis but suffers from typical drawbacks of using proteins as therapeutics (discussed below). Angiogenesis promotes tumor formation and cancer metastasis and is a target of anti-cancer drugs. A small peptide derived from endostatin was found to inhibit angiogenesis along with tumor progression in a process involving β1-integrin, which is one of several native binding partners of wild-type endostatin. (Wickstram 04)

Peptide therapeutics have been successfully used to inhibit HIV infection. The HIV envelope glycoprotein gp41 is recognized by CD4 receptors on target cells, allowing for membrane fusion and subsequent viral entry. The structure of Gp41 includes three-helix bundle core (NH-R) containing hydrophobic grooves to which three other helices (CH-R) tightly interact. Designed using the CH-R sequence as a template, the FDA-approved drug Enfuviritide inhibits this interaction by competing for N-HR binding. (Naider 2009) Another insightful study preventing HIV infection focused on HIV-

1 integrase (IN) which catalyzes integration of viral DNA into the host genome (Goss, 2007). IN shifts between dimeric and tetrameric states; two dimers bind at each end of the viral DNA and form a tetrameric complex in the presence of the host cell growth factor LEDGF/p75. The authors designed small peptides mimicking LEFGF/p75; these peptides induced early IN oligomerization which resulted in an inactive tetramer and abolished the ability of HIV to replicate *in vitro* (Goss, 2007).

Addtionally, peptides are being used in other therapeutic contexts, including as probes that bind to biomarkers for *in vivo* early detection of cancer (Hao 2008), vaccine development (Purcell, 2007), and enzyme function inhibition (Ron, 1995). In these contexts as well, knowledge of protein-peptide interaction specificity is crucial to success.

There are unique obstacles in using peptides as drugs. As with other unfolded proteins in humans, they are prone to clearance by the immune system as well as ubiquitin-mediated degradation. To address this obstacle, non-native peptides have been designed that retain the binding specificity of peptides but the chemical composition of which has been modified to avoid recognition by endogenous factors. Examples include peptoids (Zuckerman, 1992) and β-peptides (Baldauf, 2008), in which the side chain is appended, respectively, to the nitrogen and β-carbon atoms of the peptide backbone rather than to the α-carbon. Another approach is to use naturally-occurring small peptides containing one or more disulfide bonds. These so-called disulfide-rich peptides (DRPs) comprise many toxins in animal venoms and are resistant to heat denaturation and degradation in the bloodstream (Hartig, 2005).

1.6. Experiments can discover peptide-mediated interactions on large scales

A variety of experiments have been developed to characterize protein-peptide interaction specificity. While there is some overlap with those used to study globular protein binding partners, such as affinity purification and co-immunoprecipitation (Ceol, 2007), the most widely used experiments are designed specifically for peptides. One of the most powerful experimental techniques for determining protein-peptide interaction specificity is through phage display. This method was used to characterize the largest number of all protein-peptide interactions in the DOMINO database (Ceol, 2007). In phage display, random peptides are displayed on the surface of bacteriophage, and peptide binding specificity is determined by sequencing those phage which bind to an immobilized target protein of interest (Watson, 2002). This protocol can assess binding between the protein and more than 10^{10} peptides to create a specificity profile. It has been used to determine such profiles for many domains including PDZ (Tonikian, 2008), WW (Tanner, 2004), and SH3 (Haeberlein, 2005).

Another method for determining protein-peptide binding specificity is the peptide microarray (Li, 2009. Here, peptides are immobilized onto a chip, either by synthesizing the peptide directly on the chip or by covalently attaching pre-made peptides onto a functional surface. Protocols achieving the highest peptide density have reported 40,000 peptides per cm^2 (Beyer science 07). Fluorescent labels are used to detect protein binding events. This method also creates complementary negative specificity profiles, as it determines which peptides to not bind the target protein; additionally, it is capable of immobilizing post-translationally modified peptides.

A number of other experimental techniques exist, including yeast 2 hybrid

(Broderick 2003), raising antibodies against an LM epitope for determining functional sites on disordered proteins (Sampson, 2004), and mass spectrometry for detecting post-translationally modified peptides (Mahrus, 2008; discussed extensively in Chapter 4). Together, these methods have produced a wealth of knowledge of peptide interaction specificity. Despite these advancements, there are some inherent limitations. Experiments that rely on the formation of stable complexes can be inefficient, as protein-peptide interactions are often transient. Experiments can be expensive and time-consuming, and some are prone to a relatively high false positive rate. Most high-throughput methods are discovery-based and thus rely on their results to generate biologically relevant hypotheses (Diella, 2008). Additionally, even in phage display, it is difficult for experiments to obtain a complete specificity profile for a particular protein. The more subtle aspects of protein-peptide binding, such as residues at particular positions that prevent binding, or cooperativity across peptide residues, are often missed. To address these shortcomings, many computational approaches have been developed.

1.7. Computational approaches predict linear motif binding

Computational approaches to predict protein-peptide interactions generally fall into one of two categories. The first only attempt to predict whether or not a peptide binds and include statistical and machine learning approaches that benefit from training data of known positive and negative peptide binders. The second, reviewed in the following section, are protein-peptide docking algorithms that assume the peptide binds and attempt to model the conformation of the bound peptide, although the more ambitious methods also include prediction of binding as well. There have been hundreds, if not

thousands, of methods developed to carry out the first approach, and a comprehensive overview of these categories is beyond the scope of this review. A brief summary of three of the most popular follows.

Position specific scoring matrices (PSSMs) take as training input a sequence alignment of a list of peptides that are known positive binders and create a scoring function, usually represented in terms of a log odds bit score, where the score at each position for a given amino acid residue represents the frequency with which that residue was observed in that position in the alignment. Peptides to be evaluated are examined at each position in the sequence, and the scores across all positions are summed to create an overall score that is usually compared to a cutoff. PSSMs are easy to generate and conceptually simple, and can perform well when there is a high degree of specificity at each position in the sequence. However, they fail to take into account correlations between residues in the sequence, as each position is independent. Moreover, if there is a medium degree of degeneracy in the sequence, PSSMs operate with less accuracy. One way to get around this problem is to weigh the scores for certain positions in the sequence if it is known that they contribute more to binding than other positions. Methods using PSSM include the popular Prediction of Protease Specificity (PoPS) algorithm which has been generally applied to protease cleavage sites (Boyd, 2005); PePS, which is similar in concept but applied specifically to cathepsins (Lohmuller, 2003), and GrabCas, which has been applied specifically to the pro-apoptotic protease types, granzyme B and the caspases (Backes, 2005).

Hidden markov models (HMMs) represent a canonical bioinformatics class of methods that have been used extensively in sequence motif finding. A natural

application of these algorithms is to search a protein sequence for peptides likely to bind to another protein. HMMs are represented by a first-order Markov chain with a set of states; they encode transition probabilities to move from one state to another and emission probabilities which generate output during transitions from state to state. The output can represent a specific amino acid residue in a peptide sequence; HMMs are thus trained on input peptides and a test peptide is evaluated by calculating the probability of emitting the full sequence. HMMs are more robust to capture these probabilities given the training data than are PSSMs, and can model gaps in sequence alignments, but have a few drawbacks. For example, it is difficult to use them to incorporate conserved residue chemical properties at a particular position; additionally, a single HMM can only be trained and applied to predict one type of binding (*i.e.*, either positives or negatives) as opposed to discriminating between the two types (Gould, 2009). HMMs have been used extensively in searcing for linear motifs; examples include prediction of phosphorylation (Huang NAR 2005), and binding of peptides to SH2 domains (McLaughlin, 2006 JMB) and to MHC Class I molecules (Nielsen Protein Science 2003).

Finally, another popular technique for identifying linear motifs is with support vector machines (SVM). SVMs are a class of machine learning algorithms that can be trained on positives and negatives and predict into which classification a test peptide falls. *N* peptide features are encoded into an *N*-dimensional vector, which represents one data-point. These features generally correspond to residue identity, although structural aspects can also be incorporated. Data-points are plotted in *N*-dimensional space, and the SVM generates a hyperplane that separates the positives from the

negatives. Test peptides are evaluated by generating the same feature vector and determining on which side of the hyperplane it falls. The ability to discriminate positives from negatives is a powerful feature of SVMs, although they can suffer from over-training in some cases. SVMs have been used to predict peptides binding to SH3 domains (He, 2009) and MHC molecules (Liu, 2009); they are also the focus of Chapter 2, which uses an SVM to predict protease cleavage sites.

1.8. Peptide docking methods model the bound conformation of peptide complexes

The computational approaches discussed in the previous section generally rely on statistical principles of linear motifs to predict protein-peptide interaction specificity. They are largely geared toward predicting whether or not a peptide binds to a given protein. A complementary approach to predicting binding is peptide docking, which predicts the conformation of the bound peptide. This problem is a critical one in structural biology. Knowledge of the peptide conformation helps address many of the challenges outlined in previous sections. Insight into how pathogenesis occurs can be gained by observing specific contacts made in inter-species protein-peptide interactions. Understanding which residues are buried in the binding pocket allows for researchers to replicate these contacts in therapeutic design process. The optimal peptide docking algorithm would also predict whether binding occurs, either through estimation of the free energy of binding in the bound conformation, or by using a normalized score to compare peptides with different amino acid residue compositions. Additionally, transient complexes, which make up the majority of linear motif recognition events, are difficult to

solve crystallographically; therefore, there is a need for accurate peptide docking algorithms to support these types of experiments.

Many computational docking approaches have focused on small molecules, and peptide docking methods have been relatively underrepresented. Nonetheless, several approaches have been developed recently that achieve good accuracy in general biological contexts, and many others have been described to dock peptides to specific protein families. Following is a description of the former group, and a brief survey of the latter.

One recent peptide docking algorithm, Rosetta FlexPepDock *ab initio* (Raveh, 2011), demonstrated success when applied to a large benchmark set of unrelated protein-peptide complexes. This method initializes the peptide with an arbitrary conformation in the vicinity of the binding site, and uses a two-stage protocol to perform the docking. The first stage is a coarse grained approach in which the peptide is subjected to repeated rounds of rigid body transformations followed by random Monte-Carlo moves to perturb the peptide backbone conformation both through backbone dihedral adjustments and fragment insertion from the Rosetta protein fragment library. The second stage refines the protocol using a previously described method (Raveh, 2010) that performs similar local rigid body and backbone Monte-Carlo moves, but minimizes the Rosetta energy function (Rohl, 2004) and applies the Metropolis criterion to accept a move based on the energy of the system. The method was benchmarked on a set of complexes of bound peptides, a subset of which had an alternate, unbound structure available; the authors docked 18 out of 26 peptides with less than 2 Å RMSD

error when compared to the native structure in the bound cases and 7 out of 14 in the unbound cases. (Raveh, 2011)

The DynaDock algorithm also performed well when applied to a smaller benchmark set (Antes, 2009). This method also employed a two-step approach. The first component was a coarse grained procedure where the peptide was placed in the binding site of the protein and subjected to a number of steps featuring random perturbations of its structure. Each step included a random translation and rotation of the full peptide as well as a rotation of a random number of backbone and side chain dihedral angles by 10°. The full step was rejected if any pair of protein-peptide atoms overlapped by more than 90%. The second component was a high-resolution refinement procedure using molecular dynamics. A scoring function was used as the force to drive the simulation, incorporating physical force field terms in addition to coulombic and van der waals terms. These latter two restraints included a parameter that scaled their values to weaken their effects on the simulation; one of the notable features of DynaDock is that this parameter is optimized at each step of the simulation by applying conjugate gradients with respect to the parameter.

The initial conformation of the peptide was a random placement within 6.5 Å of the protein binding site, and the N and C terminal ends of this placement were constrained to be within 8.5Å of their coordinates in the native state. Protein atoms were fixed in the coarse-grained step but were allowed to move in the refinement step. The procedure was repeated for each member of a benchmark set of solved complexes, and an impressive 11 out of 15 had their best scoring pose align with the native state at less than 2.10 Å RMSD (Antes, 2009).

A third generally applicable docking algorithm uses mutually orthogonal Latin squares to calculate the scores of a subset of all conformations of the peptide, find the local minima associated with these conformations, and average the scores using a variant of the mean field technique (Prasad, 2008). The authors defined the system as having M degrees of freedom sampled at N intervals, and thus the full search space is N^M. The method samples N^2 conformations and averages them. The protocol is repeated 1,5000 times, and the authors show that this number is sufficient to identify all of the local minima in the system. Force-field terms are used to score intra-peptide interactions, and the PLP scoring function, which is a combination of physical and statistical distance-dependent terms, is used to evaluate protein-peptide interactions (Gehlhaar, 1995). The method performed well, docking 39 out of 56 of the diverse benchmark set complexes to within 2.00 Å RMSD of the native state (examining backbone atoms only; no side chain comparison was performed).

In addition to these three methods, which represent the state-of-the-art in peptide docking, there have been many studies that focus on docking peptides to a particular protein or protein family of interest. These methods apply canonical docking techniques, and often demonstrate good results, but have yet to be applied in a general context and are often optimized to work on a specific system. The most widely studied system involves MHC molecules (Lafuente, 2009 and section 1.2). One method used a Monte Carlo approach to sample peptide conformations and scored with a coarse-grained pair-wise atomic distance-dependent potential and solvation energy approximation based on the buried atomic surface upon peptide association to the binding site (Ref Liu 2004). Another technique scored peptide association solely based on the number of atomic

contacts between peptide and protein atoms, exhaustively sampled peptide dihedral angles in increments of 30°, and incorporated explicit water molecules if they were present in the solved structure of the protein binding site (Bui 2006). MHC-peptide complexes have proven to be particularly suitable for explicit free energy modeling; different studies have attempted these calculations while optimizing the system with Monte Carlo (Bordner 2006), molecular dynamics simulated annealing (MDSA) (Fagerberg 2006), and the αBB branch-and-bound approach (Schafroth 2004). Finally, one study docked the terminal ends of the peptide into the MHC molecule, applied loop-closure of the intermediate backbone atoms using MODELLER (Sali 1993), and refined the peptide with Monte Carlo sampling using physical energy terms, docking 33 out of 40 peptides to within 2.5Å RMSD of the native state (Tong 2004).

In addition to MHC molecules, studies have focused on other particular systems of interest. These efforts include docking peptides to PDZ domains with MDSA using a physical force-field and rotamer optimization (Niv 2005) and a Monte Carlo approach incorporating hydrogen bonding, hydrophobicity, and electrostatics scoring terms (Staneva 2009). Another group performed two complementary docking peptides to SH3 domains using homology modeling, explicit free energy calculations, and molecular dynamics simulations, in the process doing well to discriminate successfully SH3 peptide binders from non-binders (Hou 2006 Plos comp bio; Hou 2006 J Prot Res). Finally, other approaches have focused on docking peptides to kinases and phosphatases (Huan, 2009) and nuclear receptors (Kurcinski 2006), and used the Rosetta docking scoring function to study peptides binding to HIV-1 Protease

(Chaudhury 2009) and explore extending peptide fragments to replicate binding free energy of longer peptides in complex with a small, diverse set of proteins (Sood 2006).

While all of these methods have demonstrated success either on a general benchmark set or a particular system of interest, there is still room for improvement. Some complexes are more challenging than others, either due to longer peptides, receptor flexibility, or inadequacies in the scoring function to capture the physical forces mediating interactions for that particular complex. In many cases, despite modeling the peptide to less than 2 Å RMSD of the native state, side chain placement is still problematic (and often the positions of these side chains are of particular interest, as they have a strong impact on peptide binding affinity). Finally, the utility of computational docking methods would be dramatically increased if they could distinguish peptide binders from non-binders; to date, no generalized algorithm as demonstrated success in this area. In Chapter 3 of this dissertation, a new general peptide docking method is presented, as well as initial results on a small benchmark set of protein-peptide complexes.

Chapter 2. Prediction of Protease Substrates using Sequence and Structure Features

2.1. Introduction

As discussed in Chapter 1, the interactions of linear motifs with globular protein domains are transient and occur with varying degrees of specificity. Proteases are a particular class of enzymes that hydrolytically cleave their target substrates, often requiring a specific residue sequence motif at the cleavage site. This motif forms complementary contacts with the protease active site, allowing for tighter binding prior to cleavage and reduction of the K_m value of the reaction, as well as allowing the formation of a high energy transition state conformation. Understanding these sequence motifs, and the structural contexts on which they fall, will enable identification of proteolytic substrates using computational methods, which, as discussed, have many advantages and can complement experimental approaches. Here, we focus on the protein-peptide specificity of the pro-apoptotic proteases granzyme B (GrB) and caspases interacting with their respective protein substrates.

Apoptosis is a noninflammatory form of cell death that regulates tissue differentiation and homeostasis in higher eukaryotes (for a review, see Taylor et al., 2008). Since apoptotic turnover of cells lies in direct opposition to the uncontrolled growth of tumor cells, a strong link exists between apoptosis and cancer. Indeed, the terminal cellular effect of most chemotherapeutic compounds is induction of apoptosis (Kaufmann and Earnshaw, 2000).

GrB is a serine protease delivered by natural killer cells into virally-infected and tumor cells (Pardo et al., 2009; Russell and Ley, 2002). The caspases are a family of endogenous cysteine proteases activated by extracellular death ligands and environmental stresses (Nicholson and Thornberry, 2003). Both protease types recognize and cleave specific peptide sequences containing an aspartic acid residue on their target substrates, activating different pathways that lead to apoptosis. Identifying these substrates has led to a wealth of knowledge about how the proteases contribute to apoptosis, how the cleavage events lead to cell death, and which substrates to target for therapeutic purposes.

Substrates of the two protease types have been discovered with a variety of experimental techniques, ranging from low-throughput gel-based methods to proteomic efforts that can identify hundreds of cleaved proteins (Bredemeyer et al., 2005; Casciola-Rosen et al., 1999; Dix et al., 2008; Mahrus et al., 2008). However, different datasets overlap only partially, indicating that many substrates remain to be identified. For example, two proteomics studies, respectively, reported 261 and 292 caspase cleavage sequences, although the high-confidence overlap between the two sets was only 64 [Figure 1A in Johnson and Kornbluth (2008)]. Furthermore, the results presented later in this dissertation indicate that high throughput proteomics studies are able to characterize a subset of all modified peptides (see sections 4.3.6 and 4.6.2 for further details).

To reduce this gap, accurate computational techniques could be used to predict protein-peptide interactions for guiding further focused experiments. Computational methods have also been applied to predict substrates recognized by GrB and caspases.

Most of these methods are based on canonical linear motif searching techniques. These methods take advantage of both protease types having a near-absolute requirement for Asp at the P1 position, while allowing degenerate preference for different residue types in the positions immediately surrounding P1. These studies rely on fixed sequence searches. (Wilkins et al., 1999), PSSMs based on frequencies of residue types in known cleavage sites (Garay-Malpartida et al., 2005; Lohmüller et al., 2003; Verspurten et al., 2009) and positional-scanning combinatorial substrate libraries (PS-SCLs; Backes et al., 2005; Boyd et al., 2005), SVMs using residue composition around the cleavage site (Wee et al., 2006), and Bayesian neural networks (Yang, 2005). A full review of computational techniques for discovering linear motifs for these and other biological systems is presented in section 1.7.

Cleavage sequences for both GrB and caspases are generally thought to occur on flexible, disordered regions of substrates (Hubbard, 1998). However, it was previously shown in an analysis of caspase substrate structures that many of these known cleavage sites are in α-helices and even occasionally on β-strands (Mahrus et al., 2008; Timmer et al., 2009). This observation motivates the choice of a machine-learning algorithm that relies on the structure as well as sequence information. Here, we describe such a protocol incorporating SVM learning (Barkan 2010). The method is trained and benchmarked on separate pools consisting of known GrB and caspase cleavage sequences. It is then applied to the human proteome to generate a list of high-confidence predictions for experimental validation. Two such candidates are the proteins AIF-1 and SMN1, which are experimentally validated as being cleaved by GrB. The approach has the potential to provide greater coverage of substrates for both GrB

and caspases, and can be easily adapted to other protein-peptide systems through our web server that can learn from any user-supplied protein-peptide training set (see section 6.1 for web-server availability and description).

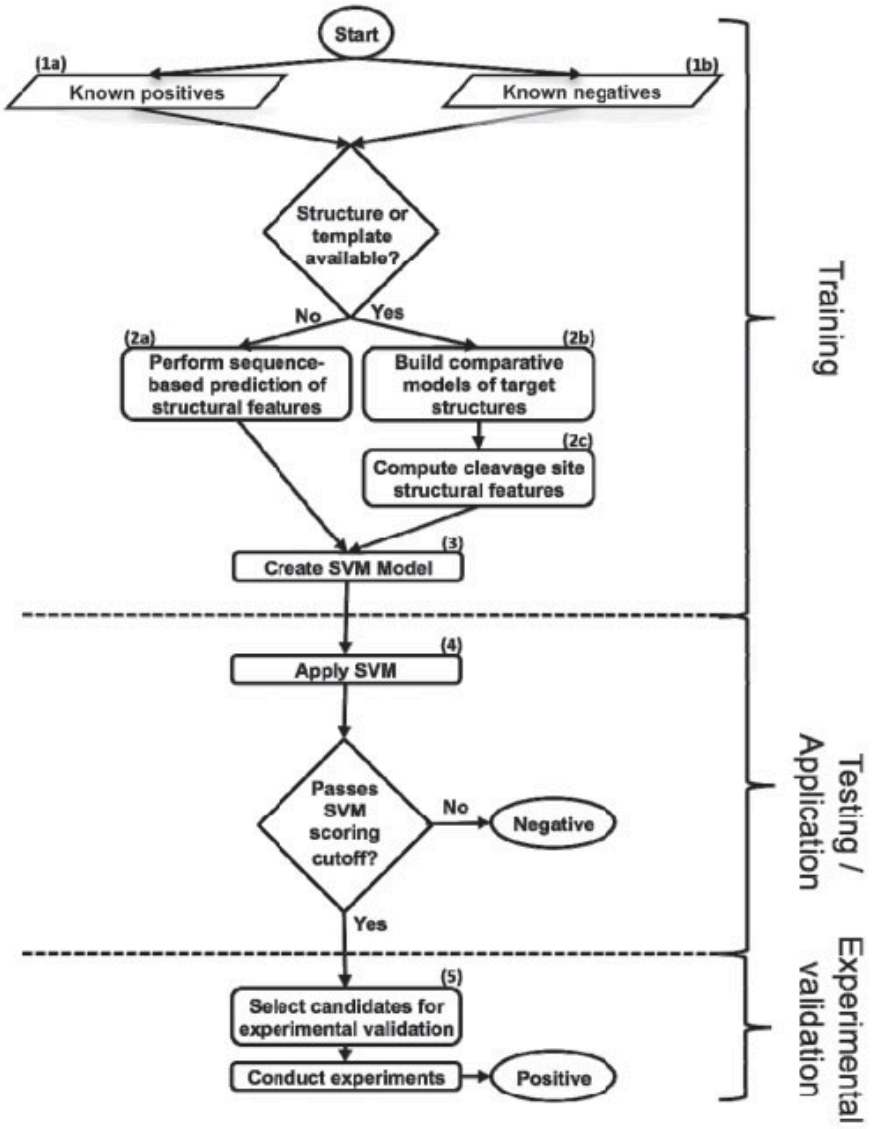


Figure 2.1 Flowchart of machine learning procedure.
Peptides are scored with the SVM trained on sequence and structure features; the peptides that pass the cutoffs derived from benchmarking are the final candidates for experimental validation.

2.2. Results

2.2.1. Benchmark sets are created from positive and negative substrates

For each protease type, two sets of octapeptides were compiled to benchmark the method (Figure 2.1, steps 1a and 1b). These sets included peptides cleaved ('positives') and not cleaved ('negatives') by the proteases, respectively (Barkan 2010, Supplemental Figure 1a). For GrB, the positives include 54 cleavage sequences from literature (i.e. our 'GrBah' dataset; and section 0) and 305 cleavage sequences from a proteomics experiment that used combined fractional diagonal chromatography for isolating peptides (Van Damme et al., 2008). These positives spanned the P4 to P4′ positions using the traditional protease nomenclature (Schechter and Berger, 1968). Positives for caspase substrates were drawn from the literature-curated Casbah dataset (Lüthi and Martin, 2007) as well as a separate proteomics dataset obtained in experiments with the Jurkat cell line (Mahrus et al., 2008 and section 4.2. The negatives for both protease types were all octapeptides in known protein substrates that are outside of the experimentally identified cleavage site and contain Asp in the fourth position (Barkan 2010, Supplementary Figure 1b). While it is possible that some of these negatives are in fact cut by the protease and were missed experimentally, many of the positives in the benchmark sets were confirmed by studies that afford a high degree of coverage. The use of octapeptides outside the cleavage site is therefore a suitable source for a statistical description of the negatives' properties.

2.2.2. Difference in peptide sequence between positives and negatives

The frequencies of amino acid residue types appearing at each position in the peptides were calculated for positives of both protease types and the combined set of negatives.

Instead of the qualitative sequence logos commonly used to plot residue-type frequencies (Crooks et al., 2004), we created a representation allowing for a more quantitative comparison of residue characteristics and identity (Figure 2.2a). A large degree of degeneracy is observed in the positives, with both GrB and caspase substrates allowing for six or more residue types appearing at frequencies >5% at six of the eight subsites in the peptide. Aside from the requirement for Asp at the P1 site, the most stringent specificities are for large hydrophobic residues at the GrB P4 site (occurring in 62% of all substrates), and for small non-polar residues at the caspase P1′ site (occurring in 74% of all substrates). Residue-type frequencies in the positives for both protease types differ from those in the negatives.

2.2.3. Enrichment of structural features in cleavage sequences

Structural features were assessed for enrichment in known cleavage sequences compared with the negatives (Figure 2.2b and c). Previous reviews of protease substrates (Hubbard, 1998) show that the cleaved sequence is more likely to be exposed to solvent, flexible, disordered and lacking secondary structure. In solved structures and comparative models, cleavage sequences are indeed more likely to be in a loop than the negatives, with 65.3% ± 13.3% of GrB sites and 65.0% ± 10.9% of caspase sites being in such a conformation compared with 52.2% ± 2.1% of the negatives. Solvent accessibility was greatest in the caspase substrates (97.3% ± 3.7% of cleavage sequences), followed by the negative set (86.5% ± 1.5%), and then by the GrB substrates (81.6% ± 10.8%). When structures or comparative models were not available, predictions gave a similar enrichment, although the magnitude of cleavage sequences in a loop conformation for all three sets was increased between 12% and

20%. This agreement in the relative distributions (Figure 2.2b and c) suggests that any errors in PSI-PRED are generally not limiting in predicting the secondary structure of cleavage sites in the substrates to which it was applied. Finally, the amount of predicted disorder (i.e. sequences that are flexible, dynamic and unresolved in an electron density map obtained by X-ray crystallography) was also greater by 12% for GrB substrates and by 37% for caspase cleavage sequences than in the negatives.

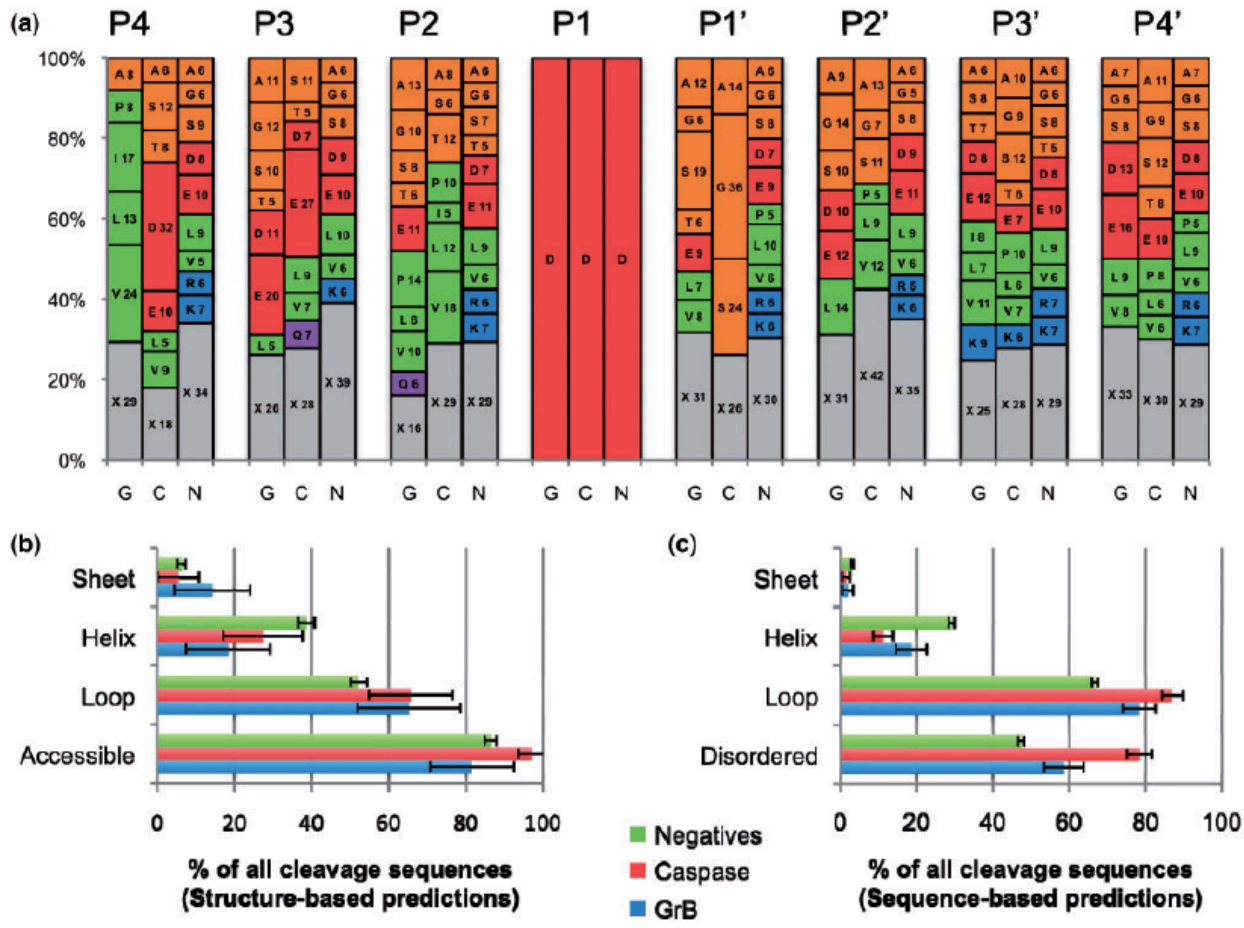


Figure 2.2 Sequence and structural properties of cleavage sequences.
(a) A stacked histogram showing the relative frequency of each residue type at each position in the cleavage sequence for substrates of GrB (G) and caspases (C), and negatives (N). The numbering spans positions from P4 to P4′. Letters on the plots represent the one-letter code for each amino acid residue type, followed by its percentage at that position. X (gray) represents the total percentage for all residue types that are present in the position at <5% relative frequency. Amino acid residue types are grouped by

general characteristic (Green: hydrophobic; orange: small non-polar; red: charged acidic; blue: basic; purple: polar).
(b) Structural properties of protease cleavage sequence positives and negatives as assessed by DSSP for substrates where a solved structure or good quality comparative model was available. Numbers may not add to 100% as some peptides did not have more than four residues in any one of the three secondary structure conformations.
(c) Structural properties as assessed by predictive methods that consider the protein primary sequence only. Disopred predicted disorder in all substrates. PSI-PRED predicted secondary structure, in cases where a structure or model of the substrate was not available.

2.2.4. Benchmarking of scoring functions

Using a jackknifing procedure and the datasets, we benchmarked a scoring function for predicting whether or not an octapeptide is a substrate of a given protease type, incorporating an SVM trained on both structure and sequence. Receiver operator characteristic (ROC) plots were generated to assess the ability of the scoring functions to distinguish between positives and negatives (Figure 2.3). The critical point of the ROC plot represents the optimal tradeoff between coverage and accuracy (i.e. the minimal combined false positive and false negative rates) and was used to compare the performance of different methods.

Due to preferences of these proteases for specific residue types around cleavage sites, as well as the enrichment of certain structural features at these sites, we hypothesized that the best classifier would incorporate these aspects of proteolysis. Indeed, the SVM trained on these features did well to discriminate between positives and negatives in the benchmark sets [Figure 2.3; 'SVM (Structure)']. The GrB benchmark set was classified with a 0.79 TPR at a 0.21 FPR at its critical point. Furthermore, these rates improved (0.87 TPR at 0.14 FPR) when the SVM was trained on all known GrB substrates but assessed on a test set consisting of only the literature-curated GrBah dataset. The caspase benchmark produced similar results on both datasets. Error bars for the FPRs across 1000 iterations were assessed and calculated

as less than 0.002 for all points; these are omitted from the figure as they are smaller than the width of the curve itself.

Due to the potential for biasing an estimate of prediction accuracy by including peptides from similar proteins in both the training and testing set, we performed the jackknifing procedure with homolog filtering. When this condition was imposed, the TPRs and FPRs did not change significantly (Barkan 2010, Supplementary Figure 2). This observation implies that including peptides from related proteins across the two sets does not significantly influence the estimate of the prediction accuracy. The likely reason is that the features used by the classifier depend on the peptides themselves and not on the proteins from which they were derived.

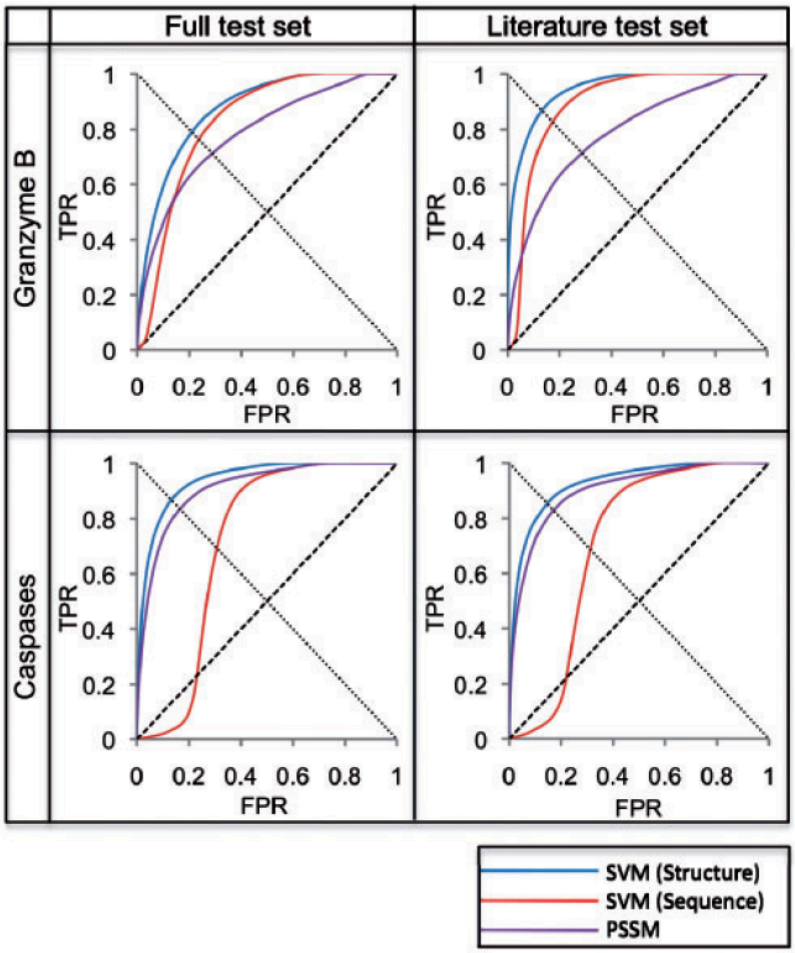


Figure 2.3 SVM benchmark results.
Results from different methods applied to four different datasets, represented by ROC curves. The line from (0,0) to (1,1) represents a random predictor; a perfect classifier would go from (0,0) to (0,1) and then to (1,1). The critical point of the ROC curve is where each curve intersects the line from (1,0) to (0,1). Full test sets included all known substrates for the respective protease type, and Literature test sets excluded the large proteomic datasets, retaining only the GrBah and Casbah substrates. SVM (Structure) was developed in the current study; SVM (Sequence) was taken from a previous study that trained on cleavage sequence residue type only (Wee *et al.*, 2006); PSSM implemented the GrabCas method for GrB substrates (Backes *et al.*, 2005), while for caspases it was trained on frequency of residue types at each position in known cleavage sequences, using the PoPS (Boyd *et al.*, 2005) algorithm. All ROC plots were interpolated through a number of points equal to the number of test set positives in each dataset (Barkan 2010, Supplementary Figure 1a).

2.2.5. Comparison with other methods

The results of the method were compared with those obtained by two previously described methods tested on the same datasets. An SVM trained on sequence only predicted GrB substrates with a 0.76 TPR at a 0.25 FPR at its critical point when assessed on the full test set [Figure 2.3, 'SVM (Sequence)']. GrabCas achieved a 0.71

TPR at a 0.29 FPR on the same test set (Figure 2.3, 'PSSM'). Similar discrepancies were observed on the GrBah test set and on both caspase test sets, here using the PoPS algorithm as the basis for the PSSM.

2.2.6. Criteria for selecting targets for experimental validation

The method was applied to all human proteome octapeptides with Asp in the fourth position to produce a score for each potential cleavage sequence. Two proteins, Apoptosis Inducing Factor 1 (AIF-1) and Survival Motor Neuron 1 (SMN1), fulfilled the following criteria for experimental followup: (i) they were not in any benchmark dataset, (ii) the corresponding mRNA was expressed in the K562 cell line (highly susceptible to granzyme-induced cell death), (iii) a validated antibody was available and (iv) evidence supported a role in apoptosis. To test whether these candidates were cleaved by GrB or caspases, K562 lysates were treated with varying concentrations of exogenous protease for either 1 h or 19 h. As the benchmark set contains substrates of both initiator and executioner caspases, a mixture composed of caspase-8 and -3 was chosen. To determine if exogenously added GrB was the causative protease, K562 lysates were pretreated with broad-spectrum caspase inhibitors before GrB addition.

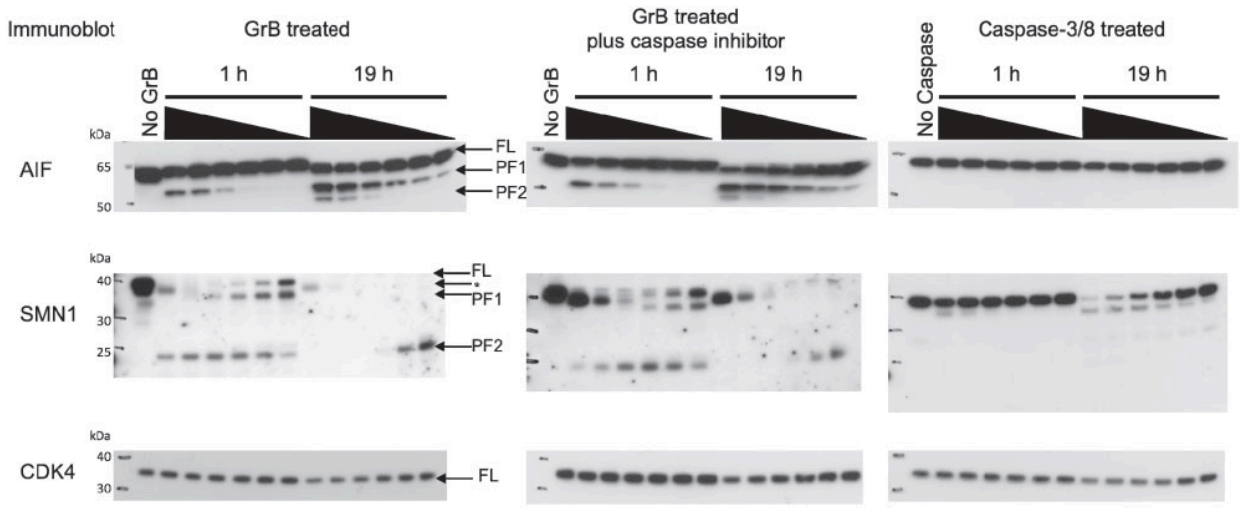


Figure 2.4 Immunoblots of predicted GrB substrates.
K562 lysates were treated with increasing concentrations of GrB or a mixture of caspase-3 and caspase-8 for either 1 h or 19 h. The final concentration of exogenously-added protease was 1 μM, 500, 250, 100, 50 and 25 nM. For caspases, the final concentration refers to the concentration of total caspase (caspase-3 plus caspase-8). The no protease controls were incubated at 37◦C for 19h to account for the activity of endogenous proteases. The caspase-inhibited lysates were pretreated with 100μM z-VAD-FMK and 100μM z-DEVD-FMK at 37◦C and then treated with GrB. Bands corresponding to full-length (FL) protein, proteolytic fragment 1 (PF1) and proteolytic fragment 2 (PF2) are indicated with arrows. Controls showed that the SMN1 antibody cross-reacts with GrB (Barkan 2010, Supplementary Figure 4). The GrB band is indicated by an arrow and asterisk.

2.2.7. Cleavage of AIF-1 by GrB

AIF-1 is a mitochondrial flavoprotein that translocates to the nucleus during apoptosis and facilitates DNA fragmentation. Interestingly, AIF-1 has a high-scoring GrB cleavage sequence (VPQD$_{126}$KAPS) that is partially solvent exposed and in a loop conformation, as determined in its X-ray structure. Addition of GrB to K562 lysates results in the appearance of a ~55 kDa proteolytic product that is both time and concentration dependent [labeled as proteolytic fragment 1 (PF1) in Figure 2.4]. A second ~50 kDa proteolytic fragment [labeled as proteolytic fragment 2 (PF2) in Figure 2.4] is detected only at the highest concentrations of GrB after 19 h. The anti-AIF1 antibody was raised against a peptide sequence derived from the C-terminus of the protein. The antibody will therefore recognize both full-length protein and any proteolytic product containing this C-terminal epitope, making cleavage at VQPD$_{126}$ the most likely explanation for the

observed 55 kDa product. AIF-1 did not contain high-scoring caspase cleavage sites. In agreement with this prediction, the same proteolysis pattern is observed when GrB is added to K562 lysates pretreated with caspase inhibitors (Figure 2.4). Furthermore, addition of exogenous caspase to K562 lysates resulted in no detectable proteolysis of AIF-1. These data indicate that proteolysis of AIF-1 is directly dependent on GrB.

2.2.8. Cleavage of SMN1 by GrB

Proteolysis of SMN1 is observed during apoptosis in neurons; one study demonstrated that cleavage occurs at ICPD$_{252}$SLDD and suggested a caspase as the causative protease (Kerr et al., 2000). When evaluated with our method, this site instead scored poorly with the caspase SVM model but scored well with the GrB SVM model (Figure 2.5). To determine if SMN1 is a GrB substrate, GrB-treated K562 lysates in the presence and absence of caspase inhibitors were immunoblotted for SMN1. Both the appearance of the ~37 kDa and ~23 kDa proteolytic products (labeled PF1 and PF2 in Figure 2.4) are caspase independent. SMN1 did contain a high-scoring caspase cleavage sequence, located six residues C-terminal to the predicted GrB cleavage site. Addition of exogenous caspase to K562 lysate resulted in the appearance of a ~37 kDa proteolytic product, consistent with cleavage near the predicted GrB site (Figure 2.4).

2.2.9. CDK4 is not cleaved by GrB

Proteins were predicted to be negatives if all candidate cleavage sequences did not score higher than a threshold defined by the SVM critical point. To determine if a predicted negative is cleaved by GrB and caspases, immunoblotting for CDK4 in protease-treated lysates was performed. In all cases, a slight reduction in the amount of

full-length protein is evident only after 19 h at 37◦C and at high concentration of exogenous protease (Figure 2.4), validating our negative predictions.

2.3. Discussion

2.3.1. Overview

In an effort to increase the coverage, accuracy and efficiency of identifying protease substrates, we developed and benchmarked a bioinformatics method that takes advantage of the current knowledge about known substrates as well as general rules of protein structure (Figure 2.1). Its predictive power was quantified by the degree to which it distinguishes between positives and negatives in a benchmark set. To demonstrate the utility of the approach, we applied it to predict novel substrates of the GrB protease and caspases, followed by experimental validation of two biologically important predictions, AIF-1 and SMN1. These results thus benefited from the synergy of sequence- and structure-based predictions combined with biological intuition to select targets for validation. The computational method has two main benefits. First, it acts as a hypothesis generator; when applied to all proteins in a proteome of interest, it produces a list of high-confidence predictions suitable for a focused and efficient experimental followup. Second, the computational method lends insight into the structural aspects that determine whether a site can be cleaved.

2.3.2. Proteome-wide prediction of protease substrates

The method was applied to all proteins in the human proteome to identify those most likely cleaved by GrB and caspases, resulting in many predictions made with high confidence. For example, the top 500 predicted caspase substrates with Gene Ontology

(GO; Ashburner et al., 2000) annotation received a score corresponding to a 0.002 FPR and a 0.110 TPR in the ROC plot (Figure 2.3). GO assignments for these sequences suggest their role in apoptosis (21 proteins), signaling (53), transcription regulation (51) and proteolysis (18), all of which are hallmarks of many known substrates targeted by caspases to induce cell death. Similar results are observed for predicted GrB substrates (Barkan 2010, Supplementary Table 2).

Once experimentally validated, these substrates lend critical insight into apoptosis. A case in point is the two GrB substrates validated in this study, AIF-1 and SMN1, which are potentially involved in two novel apoptotic pathways initiated by GrB cleavage. Prediction availability is detailed in section 6.3. Each predicted substrate site is annotated with the structural assignments that were used to make the predictions, the TPRs and FPRs for their scores, and links to the MODBASE database of comparative protein structure models to view any known structures or models of the substrate.

2.3.3. Cleavage of SMN1 and AIF-1 by GrB

The high-confidence predictions generated by this method are valuable for both streamlining experimental validation (Figure 2.4) and generating novel hypotheses regarding the roles of substrates in cell death. AIF-1 is tethered to the inner mitochondrial membrane (IMM); therefore, its translocation to the nucleus requires both mitochondrial outer membrane permeabilization (MOMP) and proteolysis of the IMM tether. The cathepsins B, S and L have been shown to proteolyze AIF-1 around residue 100, 26 residues N-terminal to the predicted GrB cleavage site (Yuste et al., 2005). The redundancy of multiple proteases liberating AIF-1 from the mitochondria might represent a strategy to overcome anti-apoptotic resistance mechanisms, such as Hsp70

overexpression. Hsp70 has been shown to inhibit import of AIF-1 to the nucleus (Ravagnan et al., 2001). GrB cleaves and inactivates Hsp70 (Loeb et al., 2006) and therefore might facilitate AIF-1 nuclear import.

SMN1 cleavage was first observed during neuronal apoptosis induced by viral infection and ischemic injury in mice (Kerr et al., 2000). Mutation of Asp_{252} to Ala abolished cleavage, leading to the speculation that caspase was the causative protease. Interestingly, SMN1 cleavage was induced by adding brain extracts from either ischemically injured or virally infected mice, raising the possibility that cytotoxic T lymphocyte (CTLs) and therefore GrB was present in the extracts.

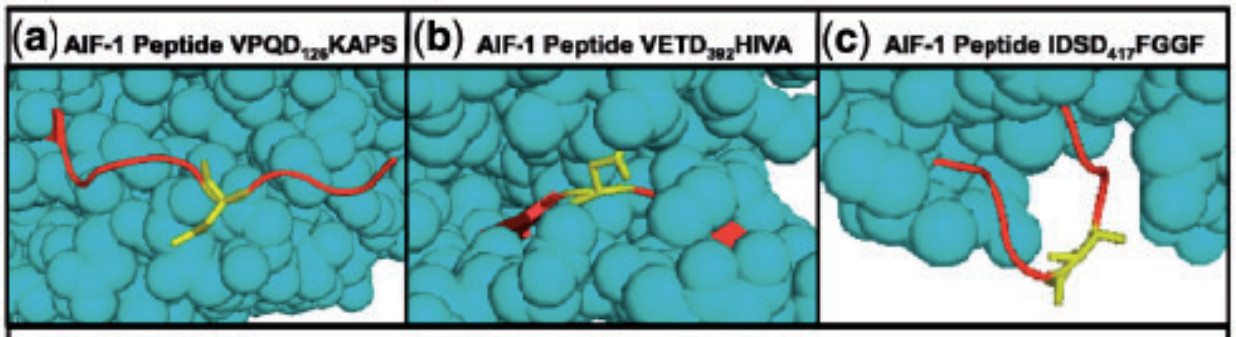
In a separate study, SMN1 cleavage has been observed in a differentiated neuronal cell line during growth factor withdrawal. CTLs are absent in this *ex vivo* study, thereby excluding GrB and implying a caspase as the causative protease (Vyas et al., 2002). Interestingly immunoblotting for SMN1 in the neuronal lysate suggested that proteolysis is inefficient, consistent with our observation that SMN1 is proteolyzed far more efficiently by GrB than the caspases. In light of evidence for a role of CTLs in both ischemic brain injury (Yilmaz and Granger, 2010) and virally infected neurons (Neumann et al., 2002), GrB should be examined as the causative protease for SMN1 cleavage in vivo.

2.3.4. Benefit of incorporating structural features in classifier training

The method was compared with several previous approaches benchmarked on the same datasets. One study using an SVM trained on sequence features did well to discriminate between positives and negatives (Wee et al., 2006), but was still outperformed by the current SVM that incorporates structure as well as sequence

features (Figure 2.3). This improvement shows that structural features of the cleavage sequence can add predictive value to a substrate identification method. Additionally, the method outperformed two other methods based on PSSMs. The first method, GrabCas, uses the results of *in vitro* small peptide libraries to predict GrB substrates (Backes et al., 2005; Thornberry et al., 1997). These *in vitro* libraries often do not fully reflect the observed protein-peptide specificity in known biological substrates. In contrast, our SVM training set does include biological substrates. The second method, PoPS, was trained only on the observed frequencies of residue types at each position in the caspase training set (Boyd et al., 2005). This PSSM does not take into account cooperativity across residue pairs. In contrast, the pair correlations can be encoded in our SVM.

It was shown previously that caspase cleavage sites can occur in regions of regular secondary structure (Mahrus et al., 2008 and section 4.2.3). Here, we show that GrB substrates display the same tendency. Indeed, >35% of known cleavage sequences in both GrB and caspase substrates fall on a region that has regular secondary structure (Figure 2.2). One possibility is that these regions undergo local unfolding prior to cleavage by the protease. These observations demonstrate the limitations of making predictions based on sequence and then filtering for expected secondary structure, as opposed to using a machine learning algorithm that makes unbiased predictions by combining sequence and structure in an integrated fashion.



(d)

SMN1 Peptide	P1	GrB	TP Rate	FP Rate	Caspase	TP Rate	FP Rate
EQEDSVLF	17	-1.38	0.97	0.59	-1.41	0.98	0.44
GQSDDSDI	29	-1.43	0.97	0.61	-2.82	1.00	0.88
QSDDSDIW	30	-1.05	0.94	0.46	-3.21	1.00	0.94
DDSDIWDD	32	-1.37	0.97	0.59	-2.01	0.99	0.66
DIWDDTAL	35	-1.60	0.98	0.67	-2.06	0.99	0.67
IWDDTALI	36	-1.02	0.93	0.45	-2.77	1.00	0.87
KAYDKAVA	44	-1.21	0.96	0.52	-2.51	1.00	0.81
KNGDICET	58	-1.62	0.98	0.68	-2.39	1.00	0.78
KVGDKCSA	96	-1.42	0.97	0.61	-2.51	1.00	0.81
WSEDGCIY	105	-1.18	0.95	0.51	-1.79	0.99	0.58
ASIDFKRE	117	-1.49	0.98	0.63	-1.55	0.98	0.49
NLSDLLSP	140	-1.50	0.97	0.64	-1.85	0.99	0.60
VSTDESEN	168	-0.71	0.88	0.33	-1.06	0.96	0.32
NKSDNIKP	181	-1.45	0.97	0.62	-2.03	0.99	0.66
ICPDSLDD	**252**	**-0.06**	**0.68**	**0.14**	**-1.55**	**0.98**	**0.49**
DSLDDADA	255	-1.42	0.97	0.61	-1.05	0.96	0.32
SLDDADAL	256	-1.57	0.98	0.66	-2.74	1.00	0.86
DDADALGS	258	-1.05	0.94	0.46	-0.43	0.73	0.32

Figure 2.5 Details of novel GrB substrates.
(a) Solved structure of AIF-1 (PDB ID 1M6I), highlighting Asp_{126}. Cleavage at this site is consistent with the observed banding patterns on the immunoblot.
(b) A peptide on AIF-1 centered on Asp_{392} that scores well when examining sequence only, but poorly when structure is considered, likely due to being largely inaccessible to solvent.
(c) Another high scoring site on AIF-1 at Asp_{417}; it is unclear why this site is not cleaved despite favorable sequence and structure properties.
(d) Scores for the SMN1 protein for all octopeptides with Asp in the fourth position, as assessed by SVMs trained on GrB and caspase substrates, respectively. The 'GrB' and 'Caspase' columns indicate the scores that the respective SVMs assigned to each peptide, and TPR and FPR signify rates that these scores would fall on in the benchmark set.

An example of the power of incorporating structure into prediction is shown by comparing two potential cleavage sequences in AIF-1, $VPQD_{126}$KAPS (Figure 2.5a) and $VETD_{392}$HIVA (Figure 2.5b). Both sites were evaluated with the sequence-based SVM (Wee et al., 2006) as well as our SVM that includes structural information.

VPQD$_{126}$KAPS, which was suggested experimentally as the GrB cleavage site (Figure 2.4), was scored with the sequence-based SVM corresponding to a 0.73 FPR. When structural features were incorporated, this site scored with much higher confidence at a 0.17 FPR. The site is on a fully exposed, flexible portion of the solved AIF-1 structure. VETD$_{392}$HIVA, on the other hand, evaluates at a 0.05 FPR when scored with the sequence-based SVM, but falls to a lower confidence 0.34 FPR when structural features are included. This site is almost completely buried and portions of it fall on a β-strand. The difference between these two sites demonstrates the importance of considering structural information when predicting protease cleavage sites. Interestingly, a third sequence at IDSD$_{417}$FGGF is not cleaved despite having favorable sequence and structure features (Figure 2.5c); further understanding of the dynamics of GrB-substrate recognition is needed to determine why this is the case.

2.3.5. General applicability of the approach

The protocol presented in this study was applied to predict substrates for GrB and caspases, two types of proteases that recognize extended, specific oligopeptide sequences possessing certain structural features. However, the approach is generally applicable to predict interaction partners for any protein that recognizes its peptide partners based on the features encoded in our method. Thus, we provide a web server (section 6.1) that allows users: (i) to construct and apply a new SVM based on a user-provided training set; (ii) to benchmark the ability of the SVM to predict interaction partners for a protein of interest; (iii) to use the newly generated SVM to make proteome-wide predictions; and (iv) to make the SVM and its predictions publically available for use by others. As a result, our approach may become a widely useful

hypothesis generator that can increase the pace of biological discovery by guiding future experiments in a variety of protein-peptide systems.

2.4. Methods

2.4.1. Structural characteristics of sequences

Datasets of known cleavage sequences were compiled for benchmarking, and all human proteome octapeptides with Asp in the fourth position were processed for the application step. Comparative models were generated by the automated modeling pipeline ModPipe (Pieper et al., 2009), and only good quality models [those predicted to have >80% of their C-α atoms within 3.5 Å of the native state, as assessed by the model evaluation algorithm TSVMod (Eramian et al., 2008)] were considered (Figure 2.1, step 2b). It has been previously shown that secondary structure features computed from accurate comparative models are similar to those for crystallographic structures (Chakravarty and Sanchez, 2004). For a solved structure or a comparative model, the DSSP program was used to assess secondary structure (mapping results 'H', 'G' and 'I' to α-helix; 'B' and 'E' to β-sheet; and 'S', 'T' and 'L' to loop) and solvent accessibility (Kabsch and Sander, 1983; Figure 2.1, step 2c). When a structure or model was not available, sequence-based algorithms were used to predict secondary structure (Figure 2.1, step 2a; Jones, 1999). A sequence-based algorithm was also used to predict disorder on all known substrates regardless of whether a structure or model was available (Jones and Ward, 2003). A cleavage sequence was defined as being in a loop if four or more of its residues were predicted to be in this conformation, devoid of regular secondary structure; similarly, a cleavage sequence was defined to be solvent accessible if four or more of its residues were >16% exposed to solvent (Kabsch and

Sander, 1983). Error bars represent two times an SD, which is calculated for a binomial experiment with $(n * p * (1 - p))^{1/2}$; values for n can be found in Barkan 2010, Supplementary Figure 1a. Training on octapeptides spanning P4 to P4′ gave the best performance relative to peptides of other lengths and positions (data not shown).

2.4.2. Scoring of potential cleavage sites by an SVM

SVMs are machine-learning algorithms that can be used for classification. They create a kernel function hypersurface that maximally separates two sets of n-dimensional training set (i.e. classified) vectors, followed by predicting an unclassified vector as falling on one side or the other of the separation. Each dimension in the vector is a feature number, which has a corresponding value. Here, a single cleavage sequence had eight features representing its oligopeptide sequence. Each residue was assigned a feature number by the formula $n*20+i$, where n represents the zero-based position in the peptide sequence of the residue and i represents the position of the residue in a zero-based alphabetical ordering of all residues. Thus, a glutamate (i=3) in the second position (n=1) would have the feature number 23. The value for all sequence features was 1.

The outputs of the structural assessment algorithms were used to create additional features for each cleavage sequence. Each of these algorithms assigned a value to each residue in the cleavage sequence. The program Disopred outputs values from 0 to 1 that correspond to the predicted degrees of disorder. DSSP outputs a calculated solvent accessibility fraction and both DSSP and PSI-PRED output a predicted structure type of loop, α-helix or β-sheet. These algorithms each added eight features to a cleavage sequence, where the structure types were assigned the values 1, 2 and 3

corresponding to loop, helix and sheet, respectively, and the other values were the raw score outputs of the algorithms.

The SVM-light software was used to execute the SVM algorithm (Joachims, 1999; Figure 2.1, step 3). A radial basis kernel function was used, sampling different values of the parameters C (selecting from 1, 10, 100 and 1000) and γ (0.01, 0.1, 1, 10 and 100) to find those that performed best in the assessment, as has been done previously (Wee et al., 2006).

2.4.3. Benchmarking of scoring by jackknifing

A jackknife procedure was employed to test different scoring functions, in which 90% of the positives for each type of protease were randomly selected into a training set, and the remaining 10% were placed in a test set, along with the known negatives. The ratio of negatives to positives in the test set was 39 : 1 for the GrB benchmark and 35 : 1 for the caspase procedure, reflecting the ratio of negatives to positives observed in respective known substrates. Scores for the peptides were ranked and the false positive rate (FPR) against the true positive rate (TPR) was assessed at different score thresholds (Figure 2.1, step 4). The jackknife procedure was repeated 1000 times and the results were averaged. Error bars for the averaged FPR μ at each TPR represent two times an SD, which is calculated over the distribution of FPRs for all iterations (x from i to N) by $((1/N)\Sigma(x_i - \mu)^2)^{1/2}$.

To ensure that random assignment of all experimentally identified peptides into different training and testing sets did not artificially influence predictive accuracy due to some similarities between the two sets, a separate jackknifing procedure was performed and compared the original to random assignment. Here, for each peptide x in the test

set, no other peptide *y* was included in the training set if *y* was derived from a protein with >25% sequence identity to the protein from which *x* was derived. These included other peptides on *x*'s protein itself. We describe this restriction as 'homolog-filtering'.

2.4.4. Comparison of the protocol to other approaches

We applied to the datasets the following published methods: (i) an SVM trained on sequence information, using the original encoding and parameter sampling scheme (Wee et al., 2006); (ii) the GrabCas method, which incorporates *in vitro* PS-SCLs into a PSSM, using default parameters; (iii) a PSSM based on the frequency of residue types appearing in each position in the training set, incorporating the generalized PoPS algorithm to score a sequence (Boyd et al., 2005).

2.4.5. Experimental validation on select substrates

The method was applied to all octapeptides in the human proteome with Asp in the fourth position. Certain peptides were selected for experimental validation using the following procedure. The expression of a predicted substrate at the mRNA level was determined by consulting the BioGPS database (https://biogps.gnf.org/; Figure 2.1, step 5). The availability of a literature-validated antibody was determined by consulting http://www.labome.com. K562 cells were grown in Iscove's modified Dulbecco's medium, 10% FBS, 1× Glutamax, 1× Penn/Strep to a density of ~5×105 cells/ml. K562 cells were harvested by centrifugation, washed in PBS, and lyzed in MPERTM (Thermo Scientific, Rockford, IL) at 1×107 cells/ml according to the manufacturer's instructions. Protein concentration was determined by BCA assay (Thermo Scientific, Rockford, IL). Pichia-expressed human GrB (Thornberry et al., 1997) and *Escherichia coli*-expressed human caspase-3 and -8 (Stennicke and Salvesen, 1999) were purified as previously

described. K562 MPERTM lysates were diluted 1 : 2 into 500 mM HEPES pH 8.0, 100 mM NaCl, 0.01% Tween-20 to raise the pH for optimal GrB activity and diluted 1 : 2 into MPER and 20 mM DTT for optimal caspase activity. GrB or a mixture of caspase-3 and -8 were added for either 1 h or ~19 h before quenching proteolysis by adding LDS sample buffer (Invitrogen, Carlsbad, CA) and incubating at 70◦C for 10min. The final concentration of exogenous protease (GrB or total caspase) was 1μM, 500, 250, 100, 50 and 25 nM. Untreated lysate was incubated for 19 h to account for the activity of endogenous proteases. Caspase-inhibited lysates were pretreated with 100μM z-VAD-FMK (Bachem, Torrance, CA) and 100μM z-DEVD-FMK (Bachem, Torrance, CA) for at least 1 h at 37◦ C and then treated with GrB as described. To verify that the exogenous protease added to the lysate was active, immunoblots against validated substrates were performed as described: pro-caspase-3 for GrB, PARP for caspase-3 and BID for caspase-8 (Barkan 2010, Supplementary Figure 3).

7μg of total protein from each protease-treated and -untreated sample were subjected to electrophoresis on denaturing and reducing NuPAGE Bis-Tris gels (Invitrogen, Carlsbad, CA). Proteins were then transferred to Polyvinylidene Fluoride (PVDF) membranes and blocked in Tris buffered saline Triton X-100 (TBST) containing 5% (w/v) milk. Membranes were then incubated with substrate-specific antibodies, washed and incubated with HRP-conjugated secondary antibodies (BioRad, Hercules, CA). Immunoblots were developed on film with the ECL Plus detection system (GE Healthcare, Piscataway, NJ). To verify that equal amounts of protein were being compared across samples, GAPDH levels were quantified in parallel with either a rabbit anti-GAPDH or mouse anti-GAPDH antibody and appropriate Cy3 or Cy5 conjugated

secondary antibody (GE Healthcare, Piscataway, NJ). Fluorescence was quantified on Typhoon Scanner (GE Healthcare, Piscataway, NJ). A representative GAPDH immunoblot is shown in Barkan 2010, Supplementary Figure 3. All primary antibodies were from either (Cell Signaling, Beverly, MA) or (Santa Cruz Biotechnology, Santa Cruz, CA).

Chapter 3. Peptide Docking

3.1. Introduction

The previous chapter presented a machine-learning algorithm that identified whether a given peptide would bind to, and thus be cleaved by, different protease families. This method addressed the question of whether binding occurred on a binary level, but described little about the mode of binding, which residue contacts occurred between molecules, and which peptide residues contributed the most to binding affinity. Additionally, the machine-learning approach relied on a large training set of known positives and negatives for its predictive accuracy. This training set is often not available in many biological systems.

A complementary approach to identify protein-peptide interaction specificity is through peptide docking. The ideal peptide docking algorithm would take as input simply the peptide sequence of interest and the protein structure and automatically determine whether the peptide binds, and if so, what the bound conformation of the peptide is. However, this problem is a challenging one due to the large degree of flexibility in a peptide, the potential for significant conformational change of an unbound protein receptor upon peptide association, and the lack of precise scoring functions to evaluate whether a bound conformation is near-native. As discussed in section 1.8, progress has been made towards accurate peptide docking, but there are still hurdles to overcome, perhaps the greatest of which is error in docking results due to optimization procedures not reaching the global minimum of the scoring function when sampling different peptide conformations. Here, we present a method that attempts to overcome this obstacle

using a divide-and-conquer scheme. This method has two main components. The first is a docking algorithm that follows a traditional optimization scheme, using molecular dynamics for sampling different conformations of the system and a combination of physical and statistical scoring restraints to evaluate each conformation. We demonstrate success with this algorithm on a small benchmark set, but note significant limitations, mostly in its ability to obtain near-native conformations for some peptide residues, but result in significant error in other areas of the peptide. The second component attempts to address this problem by employing the divide-and-conquer approach to determine accurate local regions in different steps of the MD trajectory and combine them into a global solution. While this work is still ongoing, significant progress has been made and a framework is in place to test different parameters of the algorithm. The final section of this chapter discusses the future direction this research will take.

3.2. Results

3.2.1. Benchmark complexes with different peptide lengths are selected

Protein Name	PDB Code	Peptide Chain	Peptide Length	#Peptide Atoms	#Protein Atoms
MHC Class I	1CE6	C	9	68	445
HIV Protease	1KJ4	P	9	69	337
α-Bungarotoxin	1HC9	C	13	120	221
Cyclophilin A	1AWR	I	5	40	218
Pilius FimG	3BFQ	F	15	121	423

Table 3.1 Peptide docking benchmark set statistics.

Peptide length is measured in residues; "#Peptide Atoms" and "#Protein Atoms" indicate the number peptide and protein atoms, respectively, that were used in the sampling procedure. Protein atoms represent the peptide binding site only and not the full protein.

To evaluate the performance of the algorithm, a small benchmark set was created consisting of the solved structures of five proteins each in complex with a small peptide (Table 3.1). The length in residues of the peptides ranged from five to fifteen. All peptides were in complex with a single protein chain with the exception of HIV protease where the peptide interacted with two identical protein chains. All members of the benchmark set were arbitrarily selected from the PeptiDB dataset (London, 2010), without regard to the biological context of the proteins.

3.2.2. Scoring function values are weakly correlated with RMSD Error

For each member of the benchmark set, 1,000 independent docking runs were performed, with each docking run consisting of a combination of molecular dynamics and simulated annealing to optimize the value of a scoring function based on a combination of physical and statistical restraints (Section 3.5). The resulting peptide conformation was the one scoring the lowest out of all conformations generated by the trajectory. For each of the 1,000 runs for a benchmarked complex, the score of this conformation was saved along with its RMSD error when compared with the native peptide structure. A scoring function should be designed to have its global minimum equal to the native conformation of a peptide; therefore, we first evaluated whether our scoring function potentially had this property, and also calculated the general correlation between scores of the peptides generated by a run and the RMSD error.

Protein Name	Best RMSD (Å)	RMSD of Best Score (Å)	RMSD rank of best score	Score rank of best RMSD	Score rank of native	Correlation
MHC Class I	4.64	8.76	553	12	18	0.11
HIV Protease	2.42	2.42	1	1	1	0.25
α-Bungarotoxin	4.71	7.16	126	45	5	-0.01
Cyclophilin A	2.07	5.22	267	35	700	0.26
Pilius FimG	7.40	7.98	2	11	4	0.16

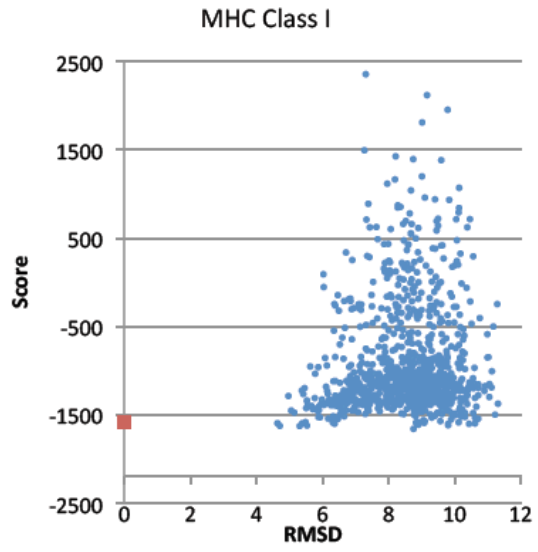
Table 3.2 Peptide docking benchmark set performance.
Best RMSD: RMSD of the conformation that is the closest to the native, regardless of score; RMSD of Best Score: RMSD of the lowest-scoring conformation. Ranks are all out of 1,000 runs. Correlation is the Pearson correlation coefficient

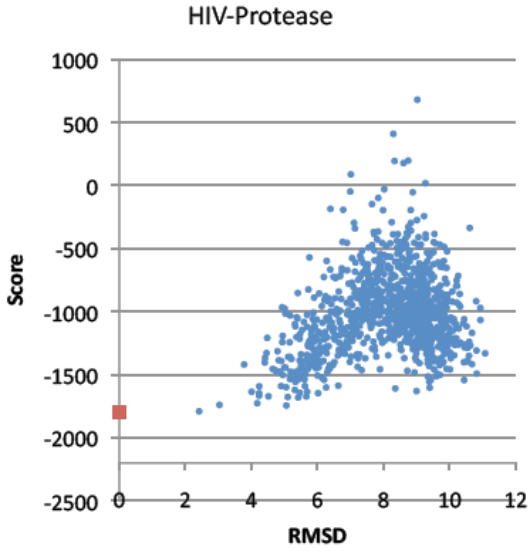
In one of the benchmark complexes (HIV protease), the native structure scored better than all peptides generated by the docking run. Evaluation on two other complexes resulted in the native peptide being outscored by fewer than five docked conformations, and a fourth native complex was outscored by nineteen docked peptides. In the final complex, cyclophilin A, the native peptide was ranked 700th, indicating that the scoring function may be insufficient to evaluate this peptide (Table 3.2, "Score rank of native").

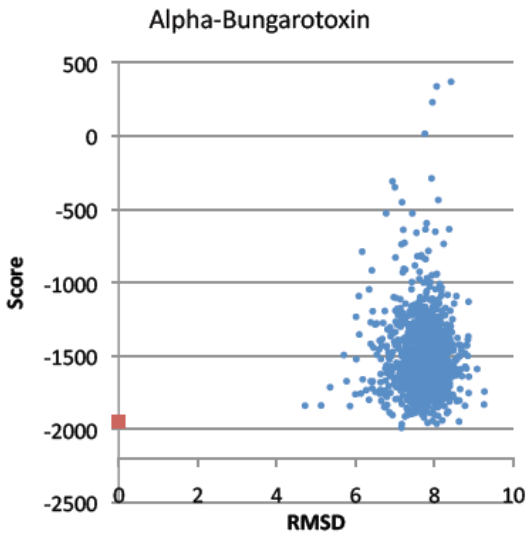
In four of the five cases, there was a weak correlation between the final docking scores and RMSD error (Figure 3.1). The value of the Pearson correlation coefficient in these cases ranged from 0.11 and 0.25 (Table 3.2, "Correlation"). The fifth case (α-Bungarotoxin) essentially had no correlation between the scores and RMSD. Within individual trajectories, as the atoms moved from initial random positions to those resembling a more biological conformation of the peptide chain, the score decreased significantly, indicating that the main inaccuracy of the scoring function is in distinguishing one near-native conformation from another.

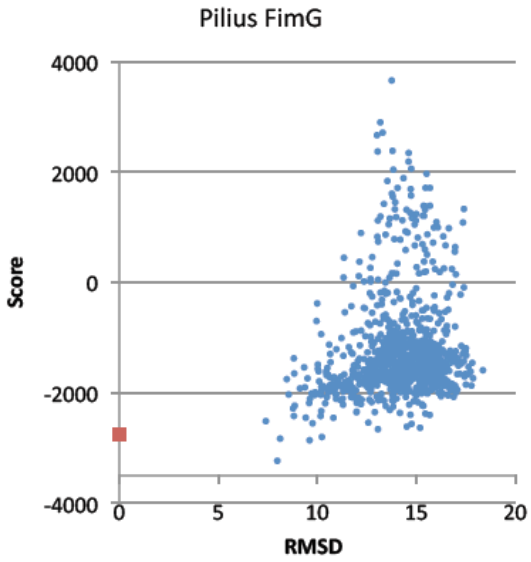
3.2.3. Regions of docked peptides are close to the native conformation

We examined the ability of the sampling procedure to find a conformation of the peptide as close to the native state as possible. For each benchmark complex, the following were evaluated: (1) the RMSD error of the lowest scoring conformation; (2) the score of the conformation with the lowest RMSD and its rank among all runs for the complex; and (3) visual inspection of the final conformation. A summary is presented in Table 3.2. An examination follows of each complex in turn.


MHC Class I
Score
RMSD
2500
1500
500
-500
-1500
-2500
0
2
4
6
8
10
12


HIV-Protease
Score
RMSD
1000
500
0
-500
-1000
-1500
-2000
-2500
0
2
4
6
8
10
12


Alpha-Bungarotoxin
Score
RMSD
500
0
-500
-1000
-1500
-2000
-2500
0
2
4
6
8
10


Pilius FimG
Score
RMSD
4000
2000
0
-2000
-4000
0
5
10
15
20

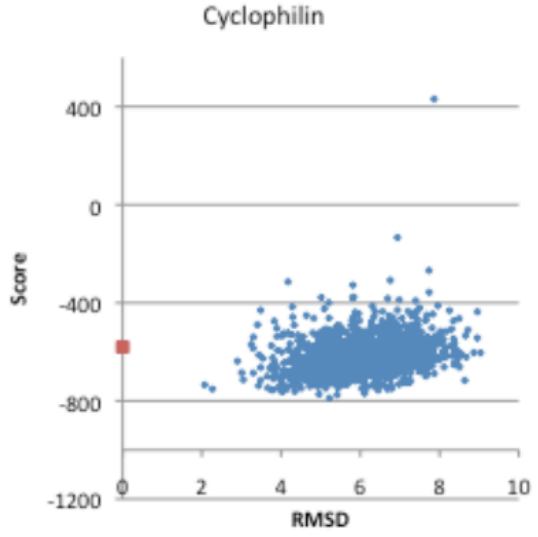

Cyclophilin
Score
RMSD
400
0
-400
-800
-1200
0
2
4
6
8
10

Figure 3.1 Scores vs RMSD of optimal docking poses
For each member of the benchmark set, 1,000 independent docking runs were performed. Each point on a plot represents the value of the best scoring conformation for one of the 1,000 runs and the corresponding RMSD error from the native complex of that conformation. All RMSDs measured in Ångstroms. Red squares indicate scores for the native complex, which by definition has an RMSD of 0Å.

α-bungarotoxin

The thirteen residue α-bungarotoxin peptide docked at an optimal conformation (defined as the conformation with the best RMSD) of 4.71Å (Figure 3.2), although the conformation receiving the best score fared worse at 7.16Å. The conformation generally preserves the overall fold of the peptide, in which the terminal ends protrude from the binding site while the center is buried. Most of the side-chains are misplaced. Despite there not being a correlation between the scores and RMSDs for α-bungarotoxin (Table 3.2), the score of the optimal conformation ranks 45th among all scores.

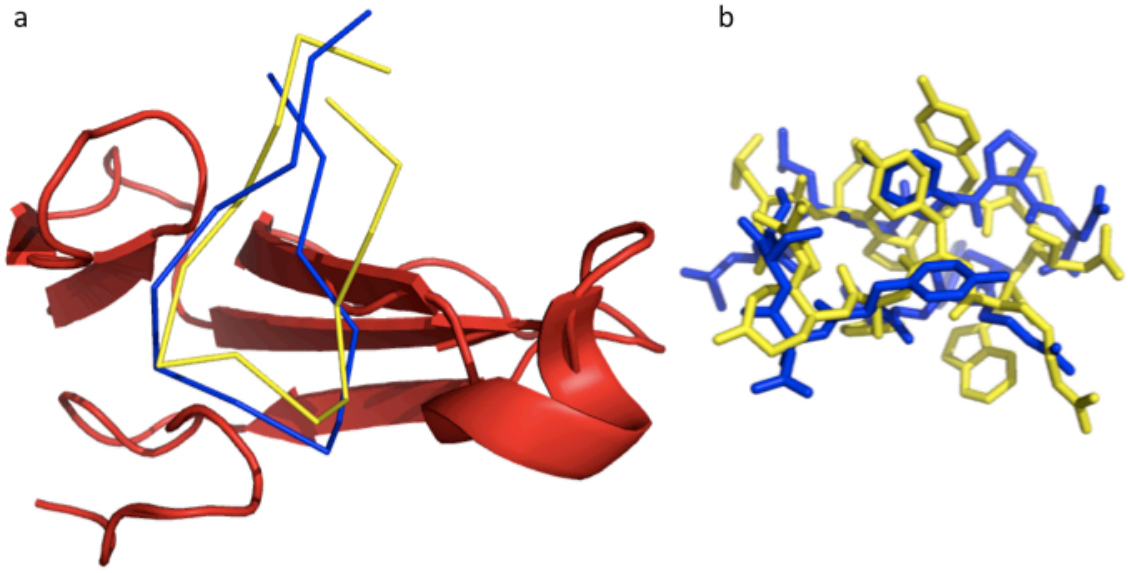


Figure 3.2 Optimal α-bungarotoxin conformation.
In this and subsequent figures, the protein is shown in red, the native peptide in blue, and the docked peptide in yellow; (a) shows the peptides in complex with the native structure of the receptor, and (b) shows the same peptide alignment without the receptor and in a different orientation.

MHC Class I

Peptides recognized by the MHC Class I receptor bind to a large pocket, here modeled as including 445 atoms in the binding site. The three C-terminal peptide residues are posed at 0.44Å from their native conformation (Figure 3.3). However, a large kink in the docked peptide at the center asparagine residue leads to the rest of the peptide not extending as long as the native, aligning with an overall error of 4.64Å. This conformation ranked 18th by score among all docking conformations.

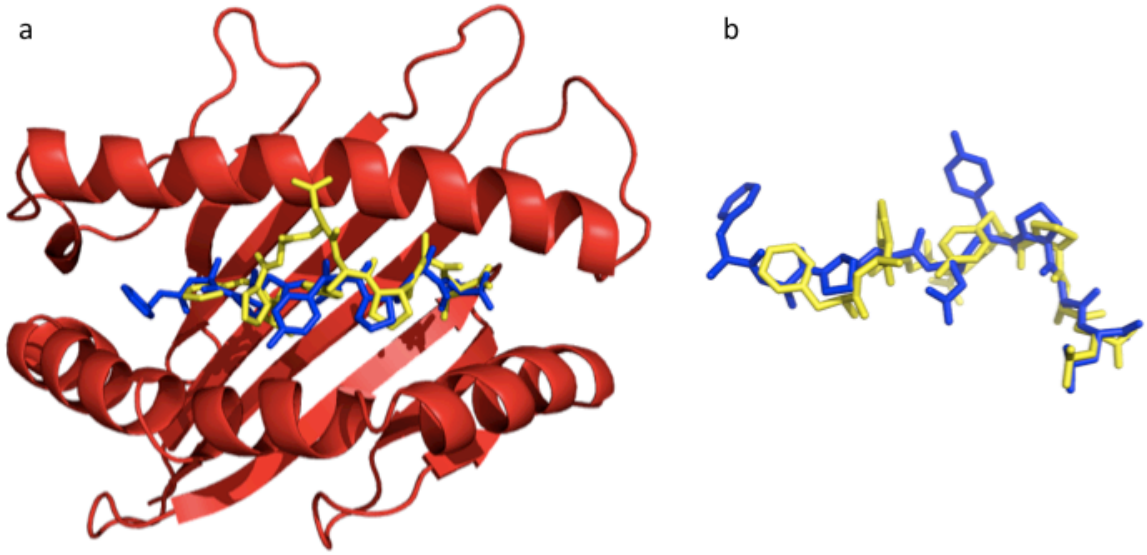


Figure 3.3 Optimal MHC Class I conformation.

HIV protease

The results for the HIV protease peptide complex docking run were the best in the benchmark set. The peptide docked at 2.42Å RMSD relative to the native, with the N-terminal glutamine residue contributing the most to the error (Figure 3.5). This residue protrudes from the peptide; the lack of restraints between peptide and protein atoms in this region is likely the reason for this error. The peptide is buried in the protease, in

contrast to other benchmark peptides that bind to an open cleft; this reduced conformational flexibility likely contributes to the accurate pose. Additionally, this optimal conformation is also the best scoring pose for the peptide.

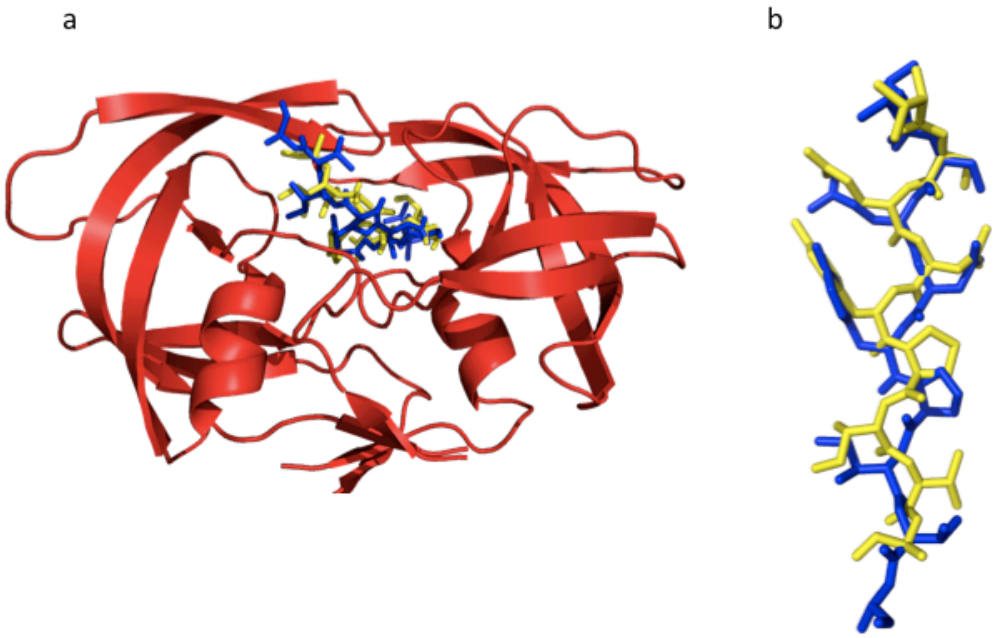


Figure 3.4 Optimal HIV protease conformation

Pilius FimG

The fimG subunit of the *E. coli* pilius assembly docks with an optimal conformation of 7.40Å RMSD error relative to the native (Figure 3.6). While this error is relatively large, there is a five residue stretch (peptide residues four to eight) that aligns at 0.56Å to the corresponding residues in the native peptide. The primary contribution to the overall error comes from the terminal ends of the peptide, which do not resemble the native conformation. FimG is a fifteen residue peptide and represents a particularly challenging docking problem, although in this case the top scoring peptide is also the second best in terms of RMSD (7.98Å).

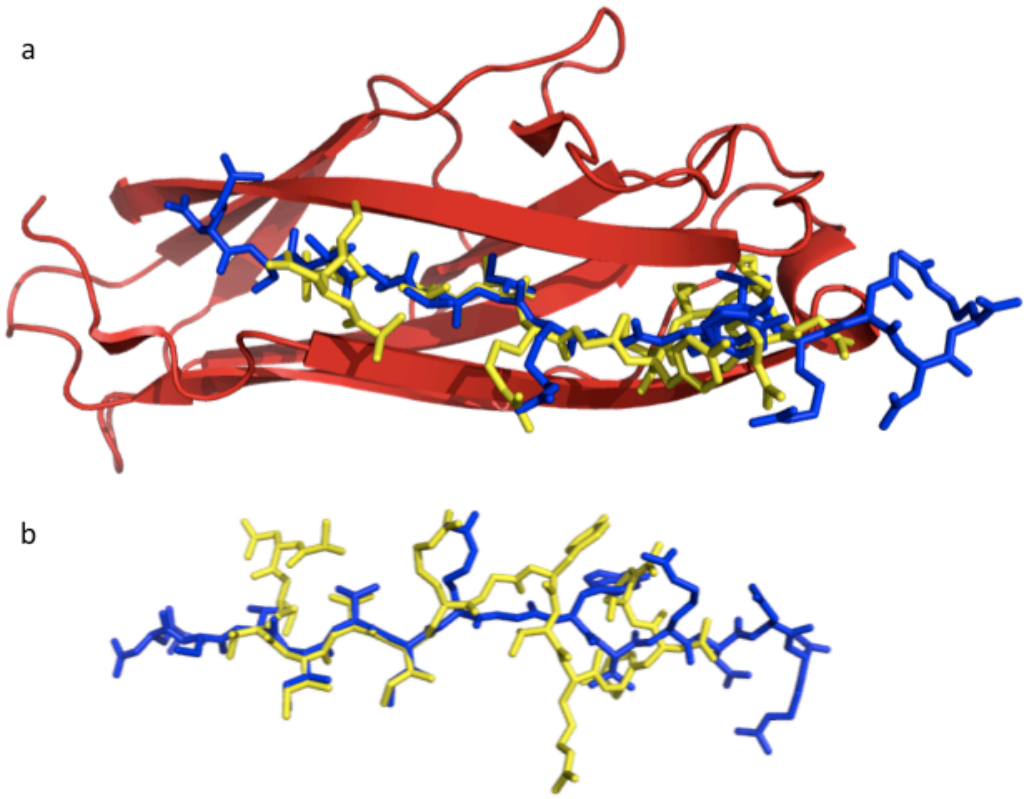


Figure 3.5 Optimal FimG conformation

Cyclophilin A

Cyclophilin A is the smallest peptide in the benchmark set at five residues. The optimal pose docks at 2.07Å RMSD relative to the native complex, making it the peptide with the lowest error among all peptides in the benchmark set (Figure 3.6). However, the best scoring peptide has an error of 5.22Å, ranking it 267th in terms of RMSD. As noted previously, the native peptide scores 700th compared to the other docking runs; therefore, the scoring function needs to be improved before a confident selection of the final peptide can be made. It is possible that the small size of the peptide, and thus a relatively small number of restraints in the system, contribute to this discrepancy between the best scores and optimal conformations.

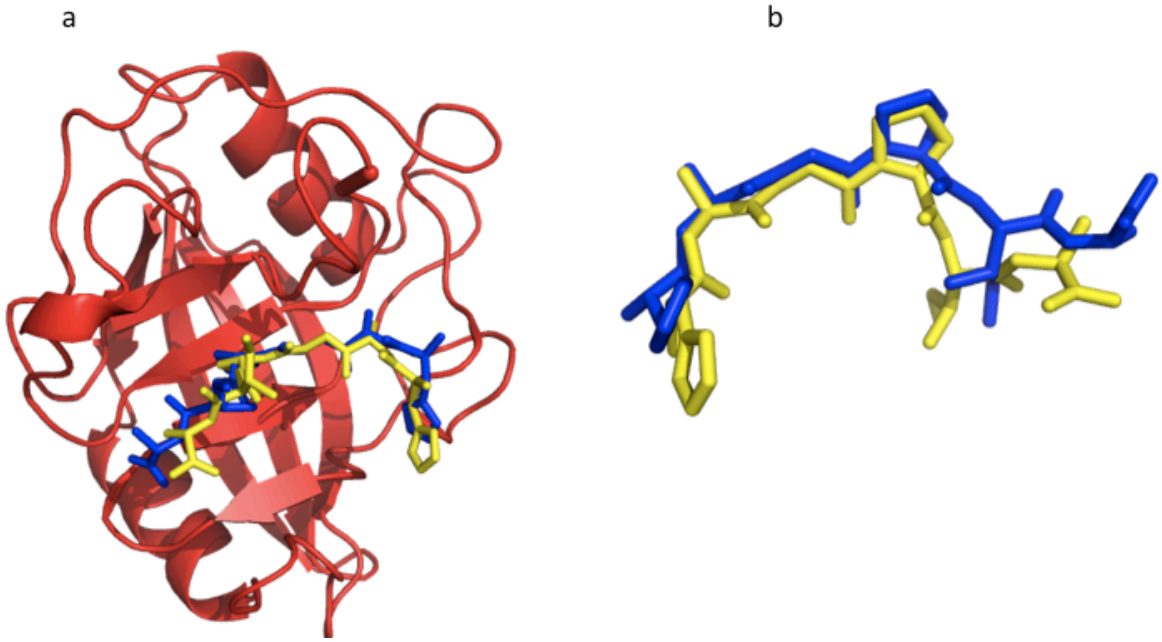


Figure 3.6 Optimal cyclophilin A conformation.

3.2.4. The DOMINO algorithm divides the system into subsets

In all of the benchmark complexes, certain regions of the peptide were close to the native conformation while others aligned with large error. While the best-scoring MD frames generally contained the former, it is possible that other frames in the trajectory contained a separate low-scoring region, but the overall score for those frames was suboptimal. A solution to this disconnect between individual low-scoring frames may come from an approach that combines individually locally optimal regions in a rigorous fashion to assemble a global conformation that scores better than any individual frame. To this end, we explored applying the DOMINO algorithm in an atomic context (Lasker, 2009).

DOMINO uses the initial restraint set as input to divide the system into overlapping subsets of interacting degrees of freedom (in this case, the three-dimensional coordinates of atoms) and evaluates the restraints acting on atoms within

each subset. The values of the restraints are drawn from the conformations generated by the trajectory. Compatible conformations of atoms across subsets are evaluated to combine the subsets into the final solution (Figure 3.7; See section 3.5 for a full description).

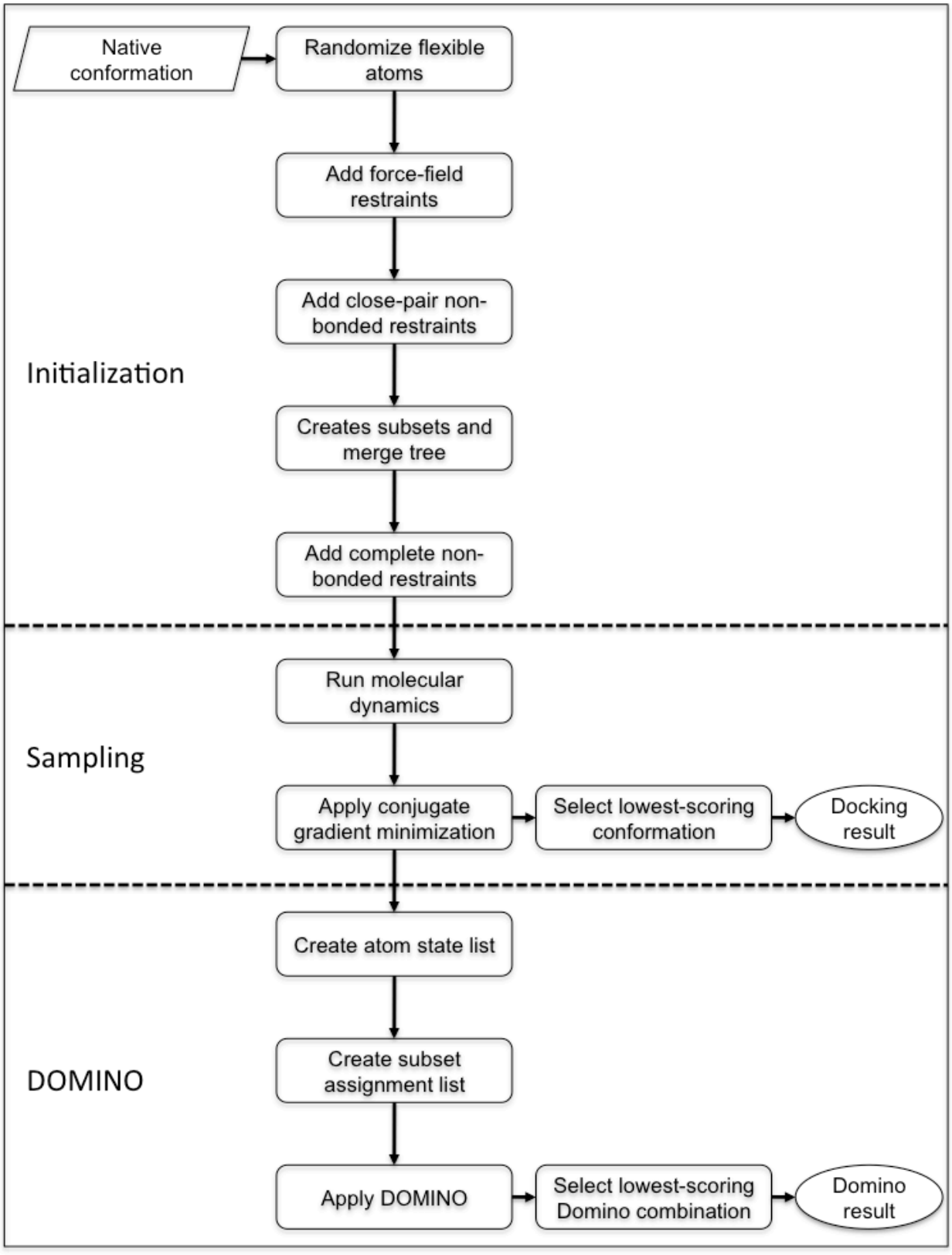


Figure 3.7 Flowchart illustrating the DOMINO procedure.

We applied DOMINO to the HIV Protease benchmark complex to evaluate its applicability. As with the MD docking procedure, we initialized the system by assigning random coordinates within the protein binding site to all peptide atoms. From here, force-field restraints were added to appropriate atoms and non-bonded restraints were added across all pairs of atoms within 6Å (Figure 3.8). This restraint graph was used to generate the subsets used in the DOMINO algorithm (Figure 3.9).

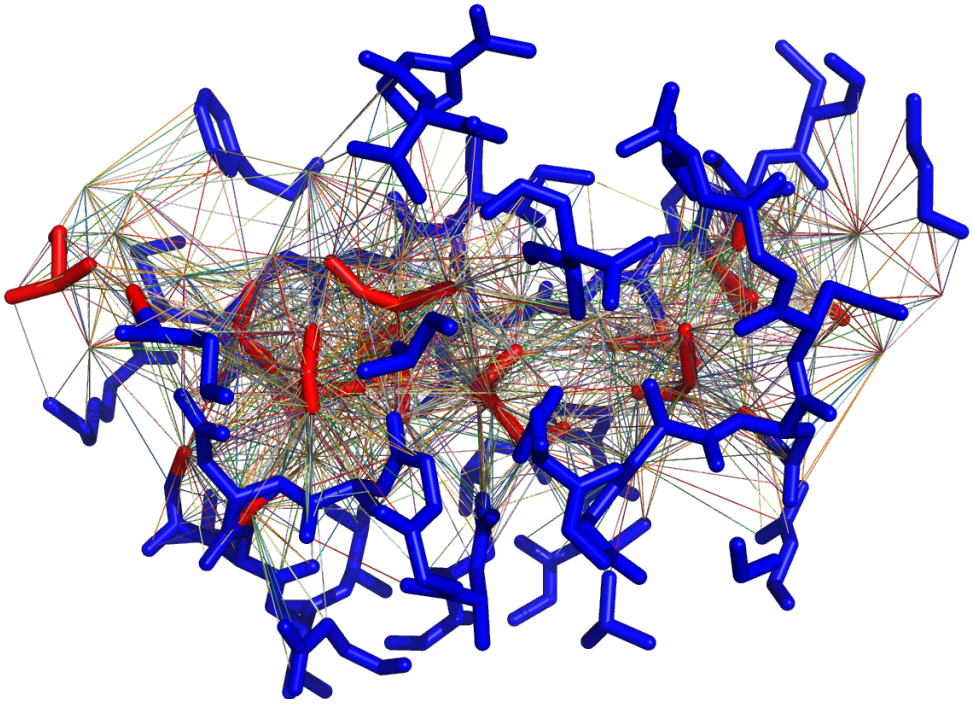


Figure 3.8 Molecular representation of the restraint graph.
Blue atoms represent the protein binding site; red are the randomized peptide atoms. Lines between atoms represent initial restraints between atoms that are used to create the junction tree. Lines terminating in empty space are associated with unbound protein atoms.

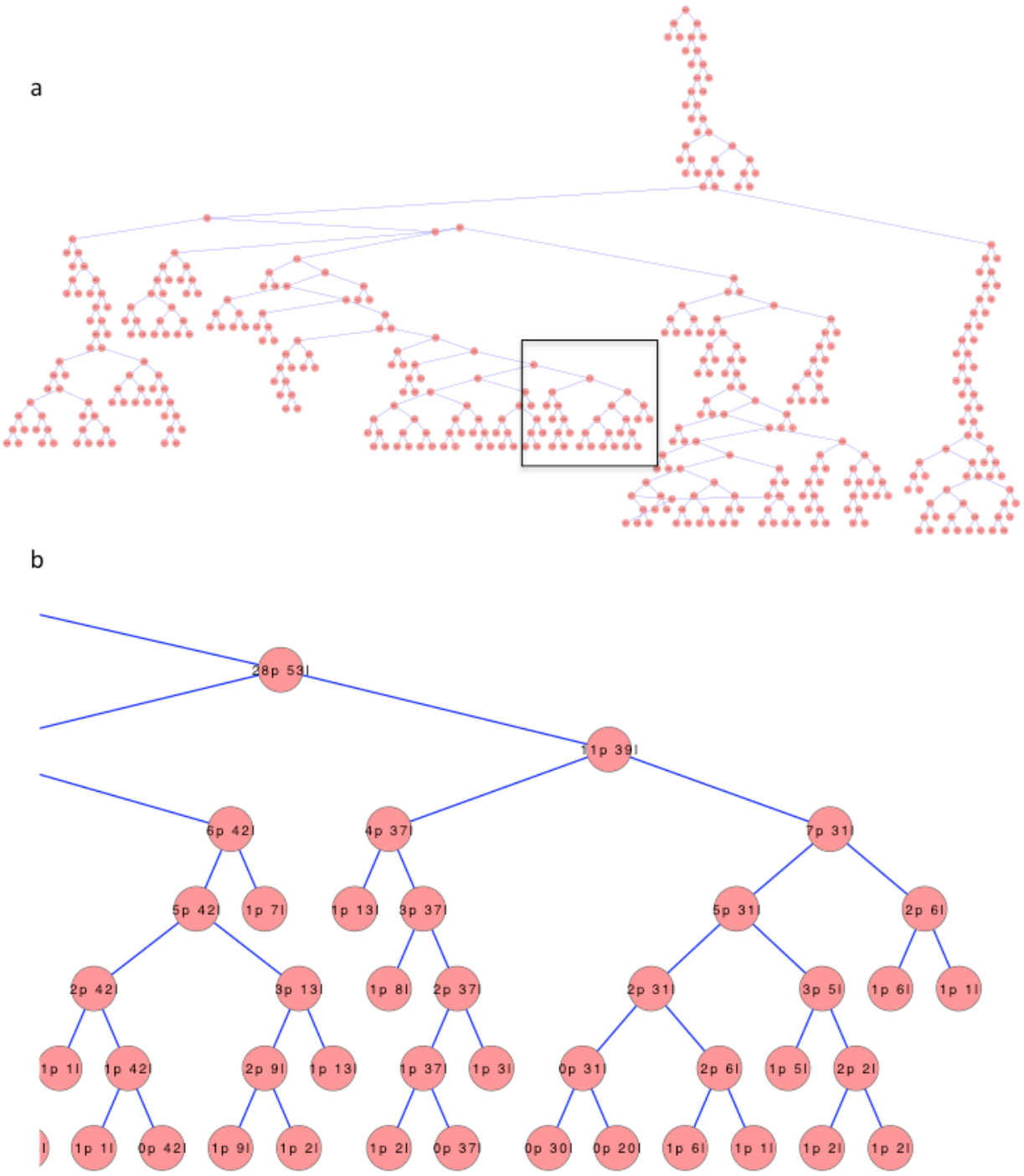


Figure 3.9 Visualization of the DOMINO Merge Tree
(a) The full merge tree, where each leaf represents a subset of atoms created in the junction tree, and each internal node is the union of atoms contained in its two children. (b) Zoomed in region of a (black square); numbers represent number of protein atoms (p) and peptide atoms (l) in that subset.

The restraint graph included 796 non-bonded restraints, which is a small fraction of the total 66,367 restraints added to all pairs of atoms prior to running the MD trajectory. The

number of subsets produced by the junction tree construction algorithm is inversely correlated with the degree of connectivity of the initial restraint graph; thus, this small number of initial restraints is appropriate. The restraint graph is used to create a junction tree, which contained 159 nodes. These nodes were set as the leaves of the merge tree (Section 3.5), which itself contained 318 nodes.

3.2.5. Domino can find a lower score better than any individual MD frame

Domino Score Rank	Best MD Score	MD RMSD	Best DOMINO Score	DOMINO RMSD	DOMINO - MD	MD - DOMINO RMSD	Assignments	Memory (GB)	Time (s)
1	-1592.55	3.47	-1592.98	3.52	-0.43	0.26	445924	3.03	6141.83
2	-1596.18	4.78	-1592.27	4.9	3.91	0.23	455083	3.21	7584.23
3	-1591.09	3.05	-1592.04	3.07	-0.95	0.05	441705	3.37	6210.02
4	-1582.97	5.59	-1579.59	5.76	3.38	0.02	445843	3.12	6428
5	-1557.82	6	-1559.49	6	-1.67	0.06	425716	3.73	5527.76
6	-1556.19	4.9	-1545.3	5.78	10.89	0.42	433562	3.38	6764.6
7	-1547.02	5.15	-1545.17	5.29	1.85	0.27	452196	3.27	8030.33
8	-1521.53	5.71	-1528.92	6.28	-7.39	0.16	444656	3.53	7396.17
9	-1522.57	6.06	-1524.9	6.07	-2.33	0.04	368800	3.01	5066.13
10	-1521.44	5.43	-1522.35	5.87	-0.91	0.92	430437	3.60	6416.14
11	-1515.81	6.16	-1515.13	6.35	0.68	0.03	408859	2.93	5806.06
12	-1514.03	5.31	-1510.04	5.37	3.99	0.41	439661	3.48	6245.24
13	-1505.86	4.51	-1509.37	4.76	-3.51	0.03	404085	2.93	5798.85
14	-1503.95	4.58	-1498.73	4.58	5.22	0.04	300720	2.45	4452.95
15	-1494.16	4.69	-1497.66	5.13	-3.5	0.1	435754	3.45	5732.88
16	-1481.48	6.14	-1483.97	6.22	-2.49	0.15	443723	3.16	6345.63
17	-1482.02	4.91	-1480.4	4.97	1.62	0.22	436912	3.53	6063.38
18	-1473.35	4.8	-1471.63	4.8	1.72	0.07	458678	3.56	6771.99
19	-1458.44	5.39	-1463.63	5.63	-5.19	0.11	424769	3.30	5524.21
20	-1456.57	6.79	-1458.14	6.96	-1.57	0.03	443448	3.40	6578.35
21	-1455.62	5.14	-1456.01	5.38	-0.39	0.05	446246	3.51	6732.83
22	-1441.29	5.29	-1439.85	6.2	1.44	0.06	300151	2.41	4815.95
23	-1441.26	5.81	-1434.52	6.2	6.74	0.58	401338	3.10	5968.79
24	-1430.2	5.39	-1431.49	5.4	-1.29	0.07	438315	3.24	6514.84
25	-1427	5.52	-1428.2	5.57	-1.2	0.03	399507	3.20	5134.92
26	-1420.52	5.17	-1427.41	5.48	-6.89	0.13	409271	3.39	4519.35
27	-1429.26	5.38	-1423.91	5.44	5.35	0.03	444909	3.19	6500.11
28	-1337.49	4.91	-1415.93	4.78	-78.44	0.78	360992	3.26	3858.47
29	-1404.45	5.49	-1412.86	5.49	-8.41	0.21	407607	3.32	8649.96
30	-1409.6	5.33	-1410.98	5.78	-1.38	0.31	447741	3.24	6788.25
251	-850.75	6.42	-1076.94	6	-226.19	1.16	206733	2.15	2618.95

Table 3.3 Domino Results.
Shown are the top 30 scoring DOMINO runs. Each is compared to the best scoring minimized MD frame for the trajectory to which the DOMINO algorithm was applied. Also shown is the RMSD of both the MD and DOMINO frame relative to the native; the difference between the two scores ("DOMINO – MD"), the RMSD between these two modeled conformations ("MD-DOMINO RMSD"), the number of conformations examined across all subsets ("Assignments") and the memory and time requirements for the run. The final line represents the example presented in Figure 3.10.

The DOMINO algorithm initializes each leaf of the merge tree with a list of all conformations of atoms in that subset as read from the MD trajectory. The set of atoms

in the interior node of a merge tree are equal to the union of atoms in each of its two child nodes, with the root node containing all atoms in the system. The protocol employs a depth-first search to set the conformations of an interior node equal to the total number of compatible conformations between the two child nodes, propagating these conformations up to the root of the tree. Any interior node thus represents conformations that are derived from multiple steps of the trajectory.

In this fashion, we used the DOMINO algorithm to combine all locally optimal conformations of the system (with each subset representing a set of local conformations) into a global conformation. We ran the algorithm 1,000 times, initializing each run with a different random configuration of peptide atoms. Overall, 36.4% of the individual runs resulted in a DOMINO configuration that had a lower score than the best scoring MD frame, including in 14 of the top 30 conformations when ranked by the score of the DOMINO solution (Table 3.3). However, only 6.73% of the 1,000 runs resulted in a DOMINO score 5% lower than the best scoring MD frame. The lowest scoring conformation from all runs was the result of a DOMINO solution that slightly decreased the optimal score from an individual frame of the trajectory.

As a proof of concept, we highlight on one iteration, which ranked 251st by DOMINO score across all runs, finishing with a 6Å RMSD error compared to the native state (Figure 3.10). In this run, the final DOMINO score was 25.6% less than that of the optimal MD frame. The two conformations deviated by 1.16Å RMSD. This result demonstrates the ability of DOMINO to generate a score significantly less than that of any individual scores in the MD trajectory.

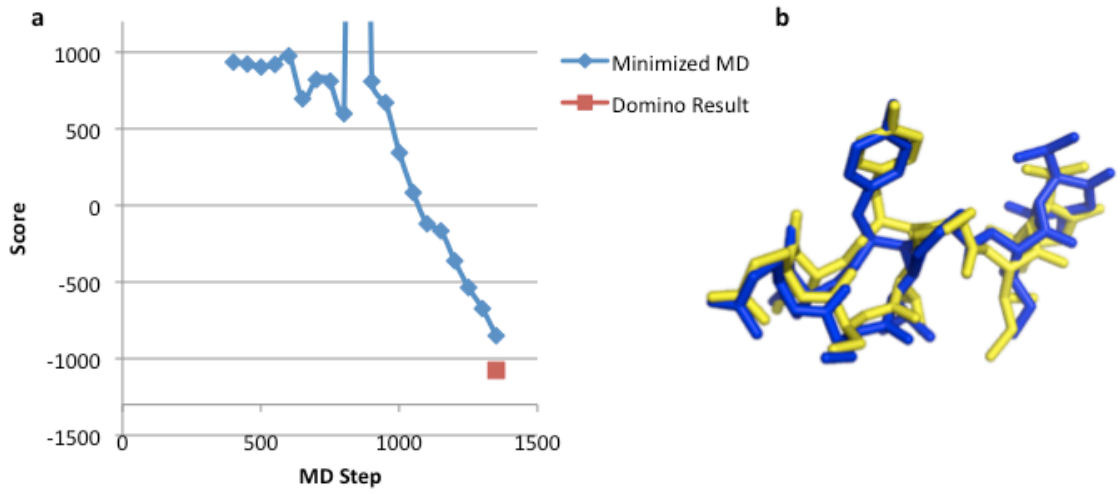


Figure 3.10 Domino proof-of-concept.
(a) For the run discussed in the text, the minimized MD over the course of the trajectory. The broken line represents a high score due to increasing the temperature in a simulated annealing procedure. The red square represents the Domino result (-1,076 compared to the lowest MD score of -850). (b) The DOMINO configuration (blue) compared to the best scoring MD conformation (yellow).

3.3. Discussion

3.3.1. Overview of progress toward an atomic level peptide docking method

We have presented two methods for docking a small peptide to the solved structure of a protein. The first is a traditional optimization procedure that attempts to minimize a scoring function using a canonical sampling algorithm, molecular dynamics. While previous docking attempts have used MD to sample peptide conformations, (Section 1.8), none has attempted a truly blind docking procedure that starts from an initial random configuration and doesn't rely on specific knowledge of the system to achieve good accuracy. In this study, the initial set of docking runs on a small benchmark set produced good results in all cases, although some were over a local region of the peptide only. Improvement to both the sampling and scoring components (discussed below) will increase the accuracy of this first docking method.

The second method is an attempt to improve the results of the sampling procedure by using the divide-and-conquer DOMINO approach to combine locally optimal and near-optimal conformations across many subsets of the system. In principle, DOMINO can take as input a trajectory generated by any sampling procedure, including the popular Monte Carlo algorithm, and apply the same merging protocol to generate a conformation scoring better than any individual conformation produced by the sampling algorithm. We have demonstrated a proof-of-concept, showing that DOMINO can indeed improve upon the scores produced by a trajectory in a large fraction of independent optimization runs, sometimes significantly. Here again, improvements to the algorithm, including junction tree construction considerations, parallelization of conformation compatibility evaluations, and iteration, will increase the accuracy of the method.

3.3.2. Fixed side chains reduce the difficulty of the problem

During benchmarking, the peptide atom positions were the only degrees of freedom in the system; the protein atoms remained fixed in their native bound conformation. In real-world docking applications, the peptide will often be docked to a native unbound conformation of the protein. While one study demonstrated that the atomic positions of most protein receptors do not vary by more than 1.5Å upon peptide complex formation, this flexibility will still be critical to account for to ensure accuracy in docking. Unfortunately, incorporating flexible side chains also increases the challenge of the problem, as it could increase the number of degrees of freedom by up to an order of magnitude (Table 3.1). This area is another where DOMINO could be useful, as more

protein atoms would be assigned to subsets and their locally optimal interactions with peptide atoms could be assessed.

3.3.3. Benchmark results illustrate the potential of DOMINO

The benchmark set demonstrated the possibility of obtaining high-accuracy local poses. In all cases, a region of the peptide docked in a conformation that was very close to the native state. For example, the three C-terminal residues of the MHC Class I structure were docked at 0.44Å relative to the native state, the HIV Protease structure aligned at less than 2Å at seven of its nine positions, and even the fifteen residue FimG structure had a stretch of five residues aligning at 0.56Å relative to the native. However, the rest of these peptides often docked with significant error. These cases are ideal for assessment by DOMINO. Assuming an improved scoring function, DOMINO has the potential to retain these conformations in a fraction of the subsets of the system, while exploring locally optimal conformations in the poor-scoring regions of the rest of the peptide. The length and flexibility of the peptide may actually be beneficial for the success of the DOMINO algorithm, as local regions of extended peptides may be generally biophysically independent from each other, and thus each region can be explored on its own and the results combined in the end.

3.4. Future Direction

3.4.1. Improvements to the scoring function

One of the primary contributions to error in the benchmark system lies in the possible inaccuracies of the scoring function. We demonstrated that the there was only a weak correlation, if any, between the score of the system and its RMSD error, with the native

peptide usually not scoring lower than all modeled conformations. A more robust scoring function will be necessary to achieve greater accuracy. One area of improvement could come from a new statistical potential to evaluate non-bonded atomic distances. The current potential, DOPE, is derived from atomic distances found in globular proteins in the PDB and is not specific to protein-peptide atomic interactions (Shen, 2006). Studies have shown that the identity and packing of atoms across the protein-peptide interface is different than in globular protein cores (Section 1.3) and a new statistical potential based on these distances could improve the accuracy of the docking procedure.

3.4.2. Improvements to the sampling procedure

In its current form, the molecular dynamics procedure proceeds in a straightforward fashion. The system is heated and cooled according to a schedule, with simple scaling of the Lennard-Jones non-bonded interaction restraints (Section 3.5). We plan several improvements to the procedure. First, an initial conjugate minimization procedure will be applied to the system to relax it before running the MD trajectory. Second, the velocities will be capped at low levels to prevent the system from exploding when the temperature is increased, similar to the caps implemented in the MODELLER protocol (Sali 1993). Third, various further scaling of Lennard-Jones restraints will be explored, with the effects of the restraints being reduced and increased at various time points. These improvements should lead to a more robust exploration of the energy landscape and result in fewer runs being immediately discarded due to the simulation not being able to handle a bad initial starting conformation.

3.4.3. Parallelization of DOMINO

One drawback of DOMINO is the large amount of memory it uses to save conformations across subsets, and the CPU time required to merge compatible conformations at step in the recursive process (Table 3.2). A solution to this problem lies in parallelization of the algorithm. Each subset contains inherent concurrency, as processing one internal node requires knowledge of the conformations of its child subsets only. Thus, each subset could be assigned to a single processor, and many subsets could be evaluated in parallel. Additionally, conformations between two subsets are evaluated in an all-vs-all pair-wise fashion. Thus, two subsets each containing *n* conformations require up to *n* vs *n* comparisons. However, as each of these comparisons is independent, groups of conformations could be assigned to different processors (for example, 100 processors could each evaluate (*n* / 10) *vs* (*n* / 10) conformations). These methods should greatly improve the speed of DOMINO as well as increase the number of conformations it can consider overall.

3.4.4. Iterative DOMINO

Many docking procedures approach the problem in an iterative fashion, with a coarse-grained approach being followed by refinement of an initial docking pose (Section 1.8). We will explore implementing DOMINO in a similar fashion. As the restraint graph is drawn based on the initial random conformation, the DOMINO subsets may not always include protein and peptide atoms that interact at close distances, which reduces the effectiveness of DOMINO. One solution is to run one iteration of sampling and DOMINO divide-and-conquer to produce an intermediate result, and then redraw the restraint graph and create new subsets based using this updated conformation. More sampling

and DOMINO can be applied, until a convergence criteria is met. Additionally, the resolution of the discrete grid that DOMINO uses can be increased as each iteration is run, allowing the system to proceed from coarse-grained to high resolution.

3.5. Methods

Here, we describe both the methods for the canonical peptide docking procedure (sections 3.5.1 to 3.5.4) as well as the DOMINO algorithm (sections 3.5.5 to 3.5.8).

3.5.1. Initialization of the system

For each member of the benchmark set described in Section 3.2.1, the peptide chain was identified and its atoms were defined as the flexible atoms in the system. The protein atoms were all kept fixed in their native conformation. The initial positions of the flexible atoms were randomized by the following procedure: for each peptide atom, set its coordinates to be equal to that of a randomly selected atom. This ensures that the initial position of each peptide atom was in the peptide binding site, but sufficiently random to make the problem difficult.

3.5.2. Generation of a scoring function

Bond lengths, angles, dihedrals, and impropers were all restrained using the CHARMM force-field for stereochemistry. The distances between all pairs of fixed atoms and flexible atoms, as well as all pairs of flexible atoms and flexible atoms, were restrained using a Lennard-Jones potential in combination with the DOPE potential (Shen, 2006). The values of the atomic radii used in the Lennard-Jones potential were scaled as discussed below. The values of these restraints were summed to produce a score for the overall conformation.

3.5.3. Sampling of the system

The system was sampled using standard molecular dynamics (MD) with a Verlet integrator. 2200 4 fs time-steps were run in a simulated annealing protocol, changing the temperature at every 200 steps using the following schedule: 250, 400, 700, 1000, 1000, 800, 600, 500, 400, 300, 200 (all temperatures in Kelvin). During this time, the values of the atomic radii in the Lennard-Jones potential were scaled to 0.1 times their normal size to permit flexible atoms to pass through one another. Following this, an additional 1000 steps were run during which time the values of the radii were scaled every 200 steps, from 0.1 to 0.6. A velocity cap of 1.0 A / fs was imposed on the system for the first 200 MD steps to prevent initially frustrated atoms restrained by harmonic potentials from moving too far away. Additionally, the system was minimized every 25 MD time-steps with up to 100 steps of conjugate gradients (CG). This full run output two optimal structures, selected from all sampled peptide conformations; the first was that receiving the best score according to the restraint set, and the second was that with the smallest RMSD value when compared to the native state.

3.5.4. Selection of final output structures

The steps described above (initialization, scoring, sampling) were repeated 1,000 times for each benchmark complex. Thus, each complex was sampled 1,000 times using a different random starting conformation, which generated two output conformations (best scoring and smallest RMSD) each time. The conformations representing the best score and smallest RMSD among these 2,000 output structures were determined to be the final output structures and were reported in Table 3.2.

3.5.5. Overview of the DOMINO procedure

The DOMINO procedure has previously been described extensively (Lasker, 2009). Briefly, the system is represented as a graph where the nodes are the degrees of freedom to be sampled (here, the three-dimensional coordinates of the atoms) and the edges are the restraints acting on the atoms. The graph is triangulated and then decomposed using an explicit junction tree construction algorithm. A junction tree is a graph created from the triangulated restraint graph, where the nodes are subsets of atoms, representing the maximal cliques (*i.e.*, fully connected atoms) from the restraint graph, and edges are added between some subsets if the subsets share an atom. If two subsets *x* and *y* share an atom but aren't connected by an edge, then the junction tree property guarantees that all subsets along the path connecting *x* and *y* will also contain that atom. In this study, a further refinement of the junction tree is included, known as the merge tree. The merge tree is a binary tree where the leaves are the subsets of the junction tree and each internal node contains the union of atoms in its two children. Thus, the root of the merge tree contains all atoms in the system. Following merge tree creation, the possible conformations of atoms in each leaf are generated and compatible conformations are propagated up through the merge tree to the root, as discussed in detail below.

3.5.6. Generation of the merge tree

The restraint graph is created following initialization of the system, where the coordinates of all flexible atoms are set (section 3.5.1). Force field restraints are added across appropriate atoms, and then DOPE restraints are added between all non-bonded pairs of flexible atoms, as well as all pairs of fixed and flexible atoms within a cutoff of

6Å. These restraints and the atoms on which they act comprise the restraint graph. Next, the restraint graph is triangulated, the system is decomposed into subsets that make up the junction tree, and the merge tree is derived from the junction tree. Finally, non-bonded restraints are created between all pairs of fixed and flexible atoms in the system exceeding the 6Å cutoff in preparation for sampling; note that these restraints are not represented in the DOMINO restraint graph.

3.5.7. Generation of assignments

Following restraint creation, the system is optimized according to the protocol described in section 3.5.3. This procedure generates a number of conformations of the system, one for each step of MD and CG. Next, for each of these conformations, the position of each atom is snapped to a point on a grid of 0.1Å resolution. This position is saved as what we denote a “state” of the atom. Thus, for each atom, a number of states is saved equal to the number of discrete grid positions that atom occupied in the trajectory; note that the same state can be found in multiple trajectory steps if the atom didn’t move far enough in a single step, or if it revisited a previous state. Finally, for each subset in the junction tree (and thus each leaf of the merge tree), the list of states of its atoms in each step of the trajectory is saved as what we denote an “assignment”. For example, if a subset as three atoms, then the assignment representing the first step of the trajectory would be [1, 1, 1]. If in the second step of the trajectory, the third atom moved into a new state, but the first two stayed in their first states, then the next assignment would be [1, 1, 2]. In this fashion, for each subset, a large number of assignments are generated, one for each step of the trajectory (duplicate assignments are discarded).

3.5.8. Merging of compatible assignments

After creation of all assignments for the leaves of the merge tree, the DOMINO inference algorithm merges compatible assignments across all merge tree nodes. By definition, the atoms in an internal node are the union of the atoms in each of its children (whether the children are leaves or are themselves an internal node). For an internal node, all assignments in its first child are compared to all assignments in its second child and a list of compatible assignments is generated. Two assignments are compatible if the overlapping atoms in the two children are in the same state in both assignments. Note that in these two assignments, the states of the non-overlapping atoms of the first child could be derived from a different trajectory step than were the states of the non-overlapping atoms of the second child. Therefore, a compatible assignment could contain states from different trajectory steps. This list of compatible assignments is saved as the assignment list for the internal node.

A recursive depth-first search is used to visit all nodes in the merge tree, propagating compatible assignments up through the tree until the root is reached. Additionally, in this procedure, only the top 10,000 scoring assignments are saved for each internal node due to memory considerations. We are exploring ways to optimize memory usage and increase this number, as the assignments discarded by this process could become optimal as the propagation proceeds.

When the root is reached, the states in each assignment are translated back into the atomic coordinates to which they were mapped. The system is evaluated using the scoring function, and the final DOMINO result is the conformation that is evaluated with the lowest score. As noted above, assignments may include states of atoms derived

from multiple trajectory steps, and in the root, this phenomenon is almost always the case. Thus, the DOMINO procedure produces a conformation that is drawn from multiple trajectory steps, and this conformation will often be evaluated with a lower score than any single trajectory step.

3.5.9. Integrated modeling platform

The docking and DOMINO procedure are both implemented in the Integrated Modeling Platform (IMP), which is a software suite for modeling protein and assembly structure using sampling algorithms and scoring functions in a modular fashion. IMP is open source and freely available; for more information, see www.integrativemodeling.org.

Chapter 4. Analysis of protein-peptide specificity determined by mass spectrometry-based proteomic experiments

The previous two chapters presented algorithms for identifying novel protein-peptide interactions. These methods are designed to complement experimental efforts, both by using existing experimental results to predict new interactions in the same system (Chapter 2) or by guiding new experiments by identifying the critical residue interactions across molecules upon analysis of accurate conformations generated by peptide docking (Chapter 3). Another important area where computational methods can contribute to experiments is in rigorous statistical analysis of large experimental datasets. This research provides insight into aspects of peptides that contribute to specificity, whether a simple result such as the distribution of residue types at certain positions in the peptide, or something more complicated such as the distance distribution of these peptides through a three-dimensional structure.

In this chapter, we perform such analysis on proteomic mass-spectrometry datasets of post-translational modifications of peptides. Mass-spec is an ideal experiment for determining protein-peptide interaction specificity, with the capacity to identify hundreds or even thousands of interactions in the proteome. While some may be not be physiologically relevant (*i.e.*, the peptide substrate is modified by the protein but this modification has no phenotype), and some may be artifacts of the experiment (for instance, if the peptide and protein are in separate cellular compartments *in vivo* but the experiment operates in a cell lysate), they are nevertheless biophysical interactions that are identified by experiments and thus can be analyzed statistically to gain insight into the forces mediating their specificity.

The two protein types studied in this chapter are the pro-apoptotic proteases caspases, which were also analyzed in Chapter 2, and the O-GlcNAc transferase, which modulates signaling pathways through the addition of a simple sugar to its target substrates. Large mass-spec datasets of modified peptides were generated using novel bioengineering technology, and the results were analyzed in a number of bioinformatic techniques. Together, these studies demonstrate how experiments and computational approaches can work together to identify aspects of protein-peptide interaction specificity on a large scale.

4.1. Introduction – Caspases and proteomics

The widespread intracellular proteolysis that is a hallmark of apoptosis is predominantly mediated by the caspase protease family. Apoptosis can be induced by extracellular death ligands, such as Fas ligand, TNF-a, or TRAIL, via the extrinsic pathway to activate caspase-8. It can also be induced by agents such as cytotoxic compounds, radiation, and other environmental stresses via the intrinsic pathway with release of proapoptotic factors from mitochondria to activate caspase-9. As discussed in section 2.1, caspase-3 can be activated through proteolysis by Granzyme B as a result of natural killer cell activity. Initiator caspases-8 and -9 in turn activate executioner caspases, among them caspases-3 and -7. Caspases then catalyze a multitude of proteolytic events to inactivate prosurvival and/or antiapoptotic proteins and activate antisurvival and/or proapoptotic proteins. This proteolysis results in apoptotic cell death and clearance of apoptotic bodies by phagocytes.

Here, we expand the focus of caspase substrate specificity introduced in Chapter 2. Because the study of apoptotic pathways has ramifications for development of

therapies for treatment of cancer, there is significant interest in gaining a better understanding of caspase activity during apoptosis. For example, identification of new targets of proteolysis in apoptosis can lead to the discovery of prosurvival and/or antiapoptotic factors, which can lead to identification of chemotherapeutic targets. Over 300 publications describing a wide variety of cell types and apoptotic inducers have reported the proteolysis of approximately 360 human proteins in apoptosis (Lu □ thi and Martin, 2007). Adding to this complexity, the nature of the apoptotic response varies widely in a cell-dependent and stimulus-dependent manner that cannot be easily predicted (Fulda et al., 2001; Stepczynska et al., 2001; Wiegand et al., 2001). Thus, combined data sets of caspase substrates from studies using varied inducers and cell types have limited use for understanding how a single inducer can cause apoptosis in a particular cell type.

We have developed an enzymatic approach for global profiling of proteolysis and sequencing of cleavage sites in complex mixtures that is based on positive selection of protein fragments containing unblocked α-amines, characteristically produced in proteolysis. This positive selection is enabled by use of an engineered peptide ligase termed subtiligase to selectively biotinylate unblocked protein α-amines with absolute selectivity over ε-amines of lysine side chains. We have used this method to sequence 333 cleavage sites in 292 different protein substrates targeted by caspase-like proteolysis in Jurkat cells after intrinsic induction of apoptosis with the classic chemotherapeutic etoposide. Through bioinformatic profiling of the proteolysis that is induced by a single agent in a single cell line, this work reveals the vastness of caspase-like proteolysis that takes place during apoptosis, sheds light on determinants

of specificity for this activity in a cellular context, and demonstrates the utility of a powerful degradomic technology to study proteolysis in biological samples.

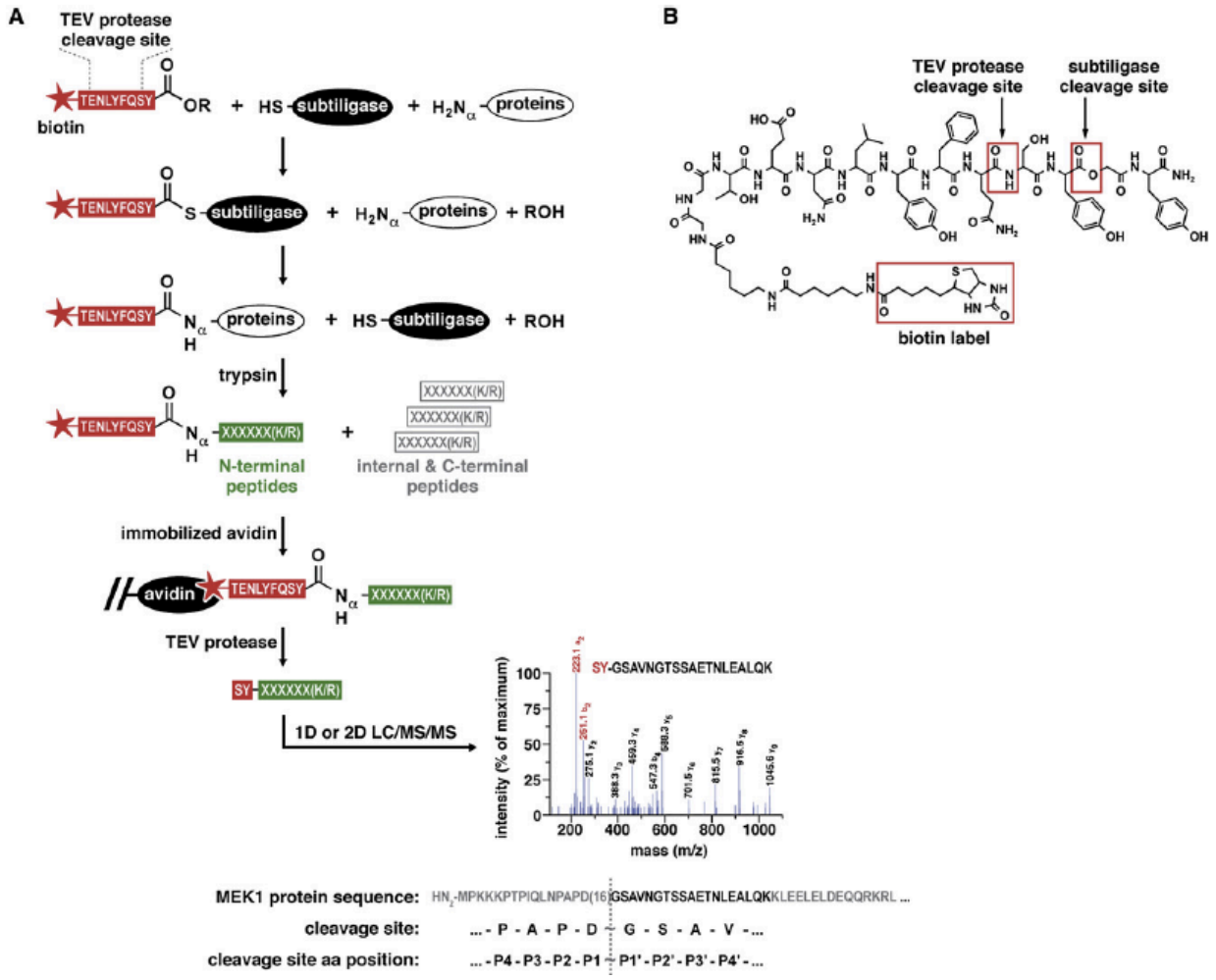


Figure 4.1 Positive selection of peptide N termini of proteins from complex mixtures
(a) Workflow for biotinylation of protein N termini in complex mixtures using subtiligase and a biotinylated peptide ester that contains a TEV protease cleavage site, trypsinization of labeled proteins, capture of biotinylated N-terminal peptides with immobilized avidin, recovery of captured peptides with TEV protease, and analysis of N-terminal peptides by 1D or 2D LC/MS/MS for identification of corresponding proteins and cleavage sites. The representative MS/MS spectrum corresponds to semitryptic peptide GSAVNGTSSAETNLEALQK from MEK1 (MP2K1_HUMAN) and identifies a previously unknown caspase-like cleavage site at Asp_{16}. The a_2 and b_2 ions at m/z 223 and 251 are characteristic hallmarks of a ligated, SY-bearing, N-terminal peptide.
(b) Structure of the biotinylated peptide glycolate ester used in the proteomic workflow.

4.2. Results – Caspase cleavage sites and analysis

4.2.1. The degradomic technology allows for positive selection of protease substrates

Direct and selective labeling of protein α-amines or α-carboxylates is a powerful approach for profiling proteolysis in complex mixtures since it permits direct identification of cleavage sites in protein substrates. Approximately 80% of mammalian proteins are known to be N-terminally acetylated (Brown and Roberts, 1976). Thus, greater signal over background can be achieved through N-terminal instead of C-terminal labeling. However, such labeling must still be extremely selective for α-amines over lysine ε-amines, which are approximately 25 times more abundant in an average protein. To achieve this selectivity, we have adopted an enzymological approach that makes use of the rationally designed protein ligase subtiligase. This engineered enzyme exhibits absolute selectivity for modification of α-amines (Abrahmse´ n et al., 1991; Chang et al., 1994).

We have developed a proteomic method utilizing subtiligase that enables capture and sequencing of N-terminal peptides found in complex biochemical mixtures (Figure 4.1a). Proteins in biological samples are N-terminally biotinylated by treatment with subtiligase and a peptide glycolate ester substrate specially tailored to our proteomic workflow (Figure 4.1b). Biotinylated samples are exhaustively digested with trypsin, and N-terminal peptides are captured with avidin affinity media. The peptide ester substrate contains a tobacco etch virus (TEV) protease cleavage site to permit facile recovery of captured peptides. An important aspect of our workflow is that recovered peptides retain an N-terminal SY-dipeptide modification, providing a key hallmark to distinguish labeled

peptides from contaminating unlabeled peptides with tandem mass spectrometry (LC/MS/MS). In standard protease nomenclature, substrates are cleaved between the P1 (N-terminal) and P10 (C-terminal) residues, with P*n* and P*n*' residues increasing in count by one in both directions away from the scissile bond (Schechter and Berger, 1968). Thus, the P*n*' residues of a cleavage site correspond to N-terminal residues of the labeled peptide identified, whereas the P*n* residues of a cleavage site can be inferred from the protein sequence preceding the identified peptide.

As a validation of this method, we analyzed endogenous N termini in nonapoptotic Jurkat cells in two small-scale experiments using one-dimensional reversed-phase (1D) LC/MS/MS and two large-scale experiments using two-dimensional strong cation exchange/reversed-phase (2D) LC/MS/MS (summarized in Tables S1 and S2 available online). Comparison of data obtained in both types of experiments is informative since 1D LC/MS/MS typically results in identification of abundant N termini, whereas the increased proteomic coverage afforded by 2D LC/MS/MS results in additional identification of lower-abundance N termini. Of the combined 131 unique N termini identified in small-scale experiments, 72% are either annotated in SwissProt as native protein N termini or correspond to cleavages within the first 50 residues of proteins, as would be expected for N-terminal signal or transit peptide processing (Mahrus 2008, Figure S1A). The remaining 28% correspond to cleavages outside of the first 50 residues, arising from additional processing or constitutive protein degradation. In support of this notion, 51% of the combined 661 unique N termini identified in large-scale experiments correspond to cleavages outside of the first 50 residues (Mahrus 2008, Figure S1A). The increased frequency of such N

termini in large-scale experiments is consistent with the expected lower abundance for products of constitutive protein degradation.

4.2.2. Degradomic analysis of apoptotic jurkat cells

For analysis of apoptosis in Jurkat cells, we conducted several small-scale (1D) and large-scale (2D) LC/MS/MS experiments (representatives are summarized in Tables S3 and S4) with cells treated with the topoisomerase II poison etoposide. The experiments with untreated cells described above serve as respective controls for the small and large-scale experiments with apoptotic cells, in which a combined 244 and 733 unique N termini, respectively, were identified. Combined data sets of all N-terminal peptides identified in untreated and apoptotic Jurkat cells, respectively, are included as supplemental data (Mahrus 2008, Tables S5 and S6). Caspases are known to exhibit strict substrate specificity for aspartate at P1, and for glycine > serine > alanine at P10 (Schilling and Overall, 2008; Stennicke et al., 2000). In small-scale experiments, 43% of N termini identified in apoptotic cells were derived from P1 aspartate cleavages, in contrast to less than 1% in untreated cells (Figure 4.2a). In large-scale experiments, 43% of N termini identified in apoptotic cells were derived from P1 aspartate cleavages, in contrast to 3% in untreated cells (Figure 4.2b). An increased frequency of glycine at the first position of N termini is also observed in apoptotic cells relative to untreated cells at both experimental scales (Figure 4.2a and b). The N termini uniquely identified in apoptotic Jurkat cells are thus consistent with induction of caspase-like activity.

Of the 3% P1 aspartate N termini detected in large-scale experiments with untreated cells (Figure 4.2b), 55% correspond to reported caspase substrates (Lu □ thi and Martin, 2007). Thus, it is likely that these originate from the small number of

apoptotic cells typically present in untreated cultures. The detection of 3% P1 aspartate N termini in large-scale experiments with untreated cells and less than 1% in small-scale experiments is consistent with the low abundance of such N termini in cultures of normal cells. Additionally, if one considers that N termini annotated in SwissProt are representative of native N termini in healthy cells, it is notable that less than 1% are derived from proteolytic processing after an aspartate residue (Mahrus 2008, Figure S2). In apoptotic samples, we find that the increased frequency of N termini located beyond the first 50 residues is solely attributable to P1 aspartate cleavages (Figures S1B and S1C). Thus, the vast majority of proteolysis we observe in apoptosis is attributable to caspases or proteases with caspase-like substrate specificity.

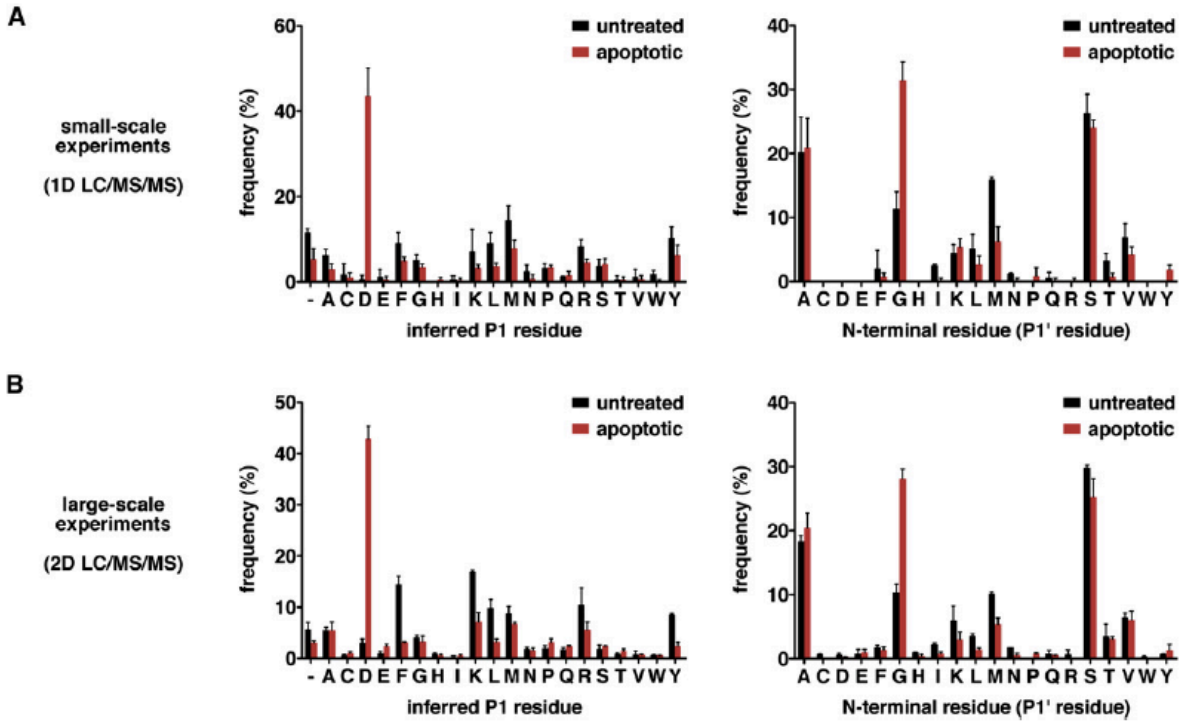


Figure 4.2 N termini derived from caspase-like cleavage are a hallmark of apoptotic cells
(a) Frequencies of P1 and P10 amino acid residues corresponding to nonhomologous N termini identified in small-scale 1D LC/MS/MS experiments with untreated and apoptotic Jurkat cells. Data are represented as mean ± SD (*n* = 2 for untreated, and *n* = 4 for apoptotic).

(b) Frequencies of P1 and P10 amino acid residues corresponding to nonhomologous N termini identified in large-scale 2D LC/MS/MS experiments with untreated and apoptotic Jurkat cells. Data are represented as mean ± SD (*n* = 2 for untreated, and *n* = 3 for apoptotic). ‘‘–’’ indicates lack of a putative P1 residue in cases where the P10 residue is an initiator methionine.

Among the total 1099 SY-labeled peptides identified in etoposide-treated Jurkat cells, 418 follow aspartate in corresponding protein sequences (Mahrus 2008, Tables S4 and S6). These peptides correspond to 333 P1 aspartate N termini and caspase-like cleavage sites (identified cleavage sites are listed in Mahrus 2008, Table S7). In turn, these cleavage sites map to 282 unique substrates and ten additional others that cannot be distinguished from homologs containing the same identified N terminus (identified substrates are listed in Mahrus 2008, Table S8). The average overlap between data sets obtained in separate experiments is 55% at the peptide level and 58% at the protein level (Mahrus 2008, Figures S3A and S3B). Similar overlap levels (~67%) have been previously observed for replicate analyses of complex mixtures of peptides with LC/MS/MS (Elias et al., 2005). We have verified 16 of the proteins identified as caspase substrates in our studies to be cleaved during apoptosis using immunoblotting (representative examples are included as Mahrus 2008, Figure S4A). We have also determined that the proteolysis of a representative set of substrates is blocked by the broad-spectrum caspase inhibitor Z-VAD(OMe)-fmk, consistent with this proteolysis being caspase-dependent (Mahrus 2008, Figure S4B).

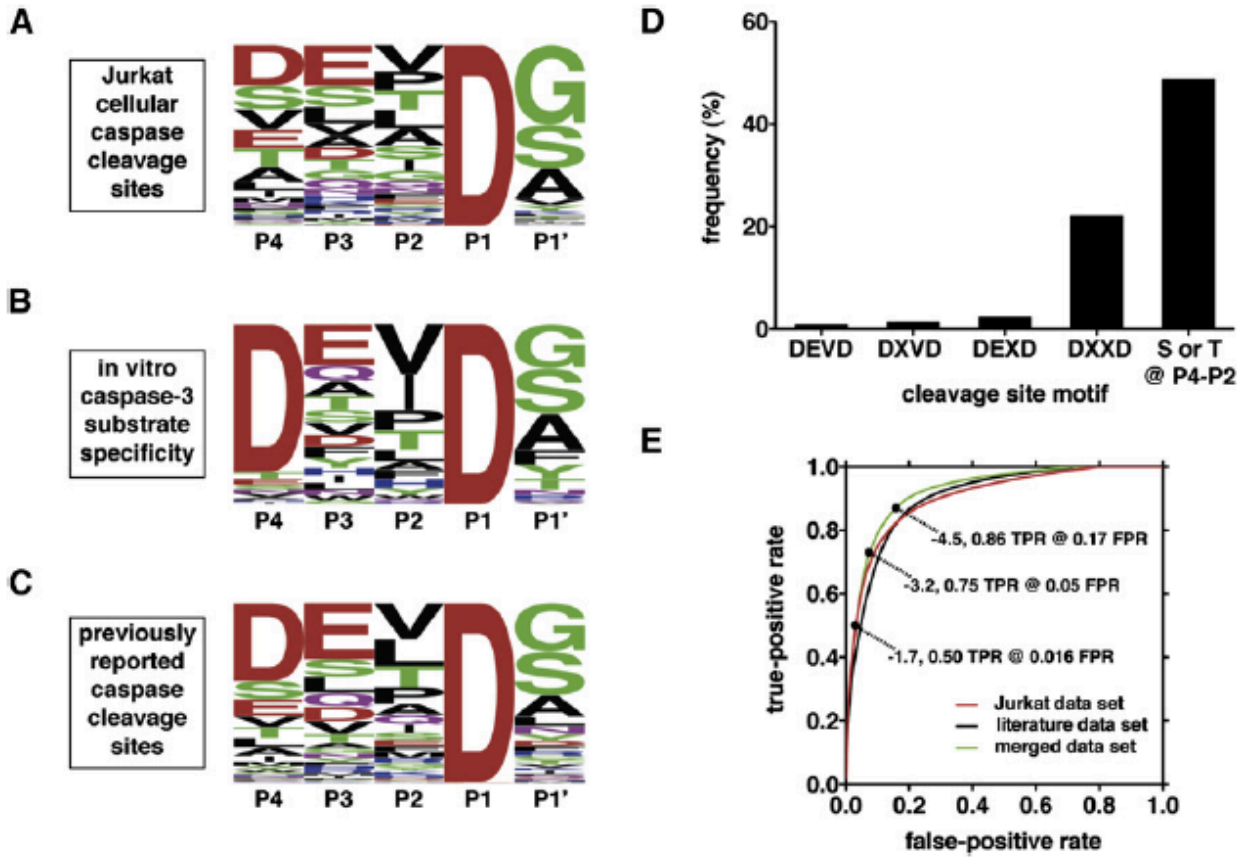


Figure 4.3 Substrate specificity of caspase-like cleavage induced in apoptotic cells
(a) Sequence logo representation (Crooks et al., 2004) of the frequency of amino acid residues in the capase-like cleavage sites identified in apoptotic cells.
(b) Sequence logo representation of the *in vitro* substrate specificity of caspase-3 (Stennicke et al., 2000; Thornberry et al., 1997).
(c) Sequence logo representation of the frequency of amino acid residues in previously reported capase cleavage sites.
(d) Frequency of P4-P1 motifs in the capase-like cleavage sites identified in apoptotic Jurkat cells.
(e) ROC curves for predictive HMMs constructed from three different cleavage site training sets (Jurkat, literature, and merged). Three representative HMM score threshold values for the merged data set are indicated (TPR, true-positive rate; FPR, false-positive rate).

The most frequent residues at the P4, P3, P2, and P10 positions of the caspase-like cleavage sites identified in apoptotic Jurkat cells are aspartate, glutamate, valine, and glycine, respectively (Figure 4.3a). Thus, an averaged composite of these cleavage sites indicates that the most common caspase activity in apoptotic cells exhibits a specificity that is most similar to the substrate specificity of executioner caspases-3 and -7, as determined with peptide substrates (Figure 4.3b) (Thornberry et al., 1997).

However, there are significant differences between the cellular cleavage sites and the *in vitro* specificity profiles. Notably, the canonical DEVD cleavage site motif is found in less than 1% of the caspase-like cleavage sites observed in apoptotic Jurkat cells, and the broader DXXD motif is still only found in 22% of the identified cleavage sites (Figure 4.3d). A distinct difference in the composite cellular profile is the high frequency of serine and threonine residues at P4, P3, and P2, which is not observed *in vitro* for any of the caspases (Mahrus 2008, Figure S5). Interestingly, a composite of all reported human capase cleavage sites and human orthologs of reported rodent caspase cleavage sites (Lu thi and Martin, 2007) is very similar to the Jurkat cellular profile reported here (Figure 4.3c).

These observations suggest that caspase substrate specificity determined with peptide substrates has limited value as a predictor of physiological caspase cleavage sites. To investigate the predictive value of a large set of known physiological caspase cleavage sites, we constructed three profile hidden Markov models (HMMs) using the cleavage sites identified in our studies, previously reported cleavage sites, and the union of these two data sets (a detailed description of this analysis is found in Mahrus 2008, Supplemental Experimental Procedures). The accuracy of these HMMs was estimated via jackknifing and plotted in a receiver operator characteristic (ROC) plot, showing the true-positive rate *versus* the false-positive rate at different HMM score thresholds. Although all three HMMs predict caspase cleavage sites relatively accurately, the HMM built from the merged substrate set performed slightly better than those built from the individual sets (Figure 4.3e). Its true-positive rate was 0.86 at the

false-positive rate of 0.15, compared to the average true-positive rate of 0.84 at the false-positive rate of 0.17 for the other two HMMs.

4.2.3. Analysis of structural determinants of caspase substrate specificity

The combined data set of the 333 caspase cleavage sites identified in our work and the approximately 300 previously identified caspase cleavage sites (Lu □ thi and Martin, 2007) allows an opportunity to expand our understanding of caspase substrate specificity from primary structure to the level of secondary and higher- order structures. To accomplish this goal, we mapped the known caspase cleavage sites onto experimentally determined atomic structures in the Protein Data Bank (PDB) (Berman et al., 2002), as well as comparative protein structure models in the ModBase database (Pieper et al., 2006). Stringent filters were applied so that only models likely to be sufficiently accurate for the analysis were used.

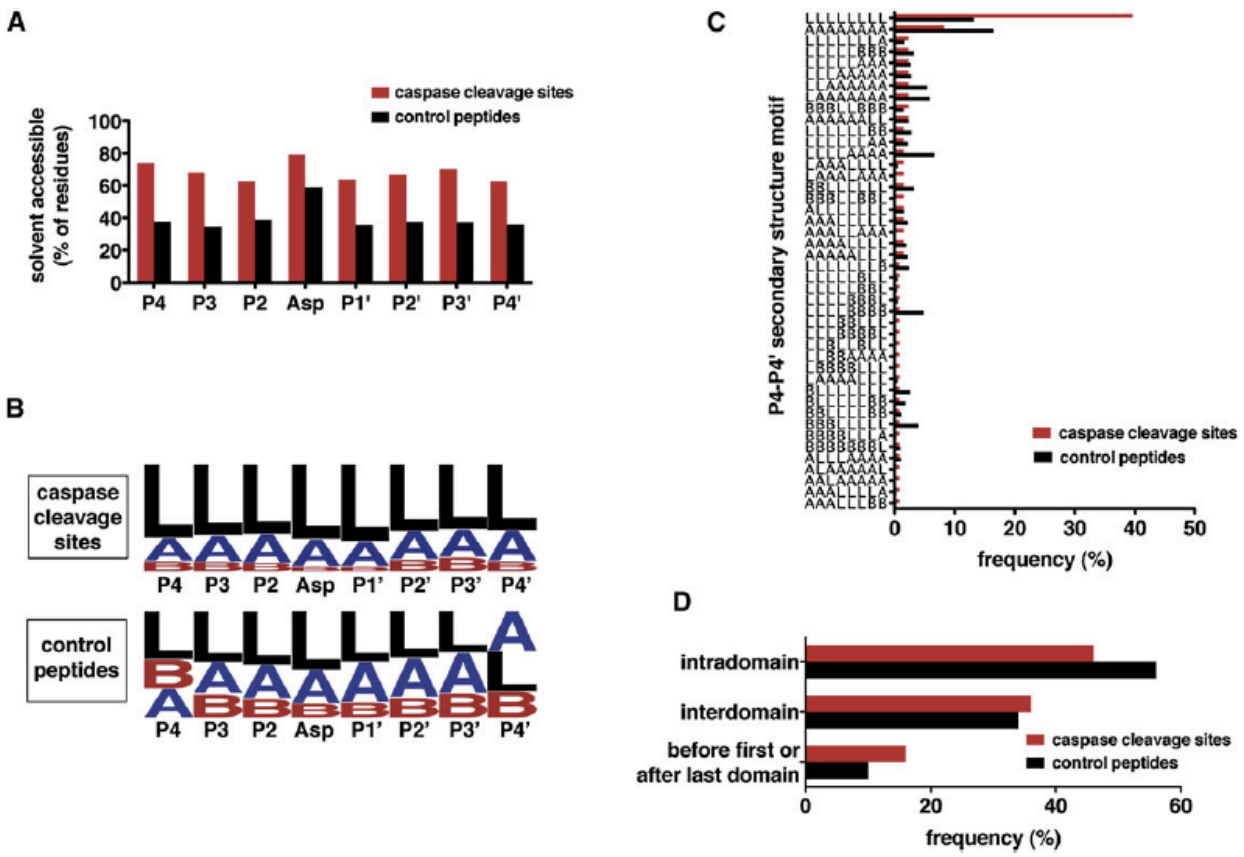


Figure 4.4 Structural Determinants of Caspase Substrate Specificity
(a) Solvent accessibility (>33% surface area exposed) at each position of all known P4–P40 positions of caspase cleavage sites and each position of all eight residue sequences containing aspartate in the fourth position found in PDB protein structures (control peptides). Differences between cleavage sites and control peptides have associated p values < 0.001 by χ-square test.
(b) Sequence logo representations of secondary structure at each position of all known P4–P40 positions of caspase cleavage sites and each position of the control peptides described above. L, loop; A, α-helix; and B, β-sheet. Differences between cleavage sites and control peptides have associated p values < 0.001 by χ-square test.
(c) Distribution of secondary structure motifs for P4–P40 caspase cleavage sites and for the control peptides described above.
(d) Localization of caspase cleavage sites in substrates relative to functional domain boundaries annotated in Pfam compared to localization of all eight residue sequences containing aspartate in the fourth position found in the human SwissProt database (control peptides). Differences between cleavage sites and control peptides have associated p values < 0.001 by χ-square test.

We identified 18 cleavage sites in known structures and 116 sites in comparative models. Depending on P4 through P40 position, between 60% to 80% of cleavage site residues are solvent accessible, as defined by solvent exposure of greater than 33% total surface area (Figure 4.4a). Averaged across P4 through P40, cleavage site

residues are 76% more exposed than a reference control of all octapeptide sequences in the PDB containing an aspartate residue at the fourth position. The type of secondary structure was assigned using DSSP (Kabsch and Sander, 1983) for P4 through P40 positions. The frequency of secondary structure types at each position reveals that caspases most frequently cleave protein substrates at loops relative to the octapeptide reference control described above (Figure 4.4b). Surprisingly, proteolysis at α-helical regions is not uncommon. Binning of cleavage sites into secondary structure motifs reveals that although an all-loop motif is the most common secondary structure motif, the second most common one is an all-helix motif (Figure 4.4c). The finding that some cleavages occur at solvent inaccessible and α-helical regions likely reflects structural dynamics of these regions. Structural examples of cleavages identified in our studies are included as supplemental data (Mahrus 2008, Figure S6).

Analysis of the location of cleavage sites in caspase substrates annotated in the Pfam database (Finn et al., 2006) indicates that 46% of them are located within an annotated functional domain, 38% are located between annotated domains, and 16% are located at protein termini, either before the first annotated domain or after the last (Figure 4.4d). This distribution is relatively similar to the distribution of a reference control of all octapeptide sequences in the human SwissProt database containing an aspartate residue at the fourth position. Thus, caspases do not exhibit a strong preference for cleavage of substrates either inside or outside functional domains. Caspase cleavage sites are also evenly distributed over the length of protein substrates (data not shown).

4.2.4. Analysis of protein-protein interactions between caspase substrates

Upon inspection of the entire data set of caspase substrates, we noted a number of instances where multiple proteins along a single biochemical pathway, or in a single protein complex, are targeted by caspases. For a more systematic analysis of this property, we utilized data from three different protein interaction databases (HPRD, IntAct, and MINT) to create a network of caspase substrate protein interactors (Chatr-aryamontri et al., 2007; Kerrien et al., 2007; Mishra et al., 2006). This network is made up solely of the substrates identified in our studies, reported human capase substrates, and human orthologs of reported rodent caspase substrates (Lu thi and Martin, 2007) but excludes the caspases themselves (binary interactions constituting this network are listed in Mahrus 2008, Table S9). A total of 415 interactors and 1253 interactions were found among the merged human caspase substrate data set of 602 proteins, for an average of 2.1 intra-data set interactions per caspase substrate. Ten data sets of 602 randomly chosen proteins from the protein interaction databases had an average of 0.2 intra-data set interactions per protein. This indicates a 10-fold enrichment in protein interactions between caspase substrates relative to randomly interacting proteins (Figure 4.5) (a detailed description of this analysis is found in Mahrus 2008, Supplemental Experimental Procedures).

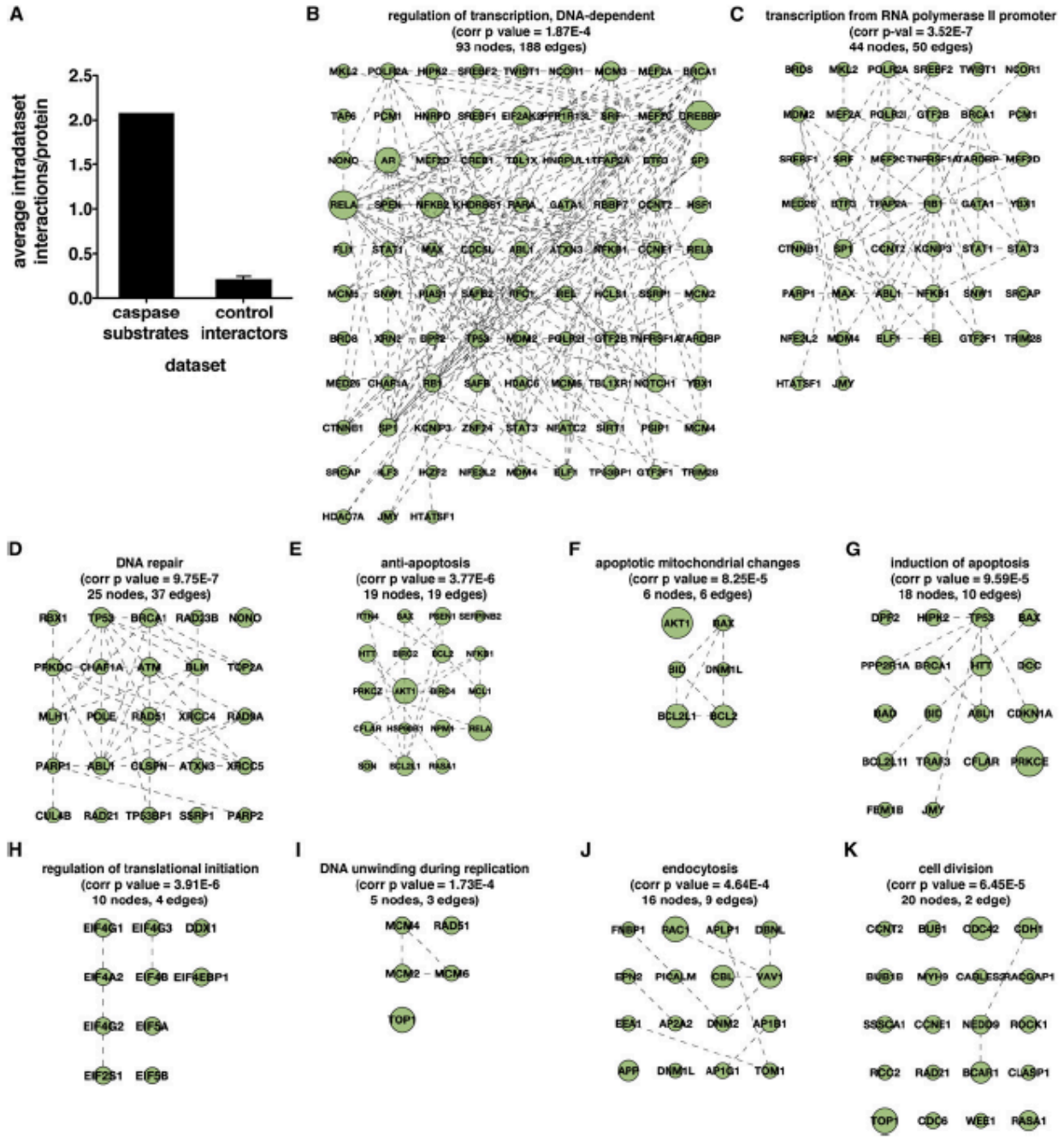


Figure 4.5 Network Analysis of Protein Interactions between Caspase Substrates.
(a) Enrichment in protein-protein interactions between the 602 total caspase substrates relative to an equally sized reference control set of protein interactors randomly selected from protein interaction databases. Data for the control set are represented as mean ± SD (n = 10).
(b–k) Caspase substrate protein interaction subnetworks encompassing substrates annotated to overrepresented GO biological process terms relative to the entire human GO annotation. Substrates are labeled with gene symbols. Corrected p values, number of nodes, and number of edges are indicated in each case. This analysis was applied to the substrates identified in this work and previously reported caspase substrates (Lu □ thi and Martin, 2007).

To determine which biological processes are preferentially targeted by caspases during apoptosis, we used the BiNGO (Maere et al., 2005) plugin of Cytoscape (Shannon et al., 2003) to find GO biological process terms that are overrepresented relative to the complete human GO annotation. We then focused on the 132 terms in the three deepest levels of the GO hierarchy to find the ten most overrepresented GO terms and the substrates annotated to those terms. This analysis yielded subnetworks of substrates involved in regulation of transcription, transcription from RNA polymerase II promoter, DNA repair, antiapoptosis, induction of apoptosis, apoptotic mitochondrial changes, regulation of translational initiation, DNA unwinding during replication, endocytosis, and cell division (Figure 4.5b-k). The regulation of transcription GO term yielded the densest subnetwork, with 188 edges among 93 nodes (Figure 4.5b). In sharp contrast to the other nine GO terms, the cell division GO term barely yielded a network at all, with only two edges among 20 nodes (Figure 4.5k).

4.2.5. The N-CoR/SMRT complex is a target of caspase proteolysis during apoptosis

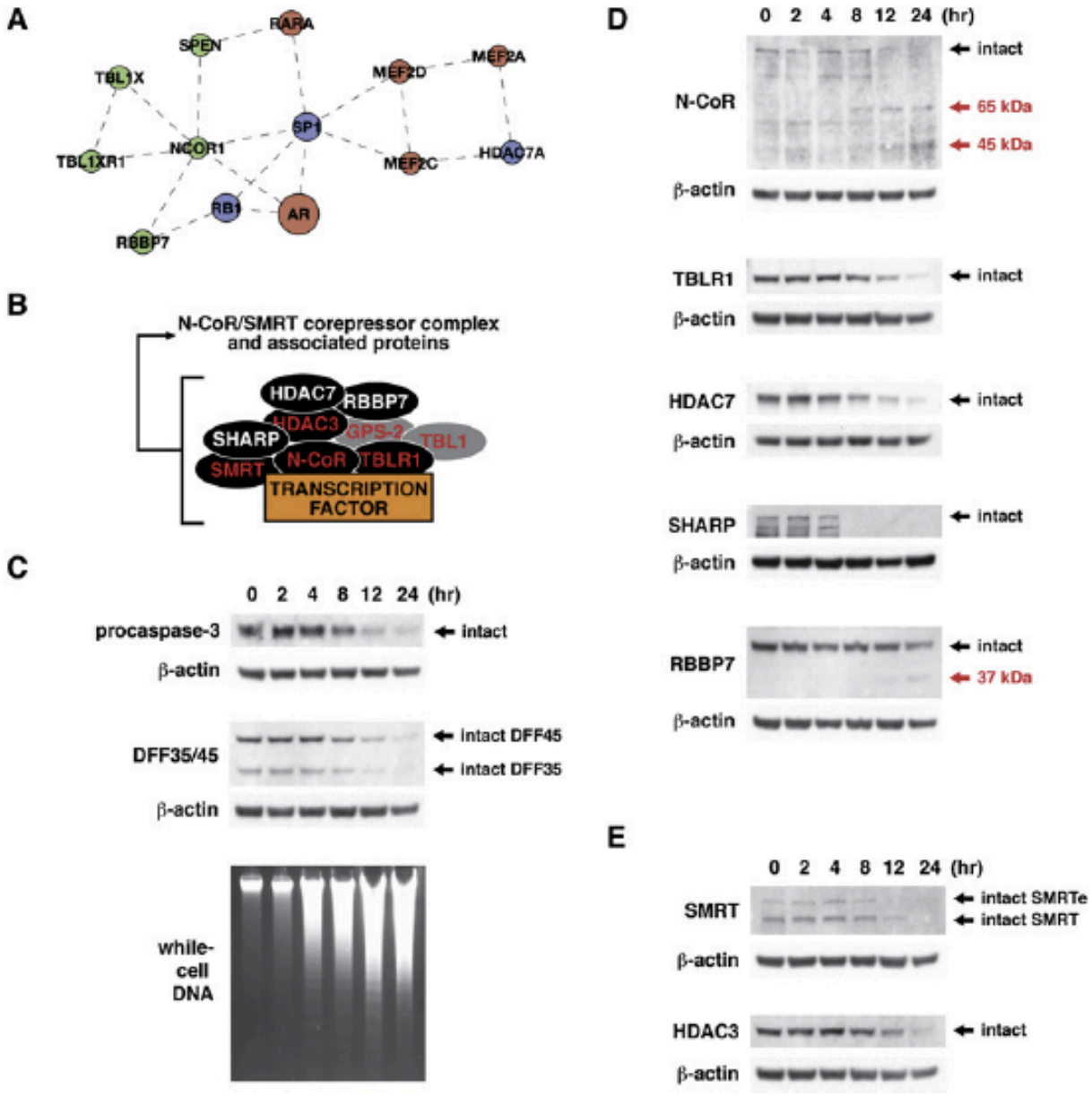


Figure 4.6 Analysis of Proteolysis of N-CoR/ SMRT Corepressor Complex Components during Apoptosis in Jurkat Cells after Treatment with 50 mM Etoposide
(a) Caspase substrate protein interaction subnetwork encompassing components of the N-CoR/ SMRT corepressor complex (N-CoR, SPEN, TBLR1, RBBP7, and HDAC7), and transcription factors such as retinoic acid receptor, androgen receptor, and SP1 (green, from this work; red, from literature; and blue, in both data sets).
(b) Schematic representation of N-CoR/SMRT corepressor complex resident components and visiting interactors (red label, resident component; white label, visiting interactor; and black fill, target of proteolysis in apoptosis).

(c) Time courses for the proteolysis of procaspase-3 and DFF35/45 and for oligonucleosomal DNA fragmentation.
(d) Full cleavage of N-CoR, HDAC7, SHARP, and TBLR1, and partial cleavage of RBBP7.
(e) Full cleavage of SMRT and of HDAC-3, a previously identified caspase substrate. Black arrows indicate full-length proteins. Red arrows indicate expected cleavage products for cleavage at the sites identified in our studies (cleavage products were not detected in all cases).

To analyze whether multiple cleavages along a pathway or in a complex occur at physiologically relevant rates, we focused on the portion of the regulation of transcription subnetwork representing N-CoR/SMRT transcriptional corepressor complex components and interactors (Figure 4.6a and b). This complex is involved in the recruitment of histone deacetylase activity to chromatin, which leads to chromatin condensation and transcriptional repression. Our studies identified N-CoR/SMRT complex resident components N-CoR and TBLR1 (Karagianni and Wong, 2007), as well as additional N-CoR/SMRT complex interactors HDAC7 (Fischle et al., 2001), MINT/SHARP/SPEN (Shi et al., 2001), and RBBP7/RbAp46 (Takezawa et al., 2007) as caspase substrates (MS/MS spectra of N-terminal peptides corresponding to cleavage sites in these proteins are included in Mahrus 2008, Figures S7–S14). We probed for cleavage of these proteins during etoposide-induced apoptosis in Jurkat cells by immunoblot in order to qualitatively determine extent of proteolysis in each case. N-CoR, TBLR1, HDAC7, and SHARP were all fully cleaved at rates similar to those observed for hallmark substrates procaspase-3 and DFF45 (Figure 4.6c and d). This proteolysis also tracked reasonably well with the time course for DNA fragmentation. In contrast, only partial proteolysis of RBBP7 was observed, suggesting it to be a possible bystander substrate (Figure 4.6d). Although not detected in our proteomic studies, we predicted the N-CoR homolog SMRT (Karagianni and Wong, 2007) to also be a caspase substrate on the basis of high sequence similarity to N-CoR cleavage sites. Indeed, SMRT was fully cleaved during etoposide-induced apoptosis in Jurkat cells

(Figure 4.6e). The previously identified caspase substrate HDAC3 (Escaffit et al., 2007), another N-CoR/SMRT complex component (Karagianni and Wong, 2007), was also fully cleaved. Organization of functional domains in these proteins indicates that proteolytic processing at the cleavage sites identified in our studies likely results in inactivation of protein function by virtue of separating functional domains from one another (Figure 4.7).

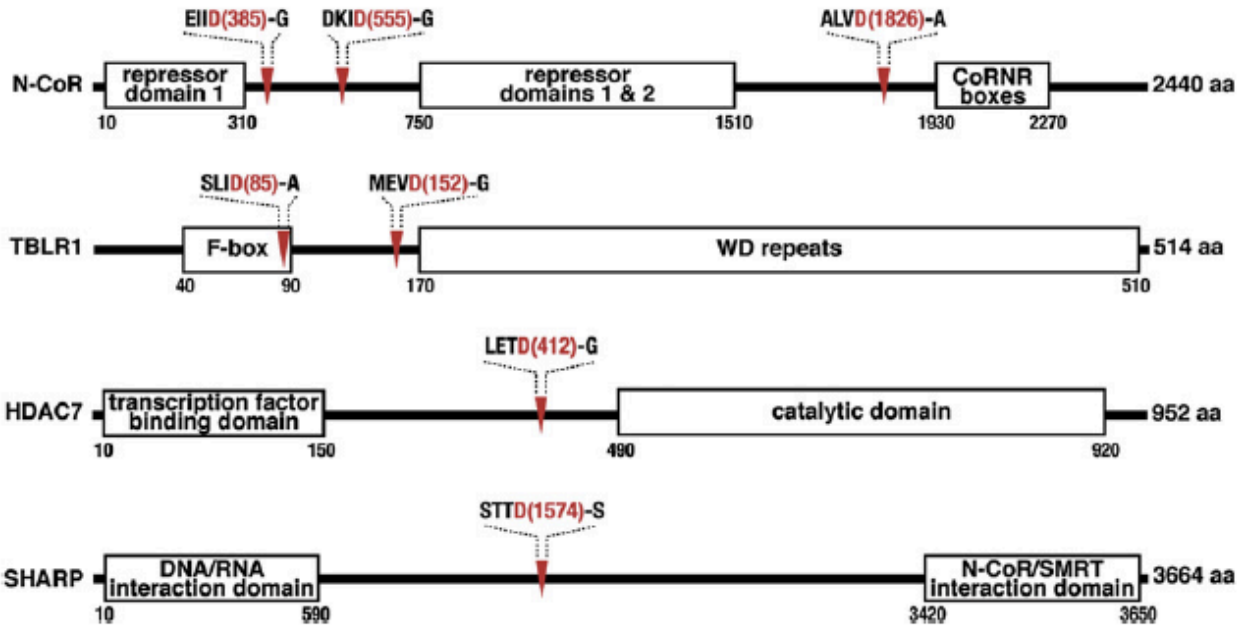


Figure 4.7 Caspases cleave resident N-COR/SMRT complex components and visiting interactors at regions leading to separation of functional domains
Functional domain organization and candidate caspase cleavage sites in SMRT and TBL1 are similar to those indicated for the respective homologs, N-CoR and TBLR1.

4.3. Discussion – The role of caspase cleavage in apoptosis

4.3.1. Caspases target specific protein hubs in certain biological pathways

One of the most striking findings of this study is that caspase substrates as a whole tend to physically interact with one or more other caspase substrates, either in protein complexes or networks. We interpret this as an indication that caspases target a limited

set of biological pathways to elicit programmed cell death, as opposed to indiscriminately targeting the entire cellular proteome. These data also suggest that caspases target protein complexes that are hubs for cell viability in essential processes such as transcription and that targeting of multiple components in each complex is required for a full commitment to apoptosis. In this regard, it is notable that active caspases are dimeric, which is rare for proteases. A dimer is well equipped for semi-processive activity consistent with targeting multiple components of protein complexes. Another reported example of targeted proteolysis of a protein complex is the cleavage of SET, HMG-2, and Ape1, three components of the SET complex, by the cytotoxic lymphocyte protease granzyme A (Lieberman and Fan, 2003). Interestingly, granzyme A is also dimeric.

4.3.2. Novel caspase substrates lead to hypotheses of apoptotic mechanisms

The discovery that several components of the N-CoR/SMRT transcriptional co-repressor complex are targets of caspase proteolysis presents a remarkable example of multiple cleavages in a single protein complex or pathway during apoptosis. Six proteins that are part of, or interact with, the N-CoR/SMRT complex are fully cleaved during etoposide-induced apoptosis in Jurkat cells, including the corepressors N-CoR and SMRT themselves. This finding was made possible by our large-scale discovery-oriented proteomic approach, as opposed to a more typical focused hypothesis-driven approach. Inactivation of the N-CoR/SMRT complex during apoptosis may achieve a result similar to the effect of HDAC inhibitors, with decreased histone deacetylation leading to opening of chromatin and transcriptional upregulation of proapoptotic genes (Bolden et al., 2006). Interestingly, HDAC 7 has recently been implicated as a

physiological substrate of caspase-8, with its proteolytic inactivation leading to upregulation of Nur77 (Scott et al., 2008).

4.3.3. Proteolytic products alter substrate functions

Our studies indicate that a change in function of proteins targeted by caspases during apoptosis must be rationalized by one or occasionally a few cuts per protein. We have found that caspase cleavages occur inside functional domains and between functional domains at approximately equal frequencies. In either case, relatively stable products must be produced after cleavage of the substrates since we detected them. Stability of these products is also consistent with the relatively strict P10 glycine, serine, and alanine specificity we observe for the cellular caspase-like activity, which creates fragments conforming to the N-end rule (Varshavsky, 1992). In addition to functional disruption of the substrate protein, such cleavages may result in products that function as dominant negatives. For example, in the case of the N-CoR and SMRT corepressors, the C-terminal cleavage products contain the CoRNR boxes known to interact with nuclear receptors (Hu and Lazar, 1999). These proteolysis products could thus inhibit interaction between N-CoR/SMRT and nuclear receptors.

4.3.4. Proteomic results represent the union of all caspase cleavage events in whole-cells

By globally identifying caspase-like cleavage sites in the proteome of apoptotic cells, this work presents a large-scale substrate specificity profile of caspase processing of endogenous proteins in intact cells. Importantly, this profile is influenced not only by the primary structure of cleavage sites but also by solvent accessibility, secondary and higher order protein structure, and possibly posttranslational modifications of substrates

(To □ zse´ r et al., 2003). Our finding that caspases often target proteins in complexes underscores the value of studying determinants of proteolysis under physiologically relevant conditions. The caspase-like cleavage sites identified in apoptotic Jurkat cells likely result from the action of several members of the caspase family of proteases. Although the aggregate substrate specificity of the observed caspase-like activity is most similar to the known specificity of executioner caspases, *in vitro* studies of caspases using peptides do not fully account for the observed cellular specificity (Schilling and Overall, 2008; Stennicke et al., 2000; Thornberry et al., 1997). Peptide-centric approaches are best suited for determination of optimal protease substrate sequence specificity, invaluable in development of sensitive synthetic substrates or potent inhibitors. In contrast, a protein-centric method such as the one presented here is best suited for characterization of endogenous proteolysis in biological samples and for studying structural context of peptide cleavage site on the native protein.

4.3.5. Proteomic results are input for further bioinformatics analysis

This work indicates that the widely used primary structural determinants of caspase *in vitro* substrate specificity are insufficient to predict physiological caspase cleavage sites. However, the cellular cleavage sites we have identified significantly expand a data set that can be used to train algorithms for predicting cleavage sites. Indeed, a proof of principle is provided by an accurate prediction of caspase cleavage sites by our preliminary HMMs. In addition to demonstrating that caspase cleavage sites are most commonly found in solvent accessible loop regions, as shown for other proteases (Hubbard et al., 1991), our analysis also indicates that a number of cleavage sites appear in partially solvent inaccessible regions and α-helices. This information

could also be incorporated into predictive algorithms (for example, as discussed in Chapter 2). Finally, based on our protein interaction analysis, predictive algorithms may also benefit from scoring that considers physical interactions of candidate substrates with other caspase substrates.

4.3.6. Experimental results represent a subset of all caspase cleavage sites

The incomplete overlap between cleavage sites and protein substrates identified in our separate experiments is not uncommon for tandem mass spectrometric analysis of complex mixtures, in which analysis of many species, whether peptidic or not, precludes complete sampling (Elias et al., 2005). The number of caspase substrates we have identified is thus likely smaller than the total number of caspase substrates in apoptotic Jurkat cells. We identified 50 of approximately 361 previously reported human caspase substrates and 50 of approximately 307 previously reported human caspase cleavage sites (Figures S3C–S3E) (Lu □thi and Martin, 2007). Incomplete proteomic sampling of caspase substrates in our studies is likely an important contributor to the modest overlap between the substrates we have identified and those previously reported. This result furthermore demonstrates a role for accurate computational methods to compliment experimental findings through predicting new substrate cleavage sites, as discussed in Chapter 2.

4.3.7. Proteomic results significantly expands understanding of caspase substrate specificity

Although the data set of substrates we have identified is not comprehensive, it doubles the number of known cleavage sites in human targets of caspase-like proteolysis in apoptosis. The study of apoptotic pathways has important ramifications for identification

of pathways that are critical for cellular homeostasis, and for development of potential anticancer therapeutics. A number of caspase targets are active or established drug targets for treating cancer, including topoisomerase II, Bcl-2, Hdm2, MEK1, and Akt, to name a few. Thus, it is possible that the list of substrates we have identified includes new candidate chemotherapeutic targets. The products of caspase proteolysis may also serve as useful biomarkers for assessment of chemotherapeutic efficacy, as demonstrated in the case of cytokeratin-18 for breast cancer (Olofsson et al., 2007). Along with MS-based quantitation, the technology we describe should enable global analysis of the apoptotic phenotype as a function of time, cellular context, and type of induction. Finally, the technology should also be broadly applicable for global sequencing of proteolytic cleavage sites in other biological settings.

4.4. Methods in profiling of caspase cleavage sites

4.4.1. Preparation of subtiligase and peptide ester substrate

Subtiligase was recombinantly expressed in *B. subtilis* and purified essentially as previously described (Abrahmse´ n et al., 1991). The biotinylated peptide glycolate ester was synthesized by solid-phase peptide synthesis as described for other subtiligase substrates (Braisted et al., 1997).

4.4.2. Cell culture, induction of apoptosis, and cell lysate preparation

Jurkat clone E6-1 (ATCC) cells at a density of 1 X 10^6 cells/ml were treated with etoposide (50 μM) for 0 or 12 hr prior to being harvested. Detergent lysates were prepared at a typical concentration of 2 X 10^8 cells/ml (approximately 20 mg/ml) with buffered 1.0% Triton X-100 in the presence of protease inhibitors.

4.4.3. Sample Biotinylation, Denaturation, Reduction, Alkylation, and Gel Filtration

Cell lysates were biotinylated by treatment with subtiligase (1 μM), biotinylated peptide ester substrate (1 mM), and DTT (2 mM). Ligation reactions were typically left to proceed at room temperature for 60 min. Samples were then denatured, reduced, alkylated, and subjected to gel filtration for removal of hydrolyzed peptide ester substrate.

4.4.4. Trypsinization and Recovery of Biotinylated Peptides

Filtered samples were subjected to solution digestion with sequencing grade modified trypsin (Promega). Biotinylated N-terminal peptides were captured from trypsinized samples with NeutrAvidin agarose (Pierce). Captured peptides were recovered by treatment of agarose resin with recombinant TEV protease (1 μM).

4.4.5. LC/MS/MS

N-terminal peptide samples were analyzed by one-dimensional reversed-phase LC/MS/MS or two-dimensional strong cation exchange/reversed-phase LC/MS/MS. In the latter case, samples were fractionated by offline strong cation exchange chromatography with a 60 min gradient on a 2.1 X 200 mm PolySULFOETHYL Aspartamide column at a flow rate of 0.3 ml/min. Reversed-phase chromatography of unfractionated or fractionated samples was carried out with a 60 min gradient on a 75 μm X 15 cm C_{18} column at a flow rate of 350 nl/min. The capillary column was coupled to a QSTAR Pulsar, QSTAR XL, or QSTAR Elite mass spectrometer (Applied Biosystems). For each acquired MS spectrum, either the single or the two most intense multiply charged peaks were selected for generation of CID spectra. A dynamic

exclusion window of 3 min was applied. CID spectra not included as supplemental data will be made available upon request.

4.4.6. Interpretation of MS/MS Spectra

Data were analyzed with Analyst QS software, and MS/MS centroid peak lists were generated with the Mascot.dll script. Data were searched against the SwissProt human database (March 2008 release) with Protein Prospector 5.0 (University of California, San Francisco). Peptide tolerances in MS and MS/MS modes were 100 ppm and 300 ppm, respectively. The digest protease specified was trypsin, allowing for two missed cleavages and nonspecific cleavage at N termini. An N-terminal SY modification and cysteine carbamidomethylation were specified as a fixed modifications, and methionine oxidation was specified as a variable modification. Peptides with scores ≥ 22 and expectation values % 0.05 were considered positively identified. False-discovery rates for peptide identifications were estimated with a target-decoy strategy.

4.4.7. Cleavage site predictions

Cleavage site prediction was assessed using 1000 jackknife trials on 473 substrates containing 603 cleavage sequences from both our caspase substrate dataset and the literature substrate dataset (Lüthi and Martin, 2007). A test set consisted of 60 randomly selected true positive cleavage sequences and 3,000 randomly selected true negative peptides derived from all octapeptides in caspase substrates with aspartate at the fourth position that have not been shown to be cleaved by caspases. A training set consisted of all cleavage sequences from respective substrate sets not present in the corresponding test set. Hidden Markov models were constructed using the "hmmbuild" command of HMMer version 2.3.2 (Eddy, 1998) and test peptides were scored using

the "hmmpfam" command.

4.4.8. Structural bioinformatics

Secondary structure analysis of cleavage sites was carried out on a set of experimentally determined structures from the Protein Data Bank (Berman et al., 2002) and “good quality” comparative models from ModBase (Pieper et al., 2006). A good quality model has either a N-DOPE score < -0.4 (Shen and Sali, 2006) or is based on ≥ 25% sequence identity to the template structure with a “model score” > 0.8 (Melo and Sali, 2007). Such models are likely to have ≥ 75% of their Cα atoms within 3.5 Å of the correct positions (Eswar and Sali, 2007). The DSSP algorithm was used to assign the type of secondary structure of each cleavage site, discriminating between α-helix, β-sheet, and loop states. The fraction of solvent accessible surface area of each residue in the cleavage sites was determined by dividing the observed exposed surface area, as also assessed by DSSP, by the maximum exposed surface area of the residue (Rose et al, 1984). Residues were considered exposed if this fraction was > 0.33. A reference control distribution of both solvent accessibility and secondary structure state was determined from the set of all octapeptides with aspartate at the fourth position from 15,787 experimentally determined structures with < 95% sequence identity to each other (Berman et al., 2002). Domain analysis was performed using domain assignments from the Pfam database (July 2007 release) (Finn et al., 2006). The reference control for this analysis was the set of all octapeptides with aspartate at the fourth found in the human Swiss-Prot database. Statistical significance of differences between caspase cleavage sites and reference controls were assessed using the χ-square test. Molecular graphics were rendered using Pymol 1.0 (DeLano Scientific).

4.4.9. DNA Fragmentation

Fragmentation of whole cell DNA was analyzed by agarose gel electrophoresis with the Apoptotic DNA Ladder Kit (Roche).

4.4.10. Immunoblotting

Jurkat cells at a density of 1 X 10^6 cells/ml were treated with etoposide (50 μM) for 0, 2, 4, 8, 12, and 24 hr prior to being harvested. Whole-cell lysates were prepared at a concentration of 2 X 10^7 cells/ml with buffered 1.0% SDS in the presence of protease inhibitors and sonication. Lysates were normalized to a protein concentration of approximately 2 mg/ml prior to analysis by SDS-PAGE and western blot. Utilized antibodies are listed in Mahrus 2008, Supplemental Experimental Procedures.

4.5. Introduction – The role of O-GlcNAcylation in the cell

O-GlcNAcylation, the addition of a single sugar (β-N-acetylglucosamine) to serine and threonine residues on specific peptides regions on intracellular domains of proteins, is a reversible and dynamic post-translational modification (PTM). The O-GlcNAcylation state of proteins is responsive to numerous cellular stimuli, including nutrient levels and stress. The addition of this PTM is catalyzed by the enzyme O-GlcNAc-transferase (OGT). This enzyme is highly expressed in the brain, and the physiological roles of protein GlcNAcylation may be particularly important in the central nervous system (Cole and Hart, 2001) (Gao et al., 2001). OGT is present in dendrites, axon terminals and is associated with microtubules (Akimoto et al., 2003). Neuron-specific deletion of OGT results in neonatal lethality due in part to abnormal neuronal development and motor deficits (O'Donnell et al., 2004).

Because O-GlcNAcylation modifies serine and threonine side chains, there is the potential for interaction between the functions of this moiety and those of phosphorylation; we denote this interaction “cross-talk”. Over 1,000 proteins have been identified as O-GlcNAc modified. While the majority of these are also phosphorylated (Copeland et al., 2008), the implications are unclear given that the majority of all cellular proteins are probably phosphorylated. In addition, in most cases the specific peptide bearing the O-GlcNAc modification within a protein is still unknown. Traditional biochemical analysis has revealed numerous proteins that have been shown to be both phosphorylated and GlcNAcylated including c-Myc (Chou et al., 1995), nitric oxide synthase (Du et al., 2001), RNA polymerase II (Kelly et al., 1993)(Comer and Hart, 2001) synapsin I (Cole and Hart, 1999), tau (Liu et al., 2004a) and amyloid precursor protein (Griffith et al., 1995). In cell culture, modulation of the global levels of phosphorylation is accompanied by changes in GlcNAcylation levels of many proteins, and vice versa (Wang et al., 2008)(Griffith and Schmitz, 1999), although the specific sites involved have not been reported. Obviously these responses are complex. For example, pharmacological inhibition of a kinase causes an increase in GlcNAcylation of some proteins and a decrease in others (Wang et al., 2007). Postulation of a clear mechanistic basis for interpretation of these types of experiments is lacking.

Driven by advances in affinity chromatography and the development of several generations of more powerful tandem mass spectrometers (Choudhary and Mann, 2010)(Thingholm et al., 2009), our knowledge of the complexity and extent of cellular phosphorylation is still growing dramatically. In contrast, analogous progress in our

knowledge of O-GlcNAcylation has lagged due to less robust enrichment methodologies and suitable, broadly applicable and sensitive mass spectrometric methodologies.

In this present work, we have established a workflow that permits the combined detection and determination of O-GlcNAcylation and phosphorylation sites from proteins in the same biological sample. This study has resulted in the identification of over 6,000 proteins, including some 1,750 sites of O-GlcNAcylation and 16,500 sites of phosphorylation. These findings correspond to some 15% and 60% of proteins being O-GlcNAcylated and phosphorylated, respectively. In addition, these results demonstrate that cross-talk between the two types of PTMs does occur at the catalytic level but is less prevalent at the structural level.

4.6. Results – Characterization and analysis of O-GlcNAcylation modifications

4.6.1. Abundance of O-GlcNAcylation and phosphorylation is quantified

We have developed a workflow to sequentially enrich O-GlcNAcylated and phosphorylated peptides from tryptic digests of mouse synaptosomes, which also allows for analysis of the protein content from the PTM-depleted sample (Figure 4.8a). O-GlcNAcylated peptides were isolated using three rounds of lectin weak affinity chromatography (Figure 4.8b), yielding a final pool containing approximately 30% GlcNAcylated peptides. Phosphorylated peptides were isolated using an automated TiO2-based enrichment step (Trinidad 2011, Figure S1A). These two PTM-enriched fractions as well as the final unbound fraction (containing non-modified peptides) were then fractionated using high pH reverse phase chromatography (Figure 4.8c). All fractions were analyzed on an LTQ-Orbitrap Velos mass spectrometer using electron transfer dissociation (ETD) for O-GlcNAc peptides and collisional dissociation (CAD or

HCD) for phosphopeptides and others. Interpretation of these mass spectral analyses resulted in the identification of 2,278 unique O-GlcNAcylated and 18,173 phosphorylated peptides. These assignments correspond to over 1,750 unique sites of O-GlcNAcylation and 16,500 unique sites of phosphorylation. Analysis of the PTM-depleted digest identified 52,208 unique peptides from 6,287 proteins, all at global FDRs of less than 1% (Trinidad 2011, Tables S1-3). As we have previously reported, our enrichment technique using the lectin wheat germ agglutinin (WGA) also enriches for N-GlcNAcylated peptides (Chalkley et al., 2009b), and in our current analysis, we found over 450 N-GlcNAcylated peptides (Trinidad 2011, Table S4). While WGA has been reported to be selective for GlcNAcylated peptides and proteins (Nagata and Burger, 1974), we have identified over 150 peptides in the WGA-enriched fractions that appear to be GalNAcylated (Trindidad 2011, Table S5).

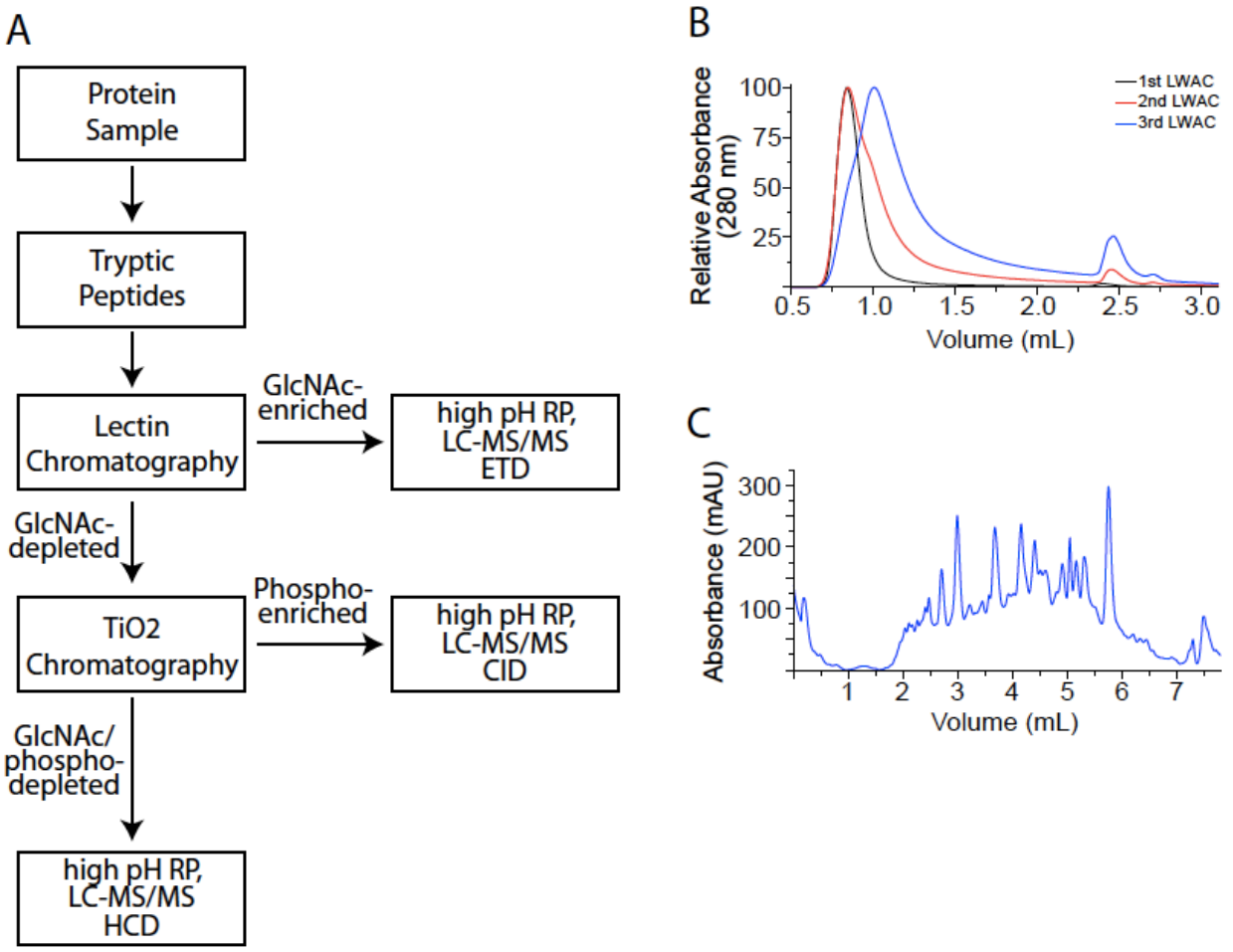


Figure 4.8 Mass spec workflow and primary data.
a) Workflow schematic for the serial analysis of GlcNAcylation, phosphorylation and protein content (see methods).
b) Peptide UV trace during three sequential runs of lectin chromatography. The UV shift to later elution times corresponds with a subsequent enrichment in the percentage or peptides that are GlcNAc-modified.
c) rUV trace of the high pH reverse phase gradient for the final GlcNAc-enriched fraction. Similar gradients are used to fractionate all peptides prior to analysis by LC-MS/MS.

4.6.2. PTM-detection efficiencies allows for estimation of total cellular PTM counts

A major factor affecting whether or not a given peptide is detected in a proteomic study is its relative abundance (Liu et al., 2004b). To estimate how efficiently we identified sites of O-GlcNAcylation and phosphorylation within our synaptosome preparation, we took advantage of the fact that we also conducted an in-depth protein analysis. The 6,287 proteins that we identified were divided into bins based upon their relative

abundance as determined by calculating exponentially modified protein abundance index (emPAI) values for each protein (Shinoda et al., 2010). We then calculated the percentage of proteins in each bin that were either GlcNAcylated or phosphorylated. For the most abundant proteins, we identified 19% and 63% of them to be GlcNAcylated and phosphorylated, respectively (Figure 4.9a and b). Proteins present at lower abundance were substantially less likely to be identified as GlcNAcylated (an average of 9.8% for the 12 lowest bins). For phosphorylation, this decrease was more modest. For 52% of the proteins in the 12 lowest bins, at least one site of phosphorylation was identified. Proteins in the most abundant bin had an average of 0.51 and 5.9 sites of GlcNAcylation and phosphorylation, respectively (Figure 4.9c and d). The average number of sites identified per protein dropped off significantly with decreased protein abundance for both PTMs. Overall, this suggests that while we were able to identify large numbers of both PTMs, we were not identifying all PTM-modified peptides present in the sample, particularly those originating from lower abundance proteins. Based upon the average modifications per protein for the most abundant/thoroughly characterized proteins, we now can postulate the existence over 3,400 O-GlcNAcylation sites and 39,000 phosphorylation sites for the more than 6,000 proteins identified in our synaptosome preparation. Using the same rationale, we estimated that we identified approximately 50% and 33% of the GlcNAcylation and phosphorylation sites in our sample, respectively. This result is another example of even large proteomic datasets finding only a subset of all modified peptides, as discussed in section 2.1.

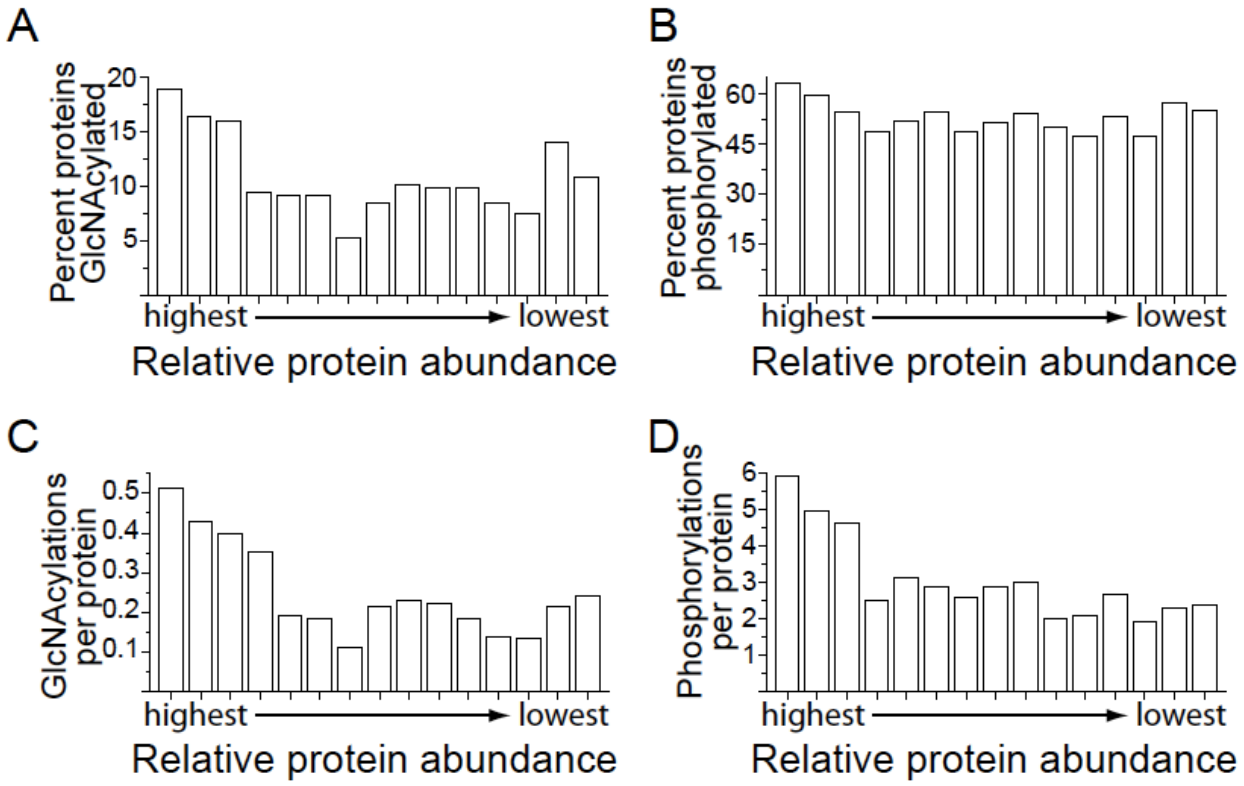


Figure 4.9 Percentage of proteins with given PTMs as a function of relative protein abundance. a-b) Overall abundance for GlcNAcylation and phosphorylation. The total number of unique non-phosphorylated and non-GlcNAcylated peptides per protein was used as an estimate of protein abundance. The ability to detect that a protein is PTM-modified increased with protein abundance. 38% and 88% of the highest abundance proteins were GlcNAcylated and phosphorylated, respectively. c-d) Average number of modification sites as a function of relative protein abundance. The number of identified PTM-sites per protein increased with protein abundance. An average of 1.5 and 17 sites of GlcNAcylation and phosphorylation per protein was observed for the most abundant proteins. Note: because the proteins bassoon and piccolo were GlcNAcylated to a much higher extent than other proteins, they were not used. Including them would raise the average number of GlcNAcylations per protein from 1.5 to 3.1.

4.6.3. Mass spectrometry allows for characterization of PTM-modified peptides

Multiple sites of GlcNAcylation were often found close together in protein primary sequence, or in close proximity to sites of phosphorylation. Figure 4.10a and b show MS/MS spectrum of two different O-GlcNAc site isomers of the peptide sequence, TAVKPTPIILTDQGMDLTSLAVEAR, from the protein bassoon. Figure 4.10c and d show MS/MS spectra of two PTM-analogs of the peptide AAVVTSPPPTTAPHK from the protein α-adducin, where the peptide is either phosphorylated or GlcNAcylated at

serine-6. Overall, we observed 137 instances when the phosphorylated peptide and the GlcNAcylated analog occurred on the same amino acid. We observed 439 instances of peptides containing two sites of GlcNAcylation. An example of one such peptide, SVTDTALPGQSSGPFYSPR, modified at serine-1 and threonine-3, is shown in Trinidad 2011, Figures S1. Trinidad 2011, Figure S2 shows an example of an N-GlcNAcylated peptide with the sequence LNGTDPIVAADSKR from the Prolow-density lipoprotein receptor-related protein 1, modified at aspargine-2.

4.6.4. PTM sequence motifs are degenerate

Previous analyses, based on a significantly smaller scale GlcNAcylated peptide dataset suggested a P-V-X-S/T motif for substrates of OGT (Vosseller et al., 2006). While this motif does exist for a subset of modified peptides in this study, the majority of GlcNAcylation sites assigned here fit poorly to this motif. In fact, less than 20% of the modified peptides we observe here can be explained using this motif. Using the present findings, Figure 4.10e shows a sequence logo representation of the amino acids surrounding the modified serine/threonine. There is a moderate preference for a proline residue either two or three amino acids N-terminal to the site of modification (-2 or -3). There is also a slight preference for valine at the -1 and -3 positions. Overall, GlcNAc appears to be targeted towards regions rich in serine/threonine residues, as evidenced by an increased frequency of these residues within five residues of modification sites. Such a preference for serine/threonine rich stretches may explain our detection of over 439 peptides with multiple GlcNAc modifications. This observation suggests a recognition mechanism in which the OGT targets a general linear motif on a protein without a strong consensus for the exact peptide sequence.

To investigate motifs within our phosphorylation dataset, we used Motif-X to look for over-represented patterns (Schwartz and Gygi, 2005). We find that a total of 56 motifs show statistically significant overrepresentation (Trinidad 2011, Table S5). To look more generally at potential motif characteristics, we grouped amino acids by chemical property (e.g. small hydrophobic, charged/polar side chains) as shown in Figure 4.3f – i. When grouped by chemical property, the most prevalent amino acids present around the site of GlcNAcylation are small/non-polar residues, indicating existence of a hydrophobic residue at the -3 position. Phosphorylation has a similar preference for small/non-polar residues. In addition, due to the prevalence of proline-directed kinases in the mammalian kinome, there was an increased probability of having a hydrophobic residue at the +1 position. Finally, we examined those serine/threonine residues showing reciprocal modification by both PTMs. This subset had a motif most similar to that of the overall GlcNAcylation motif. We compared these motifs to the population of serine/threonine residues not found to be PTM-modified. Hydrophobic residues are most prevalent at all amino acids immediately surrounding these serine/threonine residues Figure 4.3i.

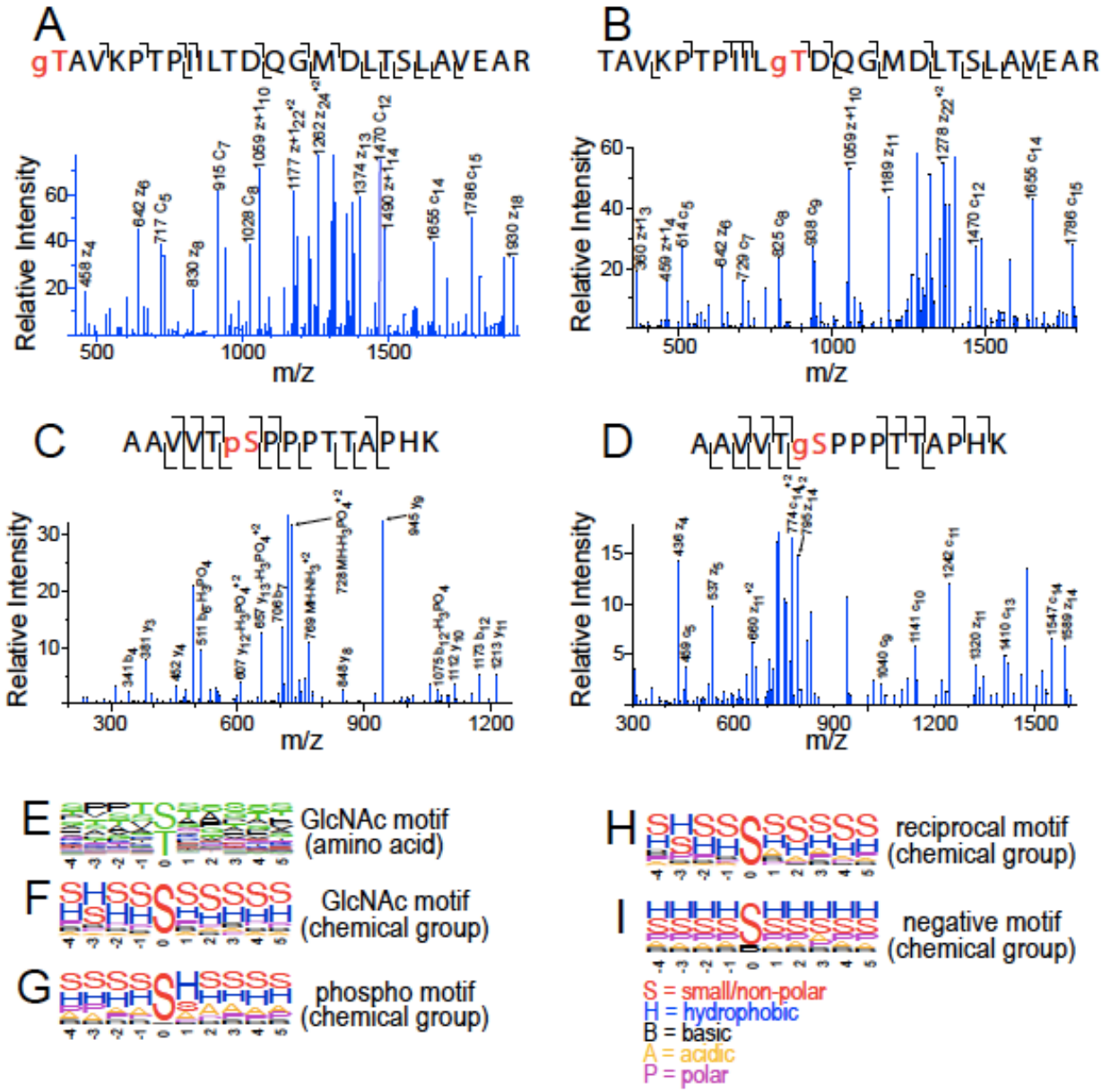


Figure 4.10 PTM-modified MS/MS spectra and motif analysis.
Peptides from the GlcNAc enrichment were analyzed using ETD, while those from the phosphopeptides enrichment were analyzed using CAD.
a-b) The peptide TAVKPTPIILTDQGMDLTSLAVEAR GlcNAcylated at the 1st or 11th amino acid, respectively.
c-d) The peptide AAVVTSPPPTTAPHK phosphorylated or GlcNAcylated at the serine in the 6th position, respectively.
e) Sequence logo for an alignment of GlcNAcylation sites identified in this study.
f-i) Sequence logo where individual amino acids are grouped by chemical property. “S” designates small/non-polar (A, G, S, T); “A” designates acidic (D, E); “B” designates basic (H, K, R); “H” designates hydrophobic (C, R< I, L, M, P, V, W); and “P” designates polar (N, Q, Y). Included are the chemical property logos for the GlcNAc motif, phospho motif, co-modified sites, and the background distributions.

4.6.5. Kinases are enriched for both types of PTMs

We identified one site of GlcNAcylation on OGT itself; however, we did not identify any phosphorylation on OGT in our synaptosome preparation despite the protein being present at relatively abundant levels, with 28 unique peptides identified. Olsen and colleagues previously identified six different phosphorylation sites on OGT from mitotically active cells (Olsen et al., 2010). On the O-GlcNAcase, we identified two sites of phosphorylation and no sites of GlcNAcylation. We identified 280 proteins annotated with protein kinase activity in the Gene Ontology (GO:0004672) and 87 protein phosphatases (GO:0004721). While 66% of kinases were phosphorylated, only 48% of proteins in this dataset were phosphorylated ($p < 3.8\ e^{-11}$, hypergeometric distribution). In addition,16% of kinases were O-GlcNAcylated, in contrast to 10% of proteins overall ($p < 3.6\ e^{-4}$, hypergeometric distribution). In contrast, protein phosphatases were not found to be PTM-modified at rates different from the overall dataset (52% phosphorylated and 8% GlcNAcylated). This evidence supports the notion that O-GlcNAcylation interacts with phosphorylation via OGT's regulation of (at least a subset of) kinases.

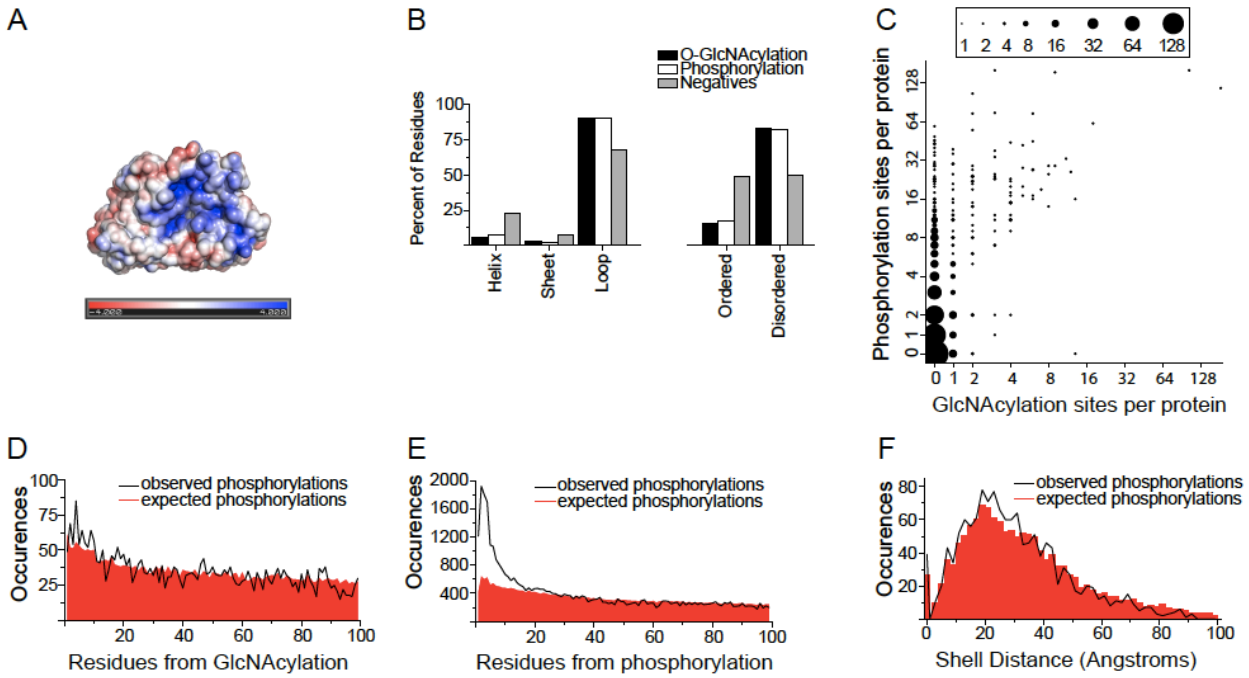


Figure 4.11 Structural aspects of OGT-substrate specificity
a) Charge distribution along the surface of the OGTs catalytic domain.
b) Structural comparison of GlcNAcylation and phosphorylation sites with a background list of non-modified serines and threonines.
c) Comparison of the number of identified GlcNAcylation sites *versus* phosphorylation sites per protein for the 20% highest abundance proteins, where our ability to identify PTMs was highest. The size of each data-point is proportional to the square root of occurrences. Note: There were some 2060 proteins at the 0-0 data-point in (c). For clarity, this data-point was given a size of 20 rather than 45.3 ($2060^{1/2}$).
d) Expected *versus* observed sequence distances between sites of GlcNAcylation and phosphorylation.
e) Expected *versus* observed sequence distances between pairs of phosphorylated residues.
f) Expected *versus* observed three-dimensional distances between sites of GlcNAcylation and phosphorylation, considering residues in solved structures or high quality homology models.

4.6.6. OGT-Substrate docking models generate hypothesis for properties mediating specificity

Recently, the crystal structure of human OGT in a complex with a model GlcNAcylated peptide has been published (Lazarus et al., 2011), which established that the transferase makes contacts primarily with the backbone of the substrate polypeptide. To investigate possible tertiary structural elements that may play a role in substrate recognition, we explored how GlcNAcylated proteins in our dataset would fit into the transferase active site using the program PatchDock, a molecular docking algorithm

based on shape complementarity principles (Duhovny et al.) (Schneidman-Duhovny et al., 2005). This was followed by use of Fast Interaction REfinement in molecular DOCKing (FIRE DOCK) (Andrusier et al., 2007) (Mashiach et al., 2008) to sort the docking solutions by energetic score. We docked crystal structures or very high quality homology models (*i.e.*, those models with greater than 85% sequence identity to the template structure used for modeling) to the human OGT structure with and without the domain containing tetratricopeptide repeats (TPR). In the structure of OGT, these TPR domains have been hypothesized to restrict access to the catalytic site (Lazarus et al., 2011). When this domain is removed (likely through a hinge-linke motion), a large basic patch, is revealed that encompasses the catalytic site of OGT (Figure 4.4a). A complementary acidic patch is present on the TPR domain that interacts with this basic patch. Constraints were applied to tether the GlcNAcylated residue within 10Å of the catalytic site in OGT. In this fashion, we docked 32 modified peptides characterized in this study to the OGT structure. It is interesting to note that only 1 of the 32 tested proteins docked to OGT when the TPR domain was attached, but 23 of the 32 tested proteins were able to dock to OGT once the TPR domain had been removed and the basic patch revealed. This is further evidence to support the theory that the TPR domain must swing out to allow substrate proteins to bind OGT.

To investigate any electrostatic interactions occurring at the protein-protein interface of OGT and docked substrate proteins, we aligned all docked solutions and color-coded residues according to side chain chemical properties. No obvious patterns emerged from this investigation. We then took every acidic and basic residue on OGT surrounding the catalytic site and identified any oppositely charged residues on docked

substrates within 10Å of the corresponding residue on OGT. Some electrostatic interactions were identified that could possibly be helping substrate proteins bind to OGT (Trinidad 2011, Figure S2A-D).

4.6.7. PTMs occur primarily on disordered loop regions

To gain insight into what secondary protein structural elements may be important for localization of these modifications, we determined the frequency with which they appeared on loops, alpha helixes, or beta sheets. Relative to the distribution of these structural elements in general, both GlcNAcylation and phosphorylation moieties were enriched within loops and relatively less prevalent within sheets or helixes (Figure 4.4b). For both PTMs, the site of modification occurred on loops approximately 90% of the time. Additionally, we calculated to what extent these PTMs were found in ordered *versus* disordered regions of protein structure. Both PTMs were approximately six-fold more likely to occur on disordered rather than ordered regions of protein structure (Figure 4.4b).

4.6.8. Respective PTM counts on individual proteins are weakly correlated

To investigate how these two PTMs might be interacting at the level of individual proteins, we examined the number of phosphorylation sites per protein as a function of the number of O-GlcNAcylation sites per protein (Figure 4.4c). There is a rough correlation between the frequencies of these two PTMs ($r^2 = 0.25$). Interestingly, the vast majority of proteins partitioned to the top left half (i.e. with a phospho:GlcNAc ratio > 1). The number of GlcNAc sites identified per protein was roughly equal to the minimum number of phosphorylation sites identified per protein (particularly when the

number of GlcNAc sites was > 2). However, for many proteins we observed extensive phosphorylation and only a limited number of O-GlcNAcylation sites.

In contrast, the only heavily GlcNAcylated protein that was not also heavily phosphorylated was CCR4-NOT transcription complex subunit 1. While estimated to be relatively abundant in our preparation, as a transcription factor, this protein likely partitions between the nucleus and cytoplasm (for a review see (Collart, 2003)). Regulation of gene transcription is a protein functional class known to be preferentially GlcNAcylated (Jackson and Tjian, 1988). A single site of phosphorylation on CCR4-NOT has been reported (Tang et al., 2007). Since only a minor fraction of this protein was present in our synaptosome preparation, it is possible that analysis of a total cell lysate (rather than of a specific organelle) would reveal additional sites of CCR4-NOT phosphorylation.

4.6.9. Single residues show no cross-talk between PTM types

As noted above, in 137 instances we observed phosphorylated peptides where the site of phosphorylation was the same as the site of GlcNAcylation observed on a GlcNAcylated peptide, representing 8% of the GlcNAcylation sites identified. While this number of reciprocally modified sites suggests cross-talk between these two PTM systems, given the extensive number of phosphorylation and GlcNAcylation sites we identified in this study, it is expected that both PTMs would map to the same amino acid residue at some frequency by chance alone. If these two PTM systems have evolved to cross-talk functionally, the observed frequency with which the same residue was found modified by both PTMs should substantially exceed the frequency predicted by chance alone. For a given protein, the number of co-modified sites expected by chance alone is

a function of the total number of serine and threonine residues on that protein as well as the phosphorylation and GlcNAcylation frequencies for that protein (*i.e.* observed modification sites with respect to total modifiable serines and threonines). However, one potential confound of the analysis is that not all serine and threonine residues may be surface-accessible and hence able to be modified by either PTM. We therefore restricted our analysis to disordered regions of protein structure, which encompassed approximately 50% of a given protein sequence (Figure 4.4b). Summing the expected co-modifications across all proteins in our dataset resulted in a prediction of 136 instances of overlapping modification. Therefore, while both PTMs are preferentially targeted to disordered regions of protein structure, within these disordered regions we find no increased propensity for GlcNAcylation to occur on the same residue as phosphorylation.

4.6.10. PTM types show very weak cross-talk within primary structure proximity

Spatial proximity between sites of GlcNAcylation and phosphorylation has been posited as a mechanism for structural cross-talk (Copeland et al., 2008)(Hart et al., 2011). If an organism has evolved to utilize such a mechanism, we reasoned that sites of GlcNAcylation should display an increased propensity to be localized proximal to sites of phosphorylation. For each site of GlcNAcylation, we calculated the distance along the primary sequence to the nearest site of phosphorylation (Figure 4.4d). For each site, we also calculated the expected phosphorylation distance distribution. This was calculated using the native distribution of serine and threonine residiues on that protein and assuming that they could be randomly phosphorylated based upon the phosphorylation frequency of that protein (with all calculations limited to disordered regions). Within five

amino acids of a site of GlcNAcylation, we observed a very subtle increase in the presence of phosphorylation sites, relative to expected distribution. However, we also investigated the distribution of phosphorylation sites with respect to each other on multiply phosphorylated proteins (Figure 4.4e). In stark contrast to the GlcNAc-phosphorylation distribution, phosphorylation sites showed a very strong preference to cluster together with respect to the protein primary sequence. Such clustering of phosphorylation sites within a protein has been previously reported (Yachie et al., 2009) (Schweiger, Regev and Linial, Michal, 2010) (Moses, Alan M et al., 2007). This minimal increase in localization of phosphorylation near sites of GlcNAcylation suggests that the two types of modification have not evolved to cross-talk via co-localization nearby in primary structure.

4.6.11. PTM types show no cross-talk within tertiary structure proximity

Primary sequence distance is an indirect measure of inter-residue distance within a protein three dimensional structure. We therefore investigated the spatial relationship of GlcNAc to phosphorylation with respect to protein three dimensional structure. Of the 466 proteins we observed with both types of PTMs, 111 were present in ModBase with high quality three dimensional models covering both sites of modification. Using shells of increasing radii, we examined the extent to which phosphorylation sites were found to be spatially proximal to O-GlcNAcylation sites. For each protein, we then calculated the expected number of phosphorylation sites at each distance using that protein's phosphorylation frequency and the distribution of serine/threonine residues. Serine and threonine residues within 100Å of a site of GlcNAcylation show no increase in phosphorylation frequency relative to the protein overall (Figure 4.4f). This result

suggests that the two types of modification have not evolved to cross-talk via close spatial proximity.

4.7. Discussion – O-GlcNAcylation and crosstalk with phosphorylation

4.7.1. O-GlcNAcylation is a widespread phenomenon

Previous investigations of protein O-GlcNAcylation have been limited in scope and in particular have lacked analogous characterization of phosphorylation for modified proteins occurring in the same biological preparations. The results presented here represent a 20-fold increase in the number of GlcNAcylation sites identified from any sample with endogenous levels of GlcNAcylation. Our extensive GlcNAcylation coverage of both proteins modified and sites occupied, coupled with over 16,500 sites of phosphorylation allowed us to systematically characterize GlcNAc distribution on synaptic proteins and address potential cross-talk between these two post-translational modifications. The increased coverage reported in this study is mainly due to three factors: (a) the use of more sensitive mass spectrometry (an Orbitrap Velos equipped with ETD fragmentation), (b) high pH fractionation of the GlcNAc-enriched fractions prior to LC-MS/MS, and (c) improved efficiency of the lectin-enrichment step. The primary improvement in the LWAC step is the switch from an agarose-immobilized lectin to one immobilized on POROS resin(Afeyan etal., 1991) carried out in three rounds of enrichment.

4.7.2. Proteomic results demonstrate the physiological role of O-GlcNAcylation in the brain

UDP-GlcNAc, the terminal product in the hexosamine biosynthetic pathway, is used by OGT to modify its substrates. Alterations in cellular energy state that increase UDP-GlcNAc levels have been shown to increase global protein GlcNAcylation (Yao et al., 2007)(Liu et al., 2000)(Housley et al., 2008). Substrate recognition is presumed to be partially regulated via adapter protein interactions with TPR domains on OGT. Such a mechanism could enable OGT to selectively modify certain substrates in response to global changes in UDP-GlcNAc levels (Lubas and Hanover, 2000)(Yang et al., 2002). GlcNAc levels in discrete subcellular compartments respond differentially in response to serum stimulation (Carrillo et al., 2011). However, as OGT has thousands of potential unique protein targets in mouse, activation of OGT (e.g. via increased UDP-GlcNAc levels) will likely result in modification of many substrates in concert.

GlcNAcylation plays a critical role in neuronal biology. Neuron-specific knockout of OGT leads to early postnatal death, which suggests a role for this enzyme in pathways basic for survival (O'Donnell et al., 2004). OGT is enriched at synapses (Cole and Hart, 2001). In addition, GlcNAc has been implicated in a diverse set of neuronal processes such as axonal branching and LTP at CA3/CA1 hippocampal synapses (Francisco et al., 2009)(Tallent et al., 2009).

We examined potential biological functions of GlcNAc using gene ontology analysis (http://amigo.geneontology.org). For this analysis, we used a background consisting of proteins in our dataset not found to be GlcNAc modified and of a similar abundance distribution to the GlcNAcylated proteins. Consistent with GlcNAc

modifications occurring on a large percentage of proteins, there were no GO categories in which GlcNAcylated proteins were significantly (greater than 50%) enriched. This suggests that in synaptic regions of the brain, modification by GlcNAc acts at a very broad level to regulate cellular function.

The protein bassoon is extensively modified by both phosphorylation and GlcNAcylation. Bassoon is a core component of presynaptic active zones, and as a component of Piccolo-Bassoon transport vesicles participates in targeting of cargo to distal axons. The binding of Bassoon to dynein light chain is thought to regulate transport of these vesicles along microtubules (Fejtova et al., 2009). Bassoon contains three functional dynein light chain binding motifs. We identified GlcNAcylation sites within two of these motifs, while none of them was found to be phosphorylated. This result indicates a potential role for GlcNacylation in regulation of vesicular transport.

4.7.3. PTMs can potentially cross-talk at multiple levels

Broadly speaking, cross-talk involving the two types of PTMs can occur *via* three distinct (yet non-mutually exclusive) mechanisms: at the structural level involving proteins modified by both PTMs; at the catalytic level involving regulating activity of one type of PTM-modifying enzyme by a second PTM; and at the sub-cellular localization level whereby PTM-mediated transport of one PTM regulates access to cellular environments containing enzymes mediating levels of the second PTM (Hunter, 2007).

Cross-talk has been defined as “the action of one posttranslational modification influencing the addition or removal of another posttranslational modification” (Hart et al., 2011). In this context, one can imagine both positive and negative cross-talk occurring at various distances within a protein’s three dimensional architecture. Primary cross-talk

(occurring at the same amino acid) will be necessarily negative when both PTMs cannot occur simultaneously at the same amino acid. Secondary cross-talk could occur either proximally or distally with respect to the initial PTM, and in principle could be both positive and negative, depending on the protein. In the case of proximal cross-talk, addition of the first PTM could obscure or complete a motif regulating addition of the second PTM, as is the situation for the phosphodegron motif (Petroski and Deshaies, 2005). In the case of negative cross-talk, the first PTM could also alter the region around the site sterically or electrostatically to impair addition of the second PTM. For distal cross-talk, the first PTM would either have to allosterically modify protein structure, or act as a recruitment site for an additional protein that in turn causes recruitment of the second PTM-modifying enzyme, or the first PTM may alter subcelluar localization of the protein (and thus modify protein localization with respect to enzymes regulating addition and removal of the second PTM).

In this study we have identified some 137 instances of individual serine and threonine residues reciprocally modified by both phosphorylation and GlcNAcylation, increasing several-fold the number of such cases reported. However, in contrast to previous studies, the scope of our analysis allowed us to demonstrate that these 137 instances are essentially what one would expect to find by chance alone given the rates with which both PTMs modify their substrates. As such, there is limited evidence that there was evolutionary pressure to increase primary cross-talk. Nevertheless, when occurring at the same amino acid, the two PTMs necessarily antagonize each others' occupancy levels, and this is therefore primary cross-talk by definition. To engage in this type of cross-talk at a functionally relevant biological level would require that the

stoichiometry of modification be sufficiently high to significantly alter the concentration of unmodified protein. While absolute stoichiometries of modification were not measured in this present study, recent reports have examined these values for both PTMs on a range of proteins (Rexach et al., 2010)(Wu et al., 2011). An examination of GlcNAcylation stoichiometry at the protein level for seven proteins showed a range of 2 to 100%, although the stoichiometries at individual sites for multiply-modified proteins will likely be lower. Wu and collegues calculated phosphorylation stoichiometries for over 5000 yeast phosphorylation sites. These values varied from 1 to 100%, with a median phosphorylation stoichiometry of approximately 25%. Based upon these results, it would appear that basal stoichiometries for both PTMs are in a range where moderate increases in one PTM would be expected to result in a decrease in the stoichiometry of the other PTM.

It has recently been reported that GlcNAc and phosphorylation levels are of similar abundance at spindles and midbodies (Wang et al., 2010b). However, without controlling for differential detection efficiency of the two PTMs, it is difficult to make such claims with a high degree of confidence. When we attempt to account for this effect, we observed 11-fold more sites of phosphorylation than GlcNAcylation in synaptosomes. While different subcellular compartments will undoubtedly have different ratios of these two PTMs, our results encompass measurements for over 6,000 proteins, which suggests that the modification rates of these PTMs in the overall proteome are very similar.

An important caveat with the current study is that it only represents a static snapshot of how these two PTMs distribute in synaptosomes. Mass pharmacological

stimulation of cells in culture clearly results in several fold changes in phosphorylation and GlcNAcylation state of many proteins. Whether physiologically relevant conditions that result in changes of similar magnitude exist *in vivo* remain to be seen. Interactions between GlcNAc and phosphorylation may exist during dynamic changes that cannot be readily discerned from static snapshots. Finally, knowledge about absolute stoichiometry of modification, in particular at those residues found to harbor both types of PTMs, may help to shed light on how these PTMs might compete for sites of co-occupancy.

4.8. Methods in characterizing O-GlcNAc modifications

4.8.1. Preparation of mouse synaptic membranes

Synaptic membrane samples were purified at 4ºC, as described previously (Trinidad et al., 2006). Briefly, brains from adult mice (strain C57BL/6J) were dissected; the cerebellum was removed and the brains immediately frozen in liquid nitrogen. Material from several animals was combined prior to the biochemical purification. The brain tissue was homogenized in a sucrose buffer containing a mixture of phosphatase inhibitors (1 mM Na_3VO_4, 1 mM NaF, 1 mM Na_2MoO_4, 4 mM sodium tartrate, 100 nM fenvalerate, 250 nM okadaic acid), and cleared by centrifugation. 10 ml of buffer was used per gram of brain. The membranous fraction was layered on a sucrose density and fractionated by centrifugation. Synaptic membranes were collected at the 1.0-1.2 M interface and harvested by centrifugation.

4.8.2. Digestion of synaptosome samples

30 mg of synaptosome was resuspended in 1 ml buffer containing 50 mM ammonium bicarbonate, 6 M guanidine hydrochloride 6X Roche Phosphatase Inhibitor Cocktails I and II, and 6X PugNAc inhibitor. The mixture was incubated for one hour at 57ºC with 2 mM Tris(2-carboxyethyl)phosphine hydrochloride to reduce cysteine side chains, these side chains were then alkylated with 4.2 mM iodoacetamide in the dark for 45 min at 21ºC. The mixture was diluted six fold with ammonium bicarbonate to a final ammonium bicarbonate concentration of 100 mM and 1:50 (w/w) modified trypsin (Promega, Madison, WI, USA) was added. The pH was adjusted to 8.0 and the mixture was digested for 12 hours at 37ºC. The digests were desalted using a C_{18} Sep Pak cartridge (Waters, Milford, MA, USA) and lyophilized to dryness using a SpeedVac concentrator (Thermo Electron, San Jose, CA, USA).

4.8.3. Preparation of the lectin weak affinity chromatography column

300 μg of POROS Al resin was reacted with 25 mg of WGA per the manufacturer's instructions. Briefly, 10 mM bicine, pH 7.5 was used as the reaction buffer and 5 mg/ml sodium cyanoborohydride was added along with 200 μl 2M sodium sulfate. The mixture was rotated at 21°C for 24 hours. The resin was spun down and washed with 10 mls bicine, then quenched with 10 mls 200 mM Tris/acetate buffer, pH 7.5 and 200 μl sodium cyanoborohydride (100 mg/ml). The resin was then packed into a 2 x 250 mm stainless steel column.

4.8.4. Enrichment of GlcNAcylated peptides using a WGA column

Peptides were resuspended in 50 μl buffer A (100 mM Tris pH 7.5, 150 mM NaCl, 2 mM $MgCl_2$, 2 mM $CaCl_2$, 5% acetonitrile). Peptides were run over the column at 125 μL/min.

GlcNAcylated peptides eluted as an unresolved smear on the right side of the flow thru tail peak. After 1.3 ml, an additional 100 µL of 20 mM GlcNAc in buffer A was injected to elute any remaining peptides. To decrease the chance of overloading the column each 10 mg portion was split into two 5 mg samples and run separately and the GlcNAc enriched fractions were combined subsequently. For subsequent rounds of enrichment, the pooled fractions were run together in a similar fashion as before.

4.8.5. Enrichment of phosphorylated peptides using titanium dioxide

Peptides were resuspended in 250 µL buffer B1 (1% TFA, 20% acetonitrile). The samples were run at 80 µL/min in buffer B1 over an analytical guard column with a 62 µL packing volume (Upchurch Scientific, Oak Harbor, WA USA) packed with 5 µm titanium dioxide beads (GL Sciences, Tokyo Japan) (Larsen et al., 2005)(Pinkse et al., 2004). The column was rinsed with H_20, then eluted with 3 x 250 µL saturated KH_2PO_4 followed by 3 x 250 µL 5% phosphoric acid. A switching valve was used to direct these elutions onto a C_{18} macrotrap peptide column (Michrom Bioresources, Auburn, CA, USA). The peptides were washed with H_20 then eluted with 50% acetonitrile, and this solution was lyophilized to dryness using a SpeedVac concentrator.

4.8.6. High pH reverse phase chromatography

High pH RP chromatography was performed using an ÄKTA Purifier (GE Healthcare, Piscataway, NJ, USA) equipped with a 1 x 100 mm Gemini 3µ C18 column (Phenomenex, Torrance, CA). Individual GlcNAc-enriched or phospho-enriched fractions loaded onto the column in 1% buffer A (20 mM NH_4FA, pH 10). Buffer B consisted of buffer A with 50% acetonitrile. The gradient went from 1% B to 21% B over 1.1 ml, to 62% B over 5.4 ml, and then directly to 100% B. 20 fractions were collected

and dried down using a SpeedVac concentrator. 1 mg of the GlcNAC and phospho depleted flow through material was separated by high pH reverse phase to collect 60 fractions.

4.8.7. Mass spectrometry analysis

All peptides were analyzed on an LTQ Orbitrap Velos equipped with a nano-Acquity UPLC. GlcNAc-enriched fractions were analyzed using electron transfer dissociation (ETD). Phospho-enriched fractions were analyzed using collision activated dissociation (CAD). Non-modified peptides were analyzed using HCD. Peptides were eluted using a 90 minutes gradient. Data was searched against the Uniprot *Mus musculus* database (downloaded January 11, 2011). To this database, a randomized version was concatenated to allow determination of false discovery rates. The cleavage specificity was set to “trypsin”, allowing for one missed cleavage. Carbamidomethylation of cysteine residues was set as a fixed modification. Acetylation of protein amino termini, oxidation of methionine residues, pyrolization of amino terminal glutamines, and loss of protein terminal methionines were set as variable modifications. For the GlcNAc search, HexNAc modification of serine, threonine and asparagines was set as variable modifications. For the phospho search, phosphorylation of serine, threonine and tyrosine was set as variable modifications. Data was searched initially with a 20 ppm tolerance of the parent ion, 0.6 Da tolerance of MS/MS measured in the ion trap (CAD and ETD) and 20 ppm tolerance for HCD MS/MS. The precursor mass tolerance was then recalibrated on a file by file basis based upon the mass accuracy of high scoring peptides. Final precursor mass tolerances were between 10 and 13 ppm.

For the resulting output, the corresponding Unigene name, gene, and entry numbers were appended (http://www.ncbi.nlm.nih.gov/unigene). Uniprot entries were grouped by their corresponding Unigene genes and redundant peptides within a gene group were removed.

For the non-modified peptide identifications, a peptide expectation value threshold ≤ 0.01 was used. A protein was considered positively identified if the most confident peptide for that protein had an expectation value ≤ $1e^{-7}$. This resulted in the identification of 6,190 Unigene entries and 58,825 unique peptides. At this threshold, the decoy database contained 6 entries and 8 unique peptides (protein FDR = 0.097%, peptide FDR = 0.013%).

GlcNAcylation and GalNAcylation both increase the mass of the modified peptide by the same amount (203.08 Da), and therefore these two PTMs are indistinguishable in the mass spectrometer. While GlcNAcylation occurs almost exclusively on intracellular protein regions, the extracelluar domain of Notch is O-GlcNAcylated (Matsuura et al., 2008). Peptides were assigned as ambiguous between GalNAcylated or GlcNAcylated based upon their annotation in Uniprot as located in extracelluar or luminal regions. These include mitochondrial proteins, which possess both complex carbohydrate modifications as well as O-GlcNAcylation (Hu et al., 2009) (Love et al., 2003).

4.8.8. Calculations of expected *versus* observed frequencies.

The expected *versus* observed cross-talk between the two types of PTMs was determined in three different contexts. (1): For cross-talk at a single residue, we counted the number of times a residue was observed to be both O-GlcNAcylated and

phosphorylated in different experiments. We also calculated the number of times this co-modification was expected to occur by chance as $n * r_g * r_p$, where n represented the number of serines and threonines in one protein and r_g and r_p were the rates of O-GlcNAcylation and phosphorylation, respectively for the same protein (calculated as the number of each modification over the total number of serines and threonines). The expected number of co-modifications were summed across all proteins and compared to the observed value using χ^2 evaluation. (2) For cross-talk at the primary structure level, we compared the observed *versus* expected values for the number of times an O-GlcNAcyation event was observed at a distance of n residues from a phosphorylation, for different values of n along the protein sequence. Thus, for each O-GlcNAcylation, we counted the number of phosphorylations at distance n to create a distribution of observed distances. Expected distances were calculated as in (1), limiting the serines and threonines to those also at distance n. Values of n were binned in intervals of five to create a larger sample size. Expected values were compared to observed values at each bin interval using χ^2. (3) For cross-talk at the spatial proximity level, we compared the expected and observed values for the number of times an O-GlcNAcylation was observed within n Å of a phosphorylation, for different values of n. Calculations proceeded as in (2). Analysis was limited to those modifications falling in a solved structure or good quality homology model of the protein. In all co-modification analysis, we limited the serines and threonines to those falling in disordered regions only.

4.8.9. Structural Analysis of PTMs

For proteins having an experimentally solved structure or good quality homology model in ModBase (ref PMID 21097780), secondary structure assignments for peptides were

created by DSSP (ref PMID 6667333). For proteins with no structure information available, secondary structure was predicted using PSIPRED (ref PMID 10493868). For all proteins, disorder was predicted using the DISOPRED algorithm (ref PMID 14579348).

Chapter 5. Host pathogen protein interactions

Pathogens have evolved numerous strategies to successfully invade their hosts, acquire nutrients, and evade their immune defenses (Munter et al. 2006). These strategies often involve direct interactions between host and pathogen molecules, including the formation of protein complexes (Stebbins 2005). Much remains to be learned about the network of interactions between host and pathogen proteins and the specificity mediating these interactions. If the intraspecies interaction network of *Saccharomyces cerevisiae* is a guide, several independent large-scale studies are likely required for a comprehensive mapping of host–pathogen interactions (Collins et al. 2007).

Interactions between host and pathogen proteins are typically studied using traditional small-scale biochemical and genetic experiments, which focus on one protein or pathway at a time. Large-scale interaction discovery methods, such as tandem affinity purification and yeast– two-hybrid experiments, enable more comprehensive detection but at the cost of significant false-negative and false-positive error rates (Hart et al. 2006). Computational methods have demonstrated utility in improving the coverage, accuracy, and efficiency of identifying protein–protein interactions in combination with experimental data sets (Jansen et al. 2003; Lee et al. 2004) and are likely to similarly complement large-scale experimental efforts to characterize host–pathogen interaction networks.

As discussed extensively in section 1.4, interactions involving peptides are of particular interest in these contexts. Peptide-mediated interactions are prevalent in host

signaling networks, which are often disrupted or mimicked by pathogens to carry out the processes described above. Thus, both protein-protein and protein-peptide interactions are important in pathogenesis. This chapter examines both of these types of interactions from a specificity standpoint. First, an application of a statistical method to predict protein-protein interactions is presented in a cross-species context. These interactions can be between two proteins, but a subset involves protein-peptide association. Second, a specific examination is conducted of protease inhibition by the *P. falciparum* falcipain-2 prodomain, which is autoinhibitory as well as selective for certain human cathepsins. Together, these studies demonstrate the predictive and explanatory aspects of protein-peptide interaction specificity in pathogenic contexts.

5.1. Introduction – High throughput prediction of host-pathogen interactions

Genome sequencing has changed the scale and diversity of biomedical problems amenable to investigation as complete sequences are now available for many species, including human and a number of biomedically relevant microbes (Guttmacher and Collins 2005). 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 (Westbrook et al. 2002; Bader et al. 2003). 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–pathogen interactions.

Here we hypothesize that host–pathogen protein interactions, knowledge of which is severely lacking, can be inferred from the growing body of experimentally observed

interactions, which is reaching saturation in some species. We previously showed that this approach can be useful in predicting intraspecies interactions (Davis et al. 2006). We now provide three additional lines of evidence that suggest the hypothesis is a valid one and that the developed protocol can predict functionally relevant host–pathogen protein interactions. The protocol identifies pairs of host and pathogen proteins with similarity to proteins known to interact, assesses the likelihood of interaction based on structural modeling, and then identifies those pairs with a greater chance of encounter as suggested by their subcellular location and expression properties. The result of the protocol is an enriched candidate set that is suitable for subsequent experimental study. We have applied the protocol to 10 human pathogens, including species of mycobacteria, kinetoplastida, and apicomplexa, which are responsible for ‘‘neglected’’ human diseases. These pathogens cause tropical diseases with a significant global burden, infecting over 1 billion people and incurring over 1 million annual deaths (World Health Organization 2003).

We first describe the protocol, detailing the data sources, the computations used, and its performance on intraspecies protein interactions in *S. cerevisiae*. We then present the predictions made for the 10 pathogens and assess them by three independent computational procedures. We then discuss the observed performance of the method and potential future improvements. We present several specific predictions that warrant experimental follow-up. Finally, we conclude by discussing the implications of these results for understanding the molecular mechanisms of pathogenesis.

5.2. Results – Generation of interaction predictions in neglected diseases

The protocol begins with the target set of host and pathogen protein sequences (Figure

5.1).

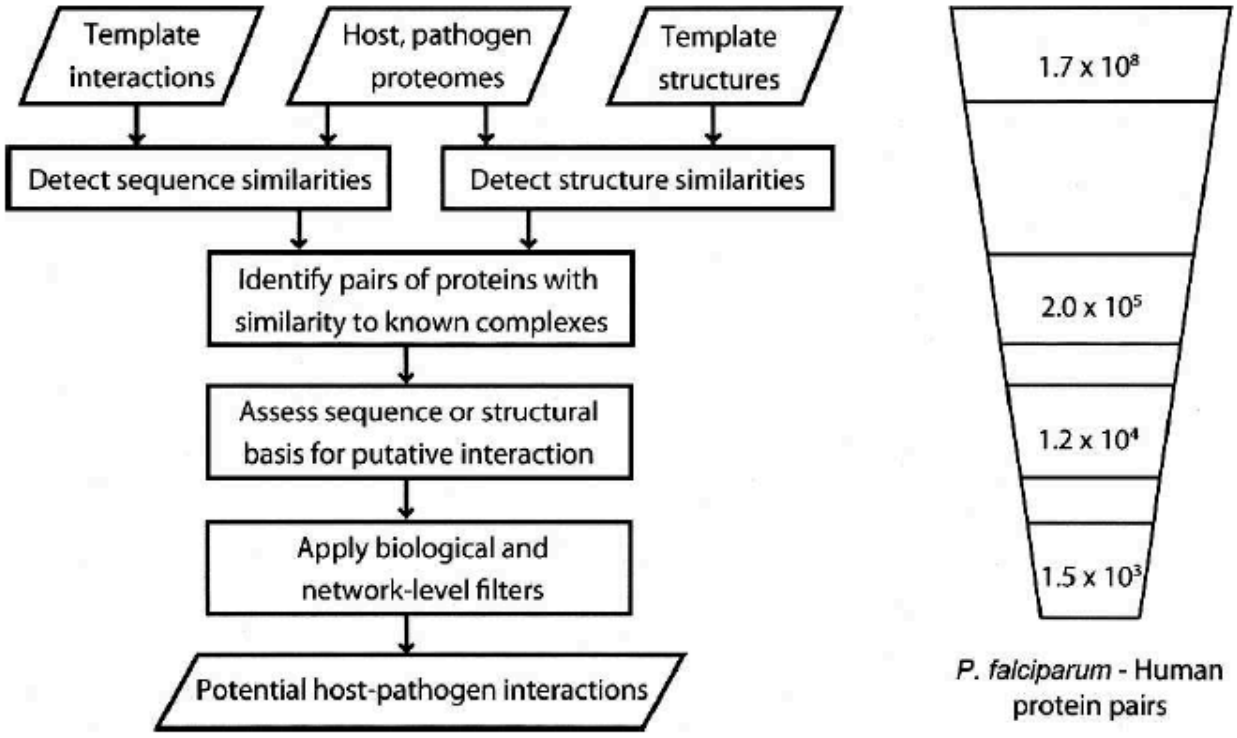


Figure 5.1 Host-pathogen interaction prediction protocol.
The protocol begins with the set of host and pathogen proteins. Sequence matching procedures are then used to identify similarities between the host or pathogen proteins and proteins with known structure or known interaction partners. A structure-based statistical potential assessment, or a sequence similarity score in the absence of structure, is then used to predict 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 about five orders of magnitude (Table 5.2).

5.2.1. Detecting sequence and structure similarities and identifying pairs of proteins with similarity to known complexes

Similarities were first detected between the target sequences and components of known protein complexes, using an automated comparative protein structure modeling pipeline. The fraction of the pathogen proteomes for which a suitable interaction template was identified varied from 16% of *Trypanosoma cruzi* sequences to 25% of *Cryptosporidium parvum* sequences, while the human proteome coverage was 34%

(Table 5.1).

Pathogen	Protein sequences	With interaction templates		With biological data	
M. leprae	1601	359	22%	1023	64%
M. tuberculosis	3954	729	18%	2551	65%
L. major	8009	1908	24%	3749	47%
T. brucei	8965	1817	20%	4040	45%
T. cruzi	19,245	3147	16%	8604	45%
C. hominis	3886	780	20%	1591	41%
C. parvum	3806	958	25%	1828	48%
P. falciparum	5342	1126	21%	4691	88%
P. vivax	5334	1131	21%	413	8%
T. gondii	7787	1311	17%	3627	47%
H. sapiens	32,010	10,993	34%	26,595	83%

Table 5.1 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 (Davis 2007, Table S1).

Pairs of host and pathogen proteins that each had detectable similarity to components of a known interaction were then identified. The number of these pairs varied widely among the pathogens, with the prokaryotes having far fewer pairs than the eukaryotes (Table 5.2, column 2). For example, 43,528 host–pathogen protein pairs were identified for *Mycobacterium tuberculosis* (3,954 sequences, 18% template coverage), while 160,952 pairs were identified for *Cryptosporidium hominis* with approximately the same proteome size and interaction template coverage (3,886 sequences, 20% template coverage). Among the eukaryotic pathogens, the number of pairs varied approximately in proportion to the proteome sizes (Table 5.1 and Table 5.2).

Pathogen	Pairs with templates		Potential interactions		Filtered interactions	
M. leprae	26,234	(6200/359)	1351	(706/101)	13	(13/1)
M. tuberculosis	43,528	(6549/729)	2474	(992/240)	45	(41/13)
L. major	411,468	(9978/1908)	22,243	(2680/656)	289	(186/29)
T. brucei	427,884	(9935/1817)	20,797	(2546/661)	0	(0/0)
T. cruzi	750,419	(10,078/3147)	33,869	(2601/1028)	914	(356/138)
C. hominis	160,592	(9118/780)	7237	(1854/257)	79	(59/8)
C. parvum	203,570	(9242/958)	10,987	(2108/335)	211	(156/13)
P. falciparum	200,428	(9554/1126)	11,655	(2291/434)	1501	(826/216)
P. vivax	211,185	(9546/1131)	12,159	(2305/399)	34	(26/4)
T. gondii	216,187	(9638/1311)	7282	(2024/261)	0	(0/0)

Table 5.2 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. The numbers in parentheses represent the number of individual host/pathogen proteins involved in the interactions.

5.2.2. Assessing the sequence or structural basis of the potential interactions

Next, the sequence or structural basis of interaction between the identified pairs was assessed using sequence similarity and statistical potential scores, respectively. This step identified ~5% of the host–pathogen pairs identified in the previous step as possible interacting partners (Table 5.2), almost all (99.5%) of which were based on structural templates. The minimal contribution of sequence-based templates to the predictions is due to the stringent joint sequence identity threshold (≥80%) required to reliably transfer interactions (Yu et al. 2004; Mika and Rost 2006). The reduction in the number of pairs by the assessment step was greatest for the *Toxoplasma 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, fewer predictions were made for the prokaryotic than for the eukaryotic pathogens.

5.2.3. Applying biological and network-level filters

The interactions were then filtered by the biological context of their component proteins, such as life-cycle stage and tissue expression, and by network-level information

regarding the template usage frequencies. Interactions that met at least one host and one pathogen biological criterion were considered to pass the biological context filter (Table 5.1 and Davis 2007, TableS3). Next, the network-level filter flagged those predictions based on templates that were used for more than 1% of the total predictions, as these predictions exhibited a low level of interaction specificity. For example, many pairs of G-protein subunits α and β were predicted to interact based on the crystal structure of the G-protein Gi heterotrimer (Protein Data Bank [PDB] 1GG2).

The filters resulted in a wide range of reductions in predicted interactions (Table 5.2), due to the different levels of biological annotation used for the genomes. For example, *Plasmodium falciparum* had the highest biological annotation coverage (88%) and, as expected, the highest fraction of interactions that passed the biological and network-level filters (13%). This final set of *P. falciparum*–human interactions is five orders 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, *Trypanosoma brucei* and *T. gondii*, resulted in removal of all interactions. The type of annotation available for the pathogen proteins is particularly important. For example, both *T. brucei* and *T. cruzi* have biological annotation for 45% of their proteomes; however, filtering results in zero interactions for the former and 914 for the latter. This difference occurs because life cycle annotation is available for 1930 (10%) of *T. cruzi* proteins but only 120 (1%) of *T. brucei* proteins (Davis 2007, Table S3). The majority of the biological annotations are GO terms that do not pass the filtering criteria.

5.2.4. 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. The structure- and sequence-based prediction methods have both been previously benchmarked in the context of intraspecies interactions (Yu et al. 2004; Davis et al. 2006), and the results are briefly described in Section 5.4. In contrast to interspecies interactions, large experimental data sets of thousands of intraspecies interactions are available and ideal for benchmarking prediction methods. These benchmarking results remain informative in the host–pathogen context as the underlying biophysical chemistry remains the same. We assessed the quality of the protocol in the host–pathogen context in three additional ways.

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

The predicted interactions were first compared with the set of known host–pathogen interactions (Davis 2007, Table S1), which although too small to assess the method rigorously, still allow insight into the performance of the method. Our protocol recovered four of the 33 host–pathogen protein interactions published in the literature for the 10 pathogen species. Other known interactions were not identified because of the lack of available templates. None of these latter cases was due to incorrect assessment by our method. As expected, this result suggests that currently, a limitation of the protocol's coverage is the restriction to interactions with an appropriate template.

No interactions have been previously identified for three of the species we studied,

Leishmania major, *C. hominis*, and *C. parvum*. The method recovered 67% (n = 2) of the known *T. brucei*–human interactions. One of these interactions, an ornithine decarboxylase (ODC) interspecies dimer whose physiological relevance has not been established, was later filtered out of the predictions because it was based on a homodimer template. For the species with the most observed interactions, *P. falciparum* and *T. cruzi*, the method recovered 9% (n = 1) and 8% (n = 1) of the previously observed interactions, respectively. In both cases, the interactions were protease–protease inhibitor interactions.

5.2.6. Assessment II: Comparison to gene expression and essentiality data

Next, we compared our prefiltered predictions to genome-scale data sets 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 source of large-scale data sets describing host–pathogen interactions, even though only weak correlation has been observed between physical protein interactions and expression data (Mrowka et al. 2001; Jansen et al. 2002).

Previous studies have identified 194 *M. tuberculosis* genes that are essential for in vivo infection (Sassetti and Rubin 2003) and 286 genes that are up-regulated in granuloma, pericavity, or distal lung infection sites compared with in vitro conditions (Rachman et al. 2006). Comparison of these two sets of genes to the set of *M. tuberculosis* proteins predicted to interact with human proteins revealed minimal overlap (Davis 2007, Table S2). In fact, only one gene occurs in both experimental data sets and our predictions: Rv3910 (GI 15611046), a probable conserved transmembrane

protein. The overlap of our predictions with the set of genes upregulated during infection (23 genes) is greater than that between the two experimental sets of up-regulated genes and genes essential for infection (18 genes).

Previous studies have identified human genes that are differentially regulated in response to a variety of protozoal infections, in particular within the macrophage and dendritic cells of the immune system (Chaussabel et al. 2003). The human proteins predicted to interact with *L. major*, *M. tuberculosis*, and *T. gondii* include, respectively, 231, 78, and 169 proteins encoded by genes differentially expressed in macrophages and dendritic cells upon infection by these pathogens (Davis 2007, Table S2B) (Chaussabel et al. 2003).

5.2.7. Assessment III: Functional overview of predicted interactions

Rank	GO ID	Function	Number	Enrichment	*P*-value
(*a*) Cellular component of all human proteins predicted to interact with *M. tuberculosis*					
1	GO:0005776	autophagic vacuole	5	17.1	3.0×10^{-4}
2	GO:0005853	eukaryotic translation elongation factor 1 complex	5	12.2	2.2×10^{-3}
3	GO:0042101	T-cell receptor complex	5	8.5	1.6×10^{-2}
4	GO:0001772	immunological synapse	7	7.7	1.0×10^{-3}
5	GO:0005884	actin filament	8	5.3	4.3×10^{-3}
6	GO:0005746	mitochondrial electron transport chain	8	4.9	7.6×10^{-3}
7	GO:0044455	mitochondrial membrane part	12	3.8	1.2×10^{-3}
8	GO:0042995	cell projection	23	2.2	1.2×10^{-3}
9	GO:0015629	actin cytoskeleton	25	2.2	5.1×10^{-4}
10	GO:0031410	cytoplasmic vesicle	22	1.9	1.5×10^{-2}
(*b*) Biological process of all human proteins predicted to interact with *M. tuberculosis*					
1	GO:0006021	myo-inositol biosynthetic process	3	34.1	1.4×10^{-2}
2	GO:0019642	anaerobic glycolysis	5	34.1	6.5×10^{-6}
3	GO:0006422	aspartyl-tRNA aminoacylation	5	24.4	1.3×10^{-4}
4	GO:0032011	ARF protein signal transduction	7	23.9	3.4×10^{-7}
5	GO:0032012	regulation of ARF protein signal transduction	7	23.9	3.4×10^{-7}
6	GO:0046847	filopodium formation	6	17.1	1.1×10^{-4}
7	GO:0051014	actin filament severing	4	17.1	1.9×10^{-2}
8	GO:0043088	regulation of Cdc42 GTPase activity	5	14.2	4.5×10^{-3}
9	GO:0032489	regulation of Cdc42 protein signal transduction	5	14.2	4.5×10^{-3}
10	GO:0032318	regulation of Ras GTPase activity	5	14.2	4.5×10^{-3}
(*c*) Molecular function of all human proteins predicted to interact with *M. tuberculosis*					
1	GO:0016872	intramolecular lyase activity	3	34.1	5.6×10^{-3}
2	GO:0004512	inositol-3-phosphate synthase activity	3	34.1	5.6×10^{-3}
3	GO:0019967	interleukin-1, type I, activating binding	4	27.3	5.9×10^{-4}
4	GO:0004909	interleukin-1, type I, activating receptor activity	4	27.3	5.9×10^{-4}
5	GO:0004739	pyruvate dehydrogenase (acetyl-transferring) activity	3	25.6	2.2×10^{-2}
6	GO:0005094	Rho GDP-dissociation inhibitor activity	3	25.6	2.2×10^{-2}
7	GO:0004738	pyruvate dehydrogenase activity	3	25.6	2.2×10^{-2}
8	GO:0004591	oxoglutarate dehydrogenase (succinyl-transferring) activity	3	25.6	2.2×10^{-2}
9	GO:0004815	aspartate-tRNA ligase activity	5	24.4	5.3×10^{-5}
10	GO:0004459	L-lactate dehydrogenase activity	7	23.9	1.4×10^{-7}

Table 5.3. Functional annotation of human proteins predicted to interact with *M. tuberculosis*. The 10 (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 with the background, are listed. The analysis was done before application of the biological filters to prevent bias in the enriched terms. The enriched terms were identified and their significance computed by GO::TermFinder using a Bonferroni correction (Boyle et al. 2004).

Finally, we evaluated the functional relevance of the predicted interactions by searching for functional annotations of proteins that were significantly enriched in the human proteins predicted to interact with pathogens, compared with the whole human proteome. This analysis was done before the application of the biological filters to prevent introduction of filter bias into the functional profile of the predictions. The human proteins predicted to interact with pathogen proteins were significantly enriched in several gene ontology terms (Table 5.3). For example, the human 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 = 10^{-3}), T-cell receptor complex (8.5-fold enrichment, P =1.6 X 10^{-2}), and autophagic vacuole (17.1-fold enrichment, P = 3 X 10^{-4}). These terms all reflect the known immunobiology of this pathogen, which elicits a T-cell response and was recently found to be eliminated through autophagy (Gutierrez et al. 2004; Deretic 2006; Singh et al. 2006; Vergne et al. 2006). Similarly, the human proteins predicted to interact with *P. falciparum* proteins are enriched in terms such as extrinsic to plasma membrane (5.2-fold enrichment, P = 9.2 X 10^{-15}) and homophilic cell adhesion (4.2-fold enrichment, P = 2.8 X 10^{-21}).

The enriched functional terms that have not been previously implicated in infection represent either novel biological insights or false positives. Distinguishing between these two possibilities 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 (Tatusov et al. 1997) could result in erroneously predicted interactions causing the enrichment in the human–*P. falciparum* network for eukaryotic translation elongation factor (7.4-fold, P = 8.4 X 10^{-4}). Similarly, terms such as pyruvate deydrogenase activity (25.6-fold, P = 2.2 X 10^{-2}) and asparate-tRNA ligase activity (24.4 fold, P = 5.3 X 10^{-5}), which 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 could be filtered.

5.3. Discussion – Confidence and limitations in interaction predictions

We presented a protocol that reduces the number of host–pathogen protein pairs to an experimentally tractable set of predicted interactions, by a series of assessments: (1) identifying template interactions; (2) assessing the putative interaction, using structure if available; and, finally, (3) filtering using biological context and network-level information. For example, the procedure resulted in a five order of magnitude reduction in the number of possible human–*P. falciparum* protein interactions. Although it is not possible to directly assess the enrichment of true interactions in the predictions, previous assessment in the context of *S. cerevisiae* interactions found an enrichment of about two orders of magnitude. In addition, assessment of the method by comparison to known host–pathogen interactions (Davis 2007, Table S1), genomics data (Davis 2007, Table S2), and functional analysis (Table 5.3) suggests that the method is capable of enriching for functionally relevant interactions. We now discuss the observed performance of the method, present several specific predictions and their support in the literature, and close by discussing future developments and applications of the method to characterize host–pathogen and other types of interspecies interactions.

5.3.1. Limitations in coverage

The performance of the method 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 main factor that limits the coverage of our method is that, like all comparative approaches, it depends on previous experimental observations of similar interactions.

Despite the limited coverage, reflected in the low number of known interactions recovered by the method (four of 33), the availability of structure enables a more rigorous assessment of the interactions than that allowed by sequence alone (Davis et al. 2006). 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 (Neduva and Russell 2006). Peptide motifs that mediate protein interactions are being identified through a combination of computational and experimental methods (Tong et al. 2002; Neduva et al. 2005), and application of these motif-based methods will likely expand the coverage of host–pathogen protein interactions.

5.3.2. Errors in accuracy

Several factors affect the accuracy of the method. These include errors in the comparative modeling process (Marti-Renom et al. 2000), the coarse-grained nature of the statistical potential used to assess the interface residue contacts (Davis et al. 2006), and consideration of only interactions between individual domains (i.e., incorrectly predicting interactions that are unfavorable in the context of the full-length proteins). While these three sources of error affect both intra-species and host–pathogen protein interactions, an additional type of error uniquely affects inter-species interactions. As the pathogen and host species are both eukaryotic for eight of the 10 pathogens studied,

many of the predicted interactions are between core cellular components, such as translation machinery, metabolic enzymes, and ubiquitin-signaling components (Table 5.3). Although these interactions could potentially occur if the host and pathogen proteins encountered one another, their availability for such an encounter is not guaranteed. We used biological data, such as known exported pathogen proteins and known host–tissue targets, to address the ‘‘accessibility’’ issue. However, the precise spatial and temporal locations of these proteins are generally difficult to characterize. We expect this last source of errors to be diminished when the evolutionary distance between pathogen and host is greater, such as between bacterial or viral pathogens and their human hosts.

5.3.3. Specific examples of validated predictions

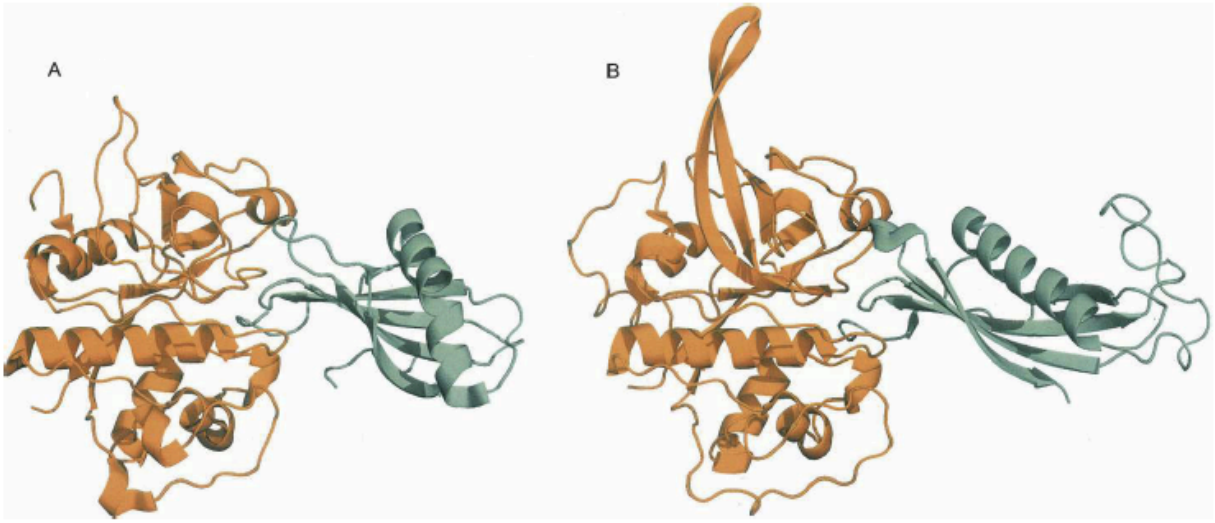


Figure 5.2 Example of a validated prediction: falcipain-2–cystatin-A.
(A) An interaction was predicted between falcipain-2 and cystatin-A based on a template structure of cathepsin-H (orange) bound to cystatin-A (teal) (PDB 1NB3). (B) The structure of falcipain-2 bound to chicken cystatin was recently 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).

We now describe two examples of predicted interactions that have been previously

observed experimentally. We predicted several interactions between proteases and protease inhibitors, the best scoring of which occurred between *P. falciparum* falcipain-2 protease and the human cystatin-A inhibitor based on a template structure of human cathepsin-H bound to cystatin-A (PDB 1NB3) (Figure 5.2). This prediction was recently experimentally validated, with chicken cystatin (PDB 1YVB) (Figure 5.2; Wang et al. 2006). This crystal structure was not present in our template set, because it has not yet been classified by the SCOP domain annotation database (Murzin et al. 1995). Thus, the predicted complex was a true blind prediction. The experimentally determined structure provides direct validation of our prediction, although it does not demonstrate relevance to infection. However, the known involvement of cysteine proteases in malaria pathogenesis and experimentally established cross-talk between host and pathogen protease and inhibitors (Pandey et al. 2006) suggests that the 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 potential relevance as a potential pharmacologic target. Falcipain-2 and cathepsin-H share only 34% sequence identity, beyond the threshold of the sequence-based method required for a reliable prediction of interaction (Yu et al. 2004). However, comparison of the experimental falcipain-2–cystatin structure with the template cathepsin-H–cystatin-A structure reveals a high degree of structural similarity at the interface (C-α RMSD of 0.43 Å). In addition, this structure can be used to search for small-molecules that may disrupt or mimic the target interaction. Falcipain-2 is discussed extensively later in this chapter.

We predicted several interspecies enzyme dimerizations, such as *T. brucei* ornithine decarboxylase (ODC) binding to human ODC. Functional dimerization of

parasitic and host enzyme subunits have been previously observed, such as in *T. brucei* and mouse ODC (Osterman et al. 1994). Although both host and pathogen ODCs have been implicated in viral and protozoal infections (Kierszenbaum et al. 1987; Das Gupta et al. 2005; Singh et al. 2007), the *in vivo* relevance of these homodimer-like complexes is not clear, and thus, we generally removed predictions based on homodimer sequence templates or template structures of subunits classified in the same domain family. This restriction also facilitates visualization and analysis of the networks, although some true positive predictions may be lost.

5.3.4. Specific examples of predicted interactions

We now describe two specific examples of predicted interactions whose indirect support in the literature warrants experimental follow-up. Two additional examples are discussed in Davis 2007,Supplemental material.

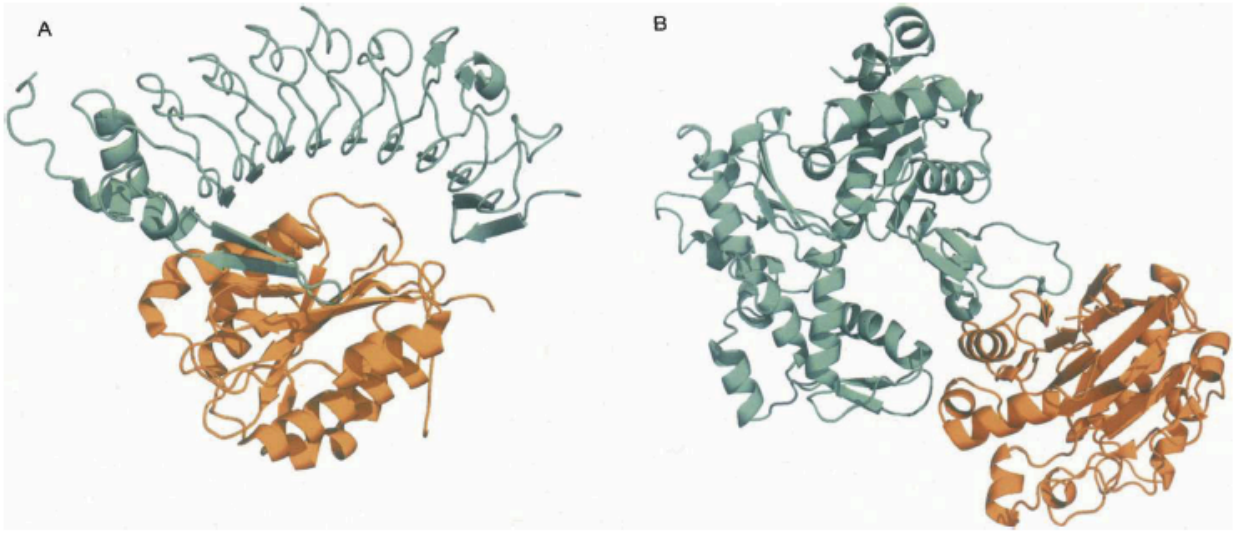


Figure 5.3 Examples of predicted interactions.
(A) *P. falciparum* thrombospondin-related adhesive protein (TRAP) was predicted to interact with human Toll-like receptor 4 (TLR4) based on a structure of glycoprotein IBa (orange) bound to von Willenbrand factor (teal), respectively (PDB 1M10). (B) *M. tuberculosis* probable exported protein Rv0888 was predicted to interact with actin based on a structure of DNAse-I (orange) bound to actin (teal), respectively (PDB 1ATN). Figures were generated by PyMOL (http:// www.pymol.org).

We predicted that *P. falciparum* thrombospondin-related adhesive protein (TRAP, SSP2, PF13_0201) interacts with human Toll-like receptor 4 (TLR4, ENSP00000346893), based on a template structure of Glycoprotein IBa bound to Von Willenbrand factor (PDB 1M10) (Figure 5.3A; Huizinga et al. 2002). TRAP, an immunogenic protein used as a component of several vaccine candidates (Hill 2006), was also predicted to interact with three other leucine-rich repeat proteins; however, the interaction with TLR4 had the most support from the biological filters. Single nucleotide polymorphisms have been observed in TLR4, a ‘‘pattern recognition module’’ involved in the innate immune response. These mutations are associated with an increased severity of malaria, although they fall outside of the region that was modeled here (Mockenhaupt et al. 2006). 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 to interact with TLR4 (Weedall et al. 2007). The possible encounter of these two proteins is also supported by the known expression of TRAP on the parasite surface during the sporozoite stage of the plasmodium life cycle and of TLR4 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 an *in vivo* role in infection.

We predicted that *M. tuberculosis* probable exported protein Rv0888 (GI 15608028) may interact with several human α-actins (ENSP00000295137) based on the template structure of DNAse I bound to actin (PDB 1ATN) (Figure 5.3B; Kabsch et al. 1990). The interaction between DNAse and actin is known to be strong enough to depolymerize actin (Kabsch et al. 1990), and so the predicted interaction could be involved in the

observed *M. tuberculosis* rearrangement of host actin (Guerin and de Chastellier 2000), which has been hypothesized to be triggered by a secreted pathogen factor (Garcia-Perez et al. 2003).

5.3.5. Future developments

The identification of protein–protein interactions is an important problem that has inspired the development of numerous algorithms to predict them (Shoemaker and Panchenko 2007). Several of these methods rely on information such as genomic proximity, gene fission/ fusion, phylogenetic tree similarity, gene co-occurrence, colocalization, 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 (Yu et al. 2004; Davis et al. 2006). Other applicable methods include those that identify peptide motifs (Neduva and Russell 2006) or sequence signatures (Sprinzak and Margalit 2001) that mediate interactions.

Another possible extension of the presented method that may aide in the interpretation of the predictions is an analysis of the genetic polymorphisms at loci encoding for the proposed interacting proteins. If the host gene exhibits polymorphisms associated with infection severity or the pathogen gene exhibits a pattern of polymorphisms suggesting antigenic variation, for example, human TLR4 and *P. falciparum* TRAP (Figure 5.3A), there may be greater reason to believe that the interaction is relevant to infection.

5.3.6. Potential impact

We developed a computational whole-genome method to study potential host–pathogen protein interactions and presented four lines of evidence that suggest it is a valid approach to enrich for these interactions. The method, like any experimental or computational method, has limitations in coverage and accuracy, as we have quantified to the best of our ability. Despite these limitations, our resource is valuable as it is the first attempt to provide large data sets enriched for 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 10 pathogens involved in human infectious diseases. The predictions are available on the Internet (see Section 6.6 for full details) 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 toward the goal of treating infectious diseases, in the ‘‘open source’’ model of the Tropical Disease Initiative, a decentralized, Web-based, community-wide effort where scientists from laboratories, universities, institutes, and corporations work together for a common cause (http://www.tropicaldisease.org) (Maurer et al. 2004). In closing, we expect our method to complement experimental methods in providing insight into the basic biology of host–

pathogen systems, as well as other interspecies relationships that fall elsewhere on the mutualism–parasitism continuum.

5.4. Methods used to predict host-pathogen interactions

The protocol began with the host and pathogen protein sequences: CryptoDB (Heiges et al. 2006), GeneDB (Hertz-Fowler et al. 2004), OrthoMCL-DB (Chen et al. 2006), PlasmoDB (Stoeckert Jr. et al. 2006), ToxoDB (Kissinger et al. 2003), TubercuList (http://genolist.pasteur.fr/TubercuList/) (Table 5.1).

5.4.1. Detecting sequence and structure similarities

First, protein structure models were calculated for all sequences using MODPIPE, our automated software pipeline for large-scale protein structure modeling (Eswar et al. 2003). MODPIPE relies on MODELLER (Sali and Blundell 1993) for its functionality and calculates comparative models for a large number of sequences using different template structures and sequence-structure alignments. Sequence-structure matches are established using a variety of fold-assignment methods, including sequence–sequence (Smith and Waterman 1981), profile–sequence (Altschul et al. 1997) (BUILD_PROFILE, a module for calculating sequence profiles in MODELLER), and profile–profile alignments (Marti-Renom et al. 2004) (PROFILE_SCAN, a module for fold-assignment using profile–profile scanning in MODELLER). 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 (Sali and Blundell 1993). The best scoring model for each alignment is then chosen using a statistical potential (Shen and Sali 2006). 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 (Melo et al. 2002; Eramian et al. 2006; Shen and Sali 2006). Only models that are assessed to have the correct fold were included in the final data sets. The models have been deposited in our database of comparative models, MODBASE (Pieper et al. 2006) (http:// salilab.org/modbase), as publicly accessible data sets.

The detected structural similarities were then used to assign structural domain boundaries to the modeled sequences, according to the SCOP classification system (Murzin et al. 1995), as previously described (Davis et al. 2006). Briefly, domain boundaries were assigned to the target proteins when the putative domain contained at least 70% of the residues in the template domain. If the template-target domain similarity was more than 30% sequence identity, the target domain was classified at the family level of the template's domain classification. If the sequence identity was more than 30% and a reliable model was built or if the sequence identity was more than 30% but MODBASE deemed only a reliable fold assignment, the superfamily was assigned. The remaining target domains received the template domains SCOP classification at the fold level, and were not used in the interaction prediction.

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

Next, pairs of host and pathogen proteins were searched for similarity to known interactions collected in PIBASE (Davis and Sali 2005) and IntAct (Kerrien et al. 2007). PIBASE (release 1.69) is a comprehensive relational database of structurally defined

protein interfaces that currently includes 209,961 structures of interactions between 2613 SCOP domain families. As previously described, these structures were clustered and then filtered to remove potential crystallographic artifacts, resulting in a set of template binary interfaces of 5275 structures (Davis and Sali 2005). IntAct (release 2006-08-18) is an open source database of protein interaction data and contains 63,276 binary protein interactions (Kerrien et al. 2007).

Putative interactions between pairs of host and pathogen proteins that contained domains classified in the same superfamily as those previously observed to interact (PIBASE) were assessed by alignment of their comparative structure models onto the corresponding domains of the template complexes and by subsequent assessment of the putative interface by a statistical potential, as previously described (Davis et al. 2006). Briefly, pairs of residues from the host and pathogen protein models whose side chains occurred within a distance of 8 Å of one another were identified and their scores summed according to a statistical potential derived from binary interface structures in PIBASE. A Z-score was calculated to assess the significance of this raw statistical potential score, by consideration of the mean and standard deviation of the statistical potential scores for 1000 sequences where all amino acid residues in the target domain sequences were shuffled.

The ability of the statistical potential to discriminate a set of 100 true protein interfaces from a background set of 100,000 sequence-randomized decoys was previously assessed using a receiver-operator-curve (ROC) analysis (Davis et al. 2006). This ROC analysis exhibited an area under the curve (AUC) of 0.993 and suggested an optimal statistical potential Z-score threshold of 1.7, which gave true-positive and false-

positive rates of 97% and 3%, respectively. 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) (Kerrien et al. 2007). Host and pathogen sequences were predicted to interact when each aligned to at least 50% of t he sequence of members of a template complex with a joint sequence identity of $(\text{sequence identity}_1 * \text{sequence identity}_2)^{1/2} >= 80\%$ (Yu et al. 2004). This threshold has been previously shown to correctly predict true protein–protein interactions (Yu et al. 2004). 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 (e.g., histones).

5.4.3. Applying biological and network-level filters

The predicted interactions were filtered using biological context and network-level information. The biological context filter was imposed at two levels, individual proteins and their interactions (Davis 2007, Table S3). The host proteins were filtered by expression in tissues known to be targeted by the pathogen (GNF Tissue Atlas [Su et al. 2004], Harrison's Principles of Internal Medicine [Kasper et al. 2004]), known expression on cell surface, and known immune system involvement (ENSEMBL [Hubbard et al. 2007], Gene Ontology Annotation [GOA] [Camon et al. 2004], IRIS [Abbas et al. 2005]). 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 [Stoeckert Jr. et al. 2006], ToxoDB [Kissinger et al. 2003], CryptoDB [Heiges et al. 2006], GeneDB [references in Davis 2007, Table S1] [Hertz-Fowler et al. 2004]). The GO terms for human protein involvement in immune system were GO:0051707, GO:0002376, and 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 one biological criterion 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 (Rachman et al. 2006); *L. major*, pairs of human proteins expressed in skin and pathogen proteins expressed in the promastigote or metacyclic life-cycle stage and 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 and human proteins expressed in liver and pathogen proteins expressed in the sporozoite life-cycle stage; and *Plasmodium vivax*, pairs of human proteins expressed in erythrocyte and pathogen proteins

predicted to be secreted.

The network-level filter removed predictions based on templates 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.
The filtering step was performed after the initial modeling and interaction prediction steps so that the filters could be easily updated to include biological annotation resulting from future experiments, without requiring re-calculation of models and interactions.

5.4.4. Assessment: Intraspecies interactions benchmark

The sequence- and structure-based prediction methods have both been previously benchmarked in the context of intraspecies *S. cerevisiae* protein interactions. For the sequence-based method, all of the interactions transferred from Caenorhabditis *elegans*, *Drosophila melanogaster*, and *Helicobacter pylori* onto *S. cerevisiae* were correct at a joint sequence identity threshold of 80% (Yu et al. 2004). For the structure-based method, 270 of 3387 (8%) predicted *S. cerevisiae* interactions overlapped with experimentally observed interactions, 90% of which exhibited less than 80% sequence identity to the their interaction template (Davis et al. 2006). The use of orthogonal biological information as filters was found to provide a significant (threefold) enrichment of previously observed interactions. The method could not predict the correct specificities in families of homologous receptor-ligand networks, such as the epidermal growth factor receptor and tumor necrosis factor-b network of ligand receptor interactions. In total, 19,424 interactions have been experimentally observed out of the possible 21,776,700 pairs of yeast proteins (0.09%; Jan 2006)

(Davis et al. 2006). Thus, the number of protein pairs was reduced by about four orders of magnitude, while the enrichment was increased by about two orders of magnitude. The analysis suggested that the method was applicable as a first pass for genome-wide predictions of protein complexes.

Pathogen	Host tissues
M. leprae	Skin, lymph node, lung
M. tuberculosis	Lung, bronchial epithelial cells, lymph node
L. major	Skin, whole blood, monocyte (Abbas et al. 2005)
T. brucei	Erythrocyte (Pasini et al. 2006), whole blood, lymph node, brain, endothelial
T. cruzi	Erythrocyte (Pasini et al. 2006), whole blood, lymph node, skeletal muscle, smooth muscle, cardiac myoctes, endothelial
C. hominis	Colorectal adenocarcinoma
C. parvum	Colorectal adenocarcinoma
P. falciparum	Erythrocyte (Pasini et al. 2006), liver, brain, whole blood, endothelial
P. vivax	Erythrocyte (Pasini et al. 2006), liver, whole blood
T. gondii	Lymph node, skeletal muscle, cardiac myoctes, placenta, brain, lung

Table 5.4 Host–tissue filters used for each pathogen.
Host–tissue expression data were obtained from the GNF Tissue Atlas (Su et al. 2004) unless noted otherwise.

5.4.5. Assessment: Functional overview of predicted complexes

The human proteins predicted to interact with pathogen proteins were analyzed for significant enrichment of gene ontology function terms using GO::TermFinder (Boyle et al. 2004). The analysis was done on the interactions before application of the biological filters to prevent introduction of filter bias into the functional profile of the predictions. 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 (Sokal and Rohlf 1995).

5.4.6. Assessment: Comparison to gene expression and essentiality data

Human genes differentially regulated (two-tailed t-test, $P < 0.05$) in macrophages and dendritic cells during infection by *L. major*, *M. tuberculosis*, and *T. gondii* were retrieved from GEO Omnibus (GDS2600) (Edgar et al. 2002; Chaussabel et al. 2003). Lists of *M. tuberculosis* genes essential for in vivo infection (Sassetti and Rubin 2003) and genes that are upregulated in granuloma, pericavity, or distal lung infection sites compared with in vitro conditions (Rachman et al. 2006) were obtained from literature.

5.5. Introduction – The role of the *P. falciparum* falcipain-2 prodomain

Plasmodium falciparum, the most virulent human malaria parasite, is responsible for hundreds of millions of illnesses and about one million deaths each year (1). The control of malaria is hindered by increasing resistance to available drugs, making it important to develop new drugs to treat this disease. Among potential new targets for antimalarial therapy are falcipain cysteine proteases (2). The best characterized of these proteases, falcipain-2 and falcipain-3, play key roles in the hydrolysis of hemoglobin by intraerythrocytic parasites (3–5). Inhibitors of falcipains demonstrate potent *in vivo* antimalarial activity, and these proteases are the targets of efforts to develop novel cysteine protease inhibitors as new antimalarial drugs (2).

Falcipains are cathepsin L-like papain-family cysteine proteases (2). Features shared with other proteases of this sub-family include a 30 kDacatalytic domain with conserved active site amino acid residues and a prodomain with potent enzyme inhibitory activity (6). We have characterized a number of unusual features of falcipains. First, folding of the mature protease is mediated by a fourteen residue N-terminal

extension, rather than the enzyme prodomain (6, 7). Second, a ten amino acid insertion near the C-terminus mediates interaction of the mature domain with its principal substrate, hemoglobin, and with the prodomain (8). Third, the prodomain does not have a typical signal sequence, but contains a membrane-spanning domain that predicts a type II integral membrane protein. Fourth, the falcipain prodomain is much larger than that of most other described papain-family proteases, with downstream sequence similar to papain and related enzymes, but unique upstream regions that mediate trafficking of falcipain-2 to the food vacuole, the site of hydrolysis of hemoglobin (9).

Considering its importance as a potential drug target, we were interested in evaluating the features of the falcipain-2 prodomain that mediate enzyme inhibition. We hypothesized that the inhibitory function is mediated by the downstream portion of the prodomain, which has an amino acid sequence similar to that of other papain family proteases. In this region, cathepsin L-like papain family proteases, including falcipains, contain a number of conserved residues that appear to mediate interaction between the prodomain and mature protease (10), including six amino acids (ERFNIN in papain) spanning nineteen residues (11, 12) and, further downstream, four conserved amino acids (GNFD in papain) spanning seven residues (13). Conservative substitutions at these motifs are common; the sequences are ERWNIN and ANFD in cathepsin L and DRWNIN and ANLD in cathepsin K. In cathepsin L, these residues appear to stabilize the prodomain structure through the formation of salt bridges (14). To determine the roles of these conserved amino acids and other portions of the falcipain-2 prodomain in enzyme inhibition, we expressed the prodomain and a series of truncated fragments, and evaluated their inhibitory activity (15). Our results define a 61 residue minimum

inhibitory domain, which includes the ERFNIN and GNFD motifs, that strongly inhibits falcipain-2 and many other cysteine proteases. Modeling of the falcipain-2 prodomain suggests that the prodomain covers the enzyme active site, and thereby inhibits activity by preventing substrate access.

5.6. Results – Characterization of prodomain inhibition

5.6.1. Identification of the inhibitory domain of falcipain-2

Falcipain-2 and homologs from related plasmodia have much larger prodomains than those of most papain-family proteases. The upstream portion of the falcipain-2 prodomain bears no obvious resemblance to sequences of non-plasmodial proteases, and mediates enzyme trafficking to the parasite food vacuole (9). In contrast, the downstream portion of the falcipain-2 prodomain is similar to that of papain, and in particular to the cathepsin L sub-family of papain-family proteases (Figure 5.4). The sequence identity for this region between falcipain-2 and human cathepsin L is 21%, and residues that have been identified as playing key roles in the functions of papain family prodomains are generally conserved in falcipain-2 and plasmodial homologs. The well characterized ERFNIN and GNFD domains (10), which contribute to proenzyme stability, are both fully conserved in falcipain-3, but falcipain-2 differs from the consensus sequence at one ERFNIN (I→V) and one GNFD (G→E) residue. Two highly conserved Trp residues (at positions 19 and 22 of procathepsin L), which also contribute to the stability of cathepsin L sub-family proteases (12), are each replaced by Phe in both falcipain-2 and falcipain-3 (Figure 5.4; falcipain-2 positions 165 and 168).

We previously showed that the prodomain of falcipain-2 is a potent reversible inhibitor of the protease (6). To characterize the requirements for inhibition, we

expressed a series of prodomain fragments in E. coli (Pandey 2009, Figure S1) and evaluated inhibition of falcipain-2 by each of the fragments (Figure 5.5). All peptides were soluble in the buffers used for our experiments and stable under our experimental conditions. As we hypothesized, the large upstream portion of the prodomain, which includes a transmembrane domain flanked by cytosolic and lumenal segments, and which mediates trafficking of falcipain-2 to the food vacuole (9), is not required for enzyme inhibition. Inhibitory potency was the same for a prodomain construct lacking only the upstream cytosolic and transmembrane domains (Tyr_{54}-Asp_{243}) and for constructs lacking the upstream 104 (Ser_{105}-Asp_{243}), 126 (Leu_{127}- Asp_{243}), or 154 (Leu_{155}-Asp_{243}) amino acids of the prodomain (Figure 5.5). All of these constructs were very potent inhibitors of falcipain-2, with K_i < 1 nM. The removal of the 27 C-terminal amino acids of the prodomain (Tyr_{54}-Asp_{216}) did not affect inhibitory potency, but removal of the 37 C-terminal amino acids (Tyr_{54}- Leu_{206}) led to a ~2000-fold loss of inhibitory potency, and removal of the 63 C-terminal amino acids (Tyr_{54}-Asn_{180}) led to a complete loss of inhibitory activity. A peptide spanning the ERFNIN and GNFD motifs (Tyr_{176}-Asp_{216}) demonstrated no inhibitory activity. These results allow identification of a minimum inhibitory domain for falcipain-2 (Leu_{155}-Asp_{216}), which includes two hydrophobic residues (Phe_{165} and Phe_{168} in falcipain-2; Phe_{182} and Phe_{185} in falcipain-3) and the ERFNIN and GNFD motifs, all of which are highly conserved among other cathepsin L sub-family proteases. We could not directly test the inhibitory activity of this minimum inhibitory peptide, as production of the recombinant peptide was unsuccessful.

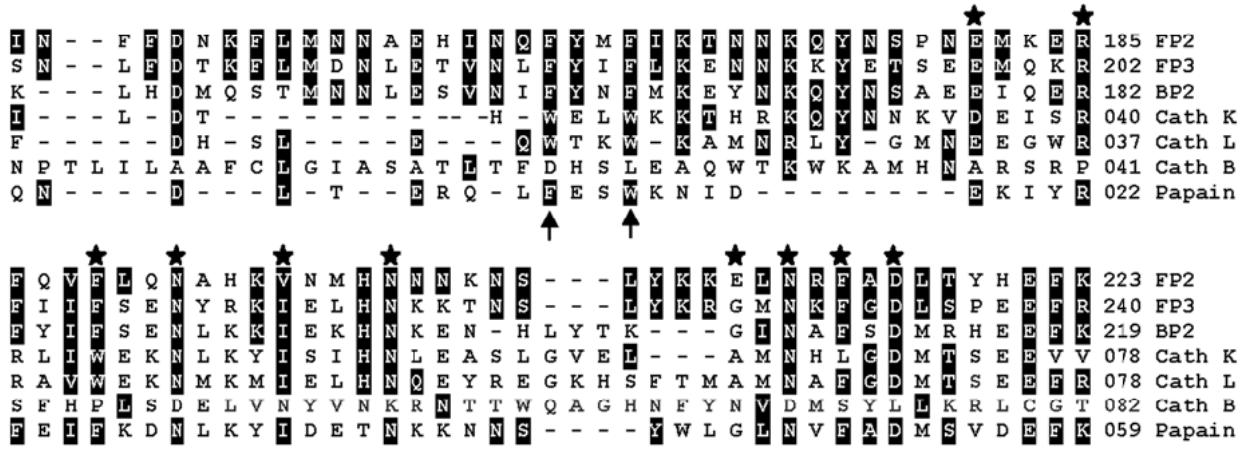


Figure 5.4 Alignment of C-terminal amino acid residues of the prodomains of falcipain-2 and related cysteine proteases.
The sequences of falcipain-2 (FP2), falcipain-3 (FP3), berghepain-2 (BP2), human cathepsin K (Cath K), human cathepsin L (Cath L), human cathepsin B (Cath B), and papain were aligned using Expassy (European Bioinformatics Institute). Amino acids comprising the ERFNIN and GNFD motifs are labeled with stars, and conserved hydrophobic residues are indicated by arrows. Amino acids that are identical or similar to those of falcipain-2 are highlighted.

5.6.2. Inhibitory Activity of the Falcipain-2 Prodomain Against Other Cysteine Proteases

Cathepsin L sub-family protease prodomains generally inhibit only closely related proteases. For example, the prodomains of cathepsin L, cathepsin K, and cathepsin S are each potent inhibitors of all three proteases, but not of cathepsin B (10). In contrast, the falcipain-2 prodomain had a rather broad inhibitory specificity, with inhibition of the falcipain-2 homolog from *Plasmodium berghei* (berghepain-2), the *Trypanosoma cruzi* protease cruzain, cathepsin L, and cathepsin B (Figure 5.6). The only tested papain-family cysteine protease that was not inhibited was the dipeptidyl peptidase cathepsin C. The aspartic protease pepsin, serine protease α-chymotrypsin, and metalloprotease collagenase were not inhibited by the falcipain-2 prodomain.

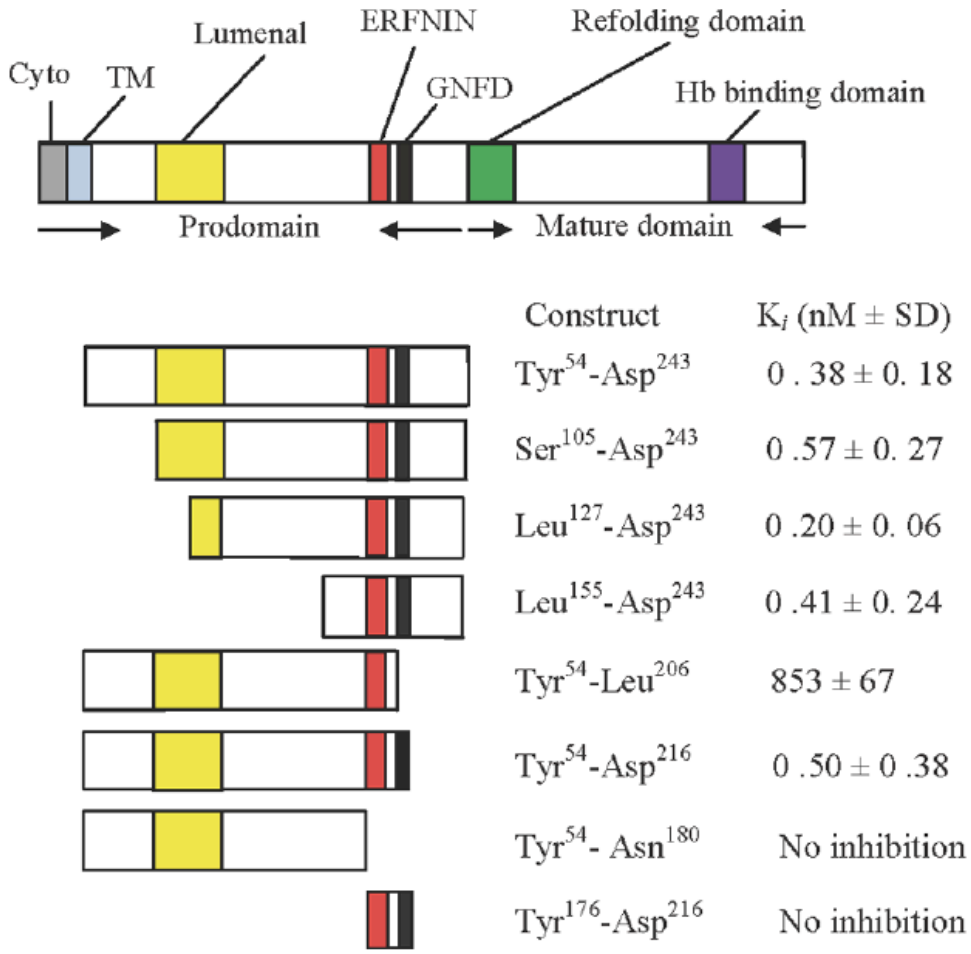


Figure 5.5 Inhibitory activity of profalcipain-2 constructs.
The domains of falcipain-2 and the studied constructs are represented diagrammatically. Abbreviations: Cyto, cytosolic domain; TM, transmembrane domain; Hb, hemoglobin. The residues contained in each construct are shown, and the inhibitory capacity of mature falcipain-2 for each construct is indicated. The data provided are the K_i values for each polypeptide construct. Results are from two experiments, each performed in duplicate.

5.6.3. Structural Explanation for Inhibitory Activity of Falcipain- 2 Prodomain Fragments

Structure-function studies identified a discrete portion of the falcipain-2 prodomain required for inhibition of the cognate mature protease. Prior work with other cathepsin L sub-family proteases suggests key roles for conserved hydrophobic amino acids as well as the ERFNIN and GNFD motifs in maintaining prodomain structure (10). We explored the roles of different domains in maintaining prodomain structure by circular dichroism

analysis (Figure 5.7). Secondary structure was seen in a fragment with potent inhibitory activity (Leu_{155}-Asp_{243}), but not in two larger constructs that lacked any sequence downstream of the ERFNIN and GNFD motifs (Tyr_{54}-Leu_{206}; Tyr_{54}-Asn_{180}) or in a peptide spanning the ERFNIN and GNFD motifs (Tyr_{176}-Asp_{216}). These results indicate that the ERFNIN and GNFD motifs and an upstream region including conserved Phe residues are required for proper folding or maintenance of secondary structure of the prodomain.

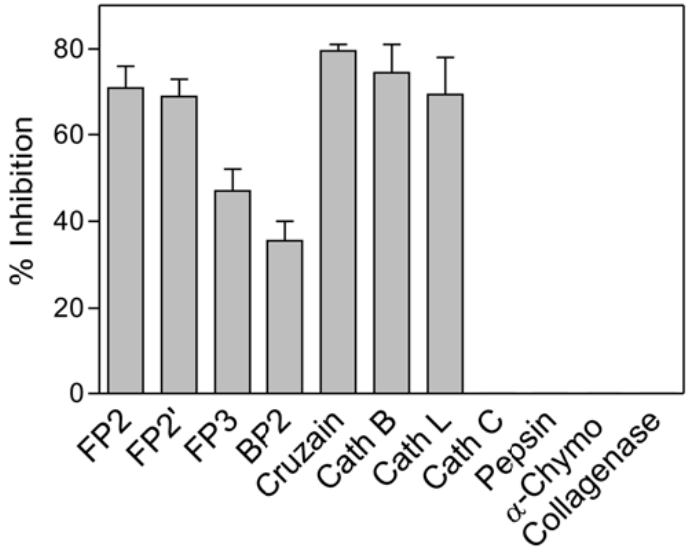


Figure 5.6 Inhibition of different proteases by the prodomain of falcipain-2.
The inhibition of falcipain-2 (FP2), falcipain-29 (FP29), falcipain-3 (FP3), berghepain-2 (BP2), cruzain, human cathepsin B (Cath B), human cathepsin L (Cath L), bovine cathepsin C (Cath C), pepsin, α-chymotrypsin (α-Chymo), and collagenase was measured as described in Methods. In each case, activity was measured with and without the prodomain and the percentage inhibition calculated. Error bars represent standard deviations from two experiments, each performed in duplicate.

5.6.4. Homology Modeling of Profalcipain-2

To explain the role of profalcipain-2 motifs in enzyme inhibition, we modeled the structure of the target falcipain-2 using the crystallographic structures of several papain-family cysteine proteases as templates. We used the software MODELLER-9v4 (16) to construct a homology model of profalcipain-2 Figure 5.8, which aligned to mature falcipain-2, procathepsin L, procathepsin K, and procaricain at sequence identities of

100% (by definition, aligned with the sequence of the mature domain only), 30.6%, 30.9%, and 32.1% respectively (14, 20–24). The model was evaluated with DOPE (Discrete Optimized Protein Energy), a pairwise atomic distance statistical potential that assesses atomic distances in a model relative to those observed in many known protein structures (17). The DOPE Z-score of the model (-0.99) is similar to the Z-scores of all templates (cathepsin L: -1.62; mature falcipain-2: -1.13; procathepsin K -0.95; procaricain -1.25); generally, a Z-score of -1 or less indicates a relatively accurate model, with more than 80% of its C-α atoms within 3.5 Å of their correct positions (17). Additionally, a separate assessment technique, TSVMod, was applied. This method predicts the native overlap (defined as the fraction of a-carbon atoms within 3.5 Å of the native structure) of a homology model in the absence of a solved structure using support vector machine learning (18, 19). The model's predicted native overlap (0.85) was similar to that of a model of mature falcipain-2 built using the mature sections of the above templates, indicating the falcipain-2 prodomain does not contribute significantly disproportionately to the overall model error. This assessment suggests that the fold of the profalcipain-2 model is correct despite the relatively low sequence identity between the falcipain-2 prodomain and the templates.

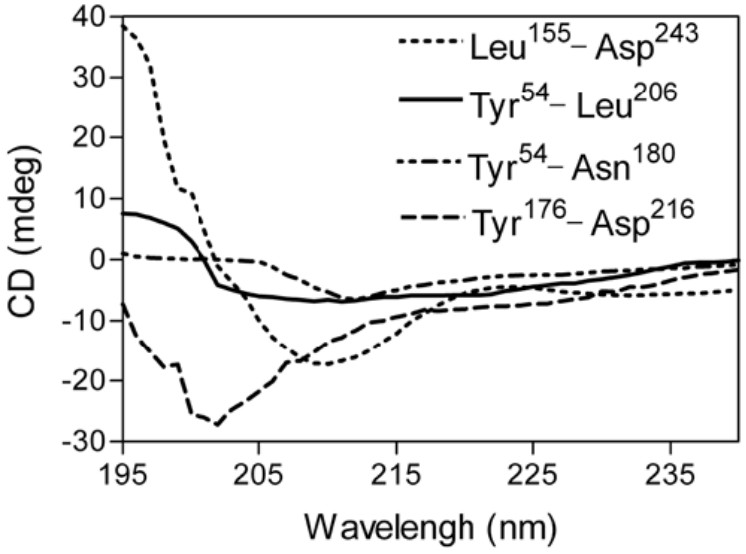


Figure 5.7 Circular dichroism analysis of prodomain constructs.
Different falcipain-2 prodomain constructs (200 μg/ml) were incubated in 20 mM sodium phosphate, pH 5.8, and absorbance between 195 and 240 nm was measured.

5.6.5. The Profalcipain-2 Model Suggests that the Conserved Residues Provide Stability to the Overall Fold

We examined the homology model for possible interactions involving residues in the conserved motifs. Several of these residues are highlighted in Figure 5.8b. (i) The charged pair Arg_{185} and Glu_{221} appears to form a salt bridge. (ii) Glu_{210} from the GNFD motif may form a separate salt bridge with Lys_{403} in the mature domain. (iii) Phe_{214} may participate in non-polar interactions, and possibly Π-bond stacking, with two tryptophan residues on the mature domain, Trp_{449} and Trp_{453}. All of these interactions are also present in at least one of the templates used to build the model, although none of them is conserved across all templates.

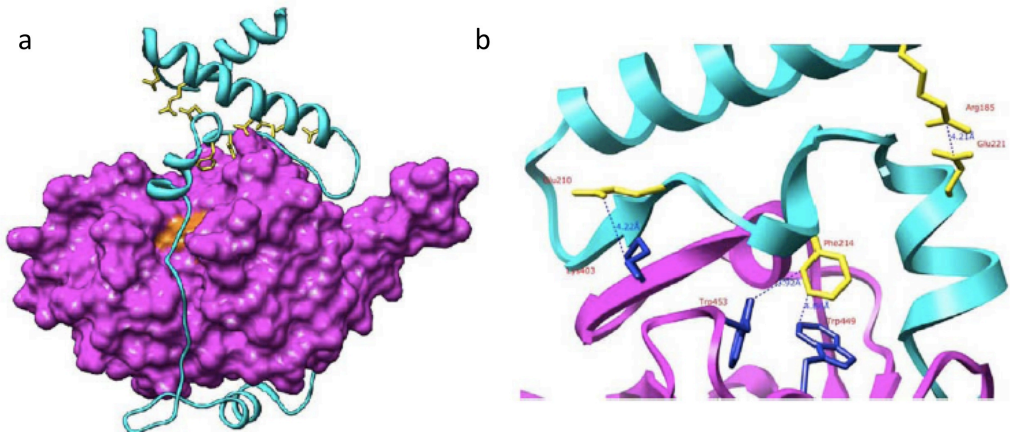


Figure 5.8 Homology model of profalcipain-2.
(a) Model created using MODELLER 9v4. The 160 N-terminal residues of the prodomain are not included in the model. The prodomain (cyan) runs up the face of the mature enzyme (purple; catalytic triad residues in orange) before forming α-helices containing the conserved ERFNIN and GNFD motifs (yellow). (b) Close-up of several predicted interactions between the mature protease and the ERFNIN (R_{185}) and GNFD (E_{210}; F_{214}) motifs. Blue dashed lines indicate presumed stabilizing interactions (both electrostatic and hydrophobic) between residues. The structure has been rotated 180° around the vertical axis from its representation in (a)

5.6.6. The Falcipain-2 Prodomain Appears to Block Substrates from Entering the Cathepsin B Active Site

A separate homology model was constructed in which the falcipain-2 prodomain and cathepsin-B mature domain were modeled as a complex (Figure 5.9a). The model was built based on an alignment of profalcipain-2 at 31.2% sequence identity with the crystallographic structure of procathepsin B (25, 26). The model received a DOPE Z score of -0.87, and a TSVMod native overlap prediction of 0.82. These scores indicate that the overall fold is correct; poor scores would have suggested that there were significant errors in the modeled structure of the prodomain, and in that case the model would not have resembled the structures of the templates on which it was based. The model suggests that the prodomain of falcipain-2 binds mature cathepsin B in a manner similar to that observed in papain family zymogens, inhibiting catalytic activity by blocking substrate access to the active site. (Figure 5.9b). While no structure has been

solved for a propeptide in complex with an inhibited mature enzyme, it is likely that these propeptides bind to the enzymes in a conformation resembling the zymogen form (14, 25–27). This hypothesis is reflected in the model, which by construction is similar to its templates, and displays favorable stereochemistry and non-bonded atom distances as evaluated by MODELLER and DOPE.

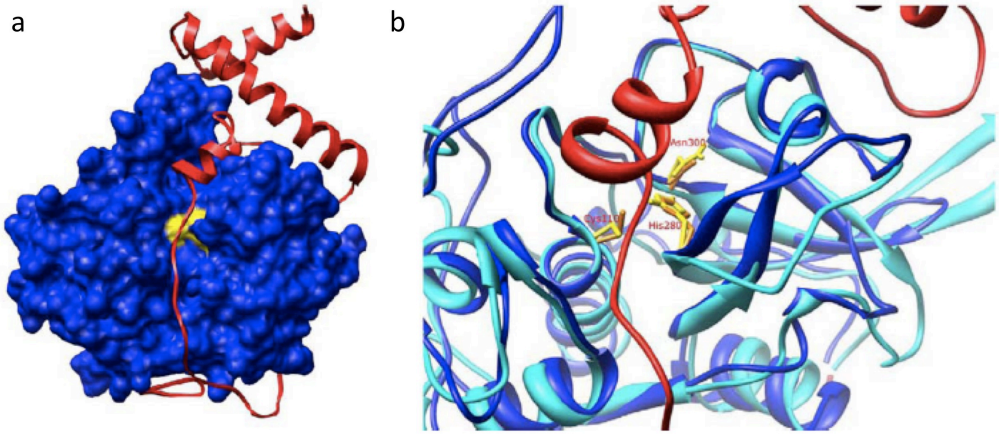


Figure 5.9 Model rationalizing the inhibition of cathepsin B by FP2 prodomain.
(a) Model of the falcipain-2 prodomain (red) and mature cathepsin B (blue; catalytic triad residues in yellow). The prodomain binds to cathepsin B in a similar fashion as zymogens of other cysteine proteases, including procathepsin L and procathepsin B. (b) Structural overlay of mature cathepsin B (blue) and falcipain-2 (cyan). Catalytic triad residues are shown in the stick representation (yellow: cathepsin B; orange: falcipain-2). Cathepsin B amino acid numbering is used.

5.6.7. Differences Between the Prodomains of Falcipain-2 and Cathepsin L

Cathepsin B activity is inhibited by the prodomain of falcipain-2 (Figure 5.6) but not cathepsin L (10). To examine the structural basis of this selectivity, we compared the sequences and structures of these two proteins. Several differences were of note (Figure 5.10). First, while the procathepsin L α1 helix clashes with the occluding loop region of mature cathepsin B, thus preventing binding, the equivalent helix in falcipain-2 does not. Second, Phe≠ in profalcipain-2 participates in polar interactions with Phe_{165} and Phe_{168}; in procathepsin K and procathepsin L, Phe_{186} is replaced by Arg. Third, a multiple sequence alignment reveals a conserved motif (LMNNAEHIN in falcipain-2) in

the plasmodial proteases falcipain-2, falcipain-3, and berghepain-2 that represents an insertion relative to the sequences of procathepsin K and procathepsin L (Figure 5.4). Finally, an apparent salt bridge (interaction not shown) is formed between Glu_{210} in the falcipain-2 prodomain and Lys_{184} in mature cathepsin B; Glu_{210} of falcipain-2 (which has replaced Gly in the GNFD motif) is replaced by Ala in cathepsin L and cathepsin K. Taken together, differences between modeled interactions for the cathepsin B mature domain with procathepsin L or profalcipain-2 appear to describe the structural basis for the observed selective inhibition of cathepsin B activity by profalcipain-2.

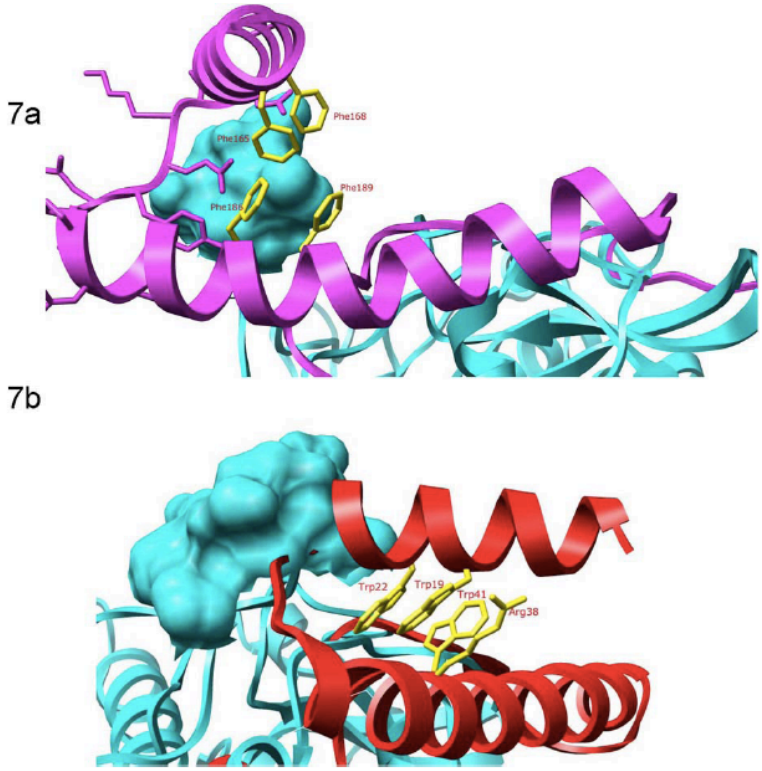


Figure 5.10 Modeled differences between falcipain-2 (a) and cathepsin L (b) prodomain binding to cathepsin B.
The model predicts a helix arrangement in the falcipain-2 prodomain (purple) that prevents steric clashes with the cathepsin B occluding loop (cyan). Phe_{186} may mediate this arrangement; in cathepsin K and cathepsin L, Phe_{186} is replaced by Arg. For cathepsin L (red), there is a large steric clash between the linker joining the two cathepsin L helices and the space-filled occluding loop (cyan).

5.7. Discussion – A structural model for prodomain inhibition specificity

We evaluated features of the falcipain-2 prodomain that mediate enzyme inhibition. Our data show that only an 11 kDa C-terminal region of the prodomain is required for potent inhibition of the protease. The region includes two hydrophobic residues (both Phe in falcipain-2) and the ERFNIN and GNFD motifs, all of which are conserved among cathepsin L- like papain family proteases. The falcipain-2 prodomain also inhibited other papain family cysteine proteases, including similar cathepsin L sub-family proteases and the more distantly related cathepsin B. We explored the relevance of conserved falcipain-2 motifs by circular dichroism; the conserved residues were required to maintain the secondary structure of the prodomain. Thus, the first prerequisite for inhibitory activity was appropriate secondary structure. We also constructed a homology model of profalcipain-2 to help explain the observed experimental results. The model identified potential interactions between the inhibitory portion of the prodomain and mature falcipain-2 that appear to explain the inhibitory activity, and also the ability of the prodomain of falcipain-2, but not that of the related protease cathepsin L, to inhibit cathepsin B. Taken together, our results identify and structurally characterize a minimum inhibitory domain of the falcipain-2 prodomain, offering a starting point for new considerations for the inhibition of key proteases of malaria parasites. Indeed, small molecules that inhibit falicipains via interactions independent of the active site might offer highly specific antimalarials without detrimental effects due to inhibition of host cysteine proteases.

Results of structure-function studies were straightforward. As expected, the upstream portion of the falcipain-2 prodomain, which mediates protein trafficking (9),

was not required for inhibitory activity. Indeed, only a small portion of the prodomain (Leu_{155}-Asp_{216}) was required for sub-nanomolar inhibition of the mature enzyme. We did not demonstrate inhibition by the isolated Leu_{155}-Asp_{216} peptide, as production of this peptide proved difficult, but consideration of inhibition by a number of overlapping constructs (Figure 5.5) clearly demonstrates that this peptide is sufficient for inhibition of falcipain-2. Circular dichroism studies suggested that the limits of the minimum inhibitory domain are dictated by requirements for appropriate folding and maintenance of a secondary structure for the inhibitory portion of the prodomain. Due to the conserved overall fold of cathepsin precursors (10), along with the high degree of structural similarity between these proteases and mature falcipain-2 (Cα root-mean-square-deviation between falcipain-2 and cathepsin K is 0.92 Å; falcipain-2 and cathepsin L is 0.81Å; and falcipain-2 and procaricain is 0.95 Å), profalcipain-2 is a good candidate for comparative modeling analysis. Our model has a good DOPE score, a pairwise atomic distance statistical potential that has been shown to perform well in evaluating errors in homology models (17). DOPE is particularly suited to determine the accuracy of the overall fold of a model. The DOPE score of the model of falcipain-2 was similar to those of mature falcipain-2 and procathepsin L, indicating that the overall fold of our homology model is accurate. A separate model assessment program, TSVMod, gave essentially the same results.

In our model, residues in the ERFNIN and GNFD motifs were involved in several interactions important to the stability of the falcipain-2 prodomain fold (Figure 5.8b). Two interactions, Arg_{185}– Glu_{221} and Phe_{214}–Trp_{449}/Trp_{453}, appear to be conserved between falcipain-2 and cathepsin L, with equivalent residues present in procathepsin L (21). A

third interaction, Glu_{210}–Lys_{403}, represents a unique charged pair interaction, as a Glu is found in falcipain-2, but not falcipain-3 or most related proteases, replacing the Gly in the GNFD motif. Side chain packing is the most difficult part of comparative modeling; however, in this case using the ERFNIN and GNFD motifs as well as the conserved Phe residues to guide the alignment resulted in conserved sequences across the downstream region of the prodomain (Figure 5.4), increasing confidence in our predictions. Many cathepsin L sub-family propeptides act in trans to inhibit related proteases (10). However, selectivity has been observed, and it has been demonstrated that the prodomains of cathepsin L and cathepsin K are unable to inhibit cathepsin B (25–27). Explanations for this observation include the following. First, cathepsin B lacks the ERFNIN motif, so that the protease lacks most of the α2 helix found in cathepsin L sub-family proteases. Second, cathepsin B contains a large occluding loop insertion, conferring dipeptidase activity, but preventing propeptides containing the ERFNIN motif from binding due to a steric clash between the occluding loop and the prodomain residues connecting α1 and α2 (Figure 5.10b). Interestingly, selectivity for prodomain inhibition was broader for falcipain-2, as the prodomain of falcipain-2 markedly inhibited cathepsin B (Figure 5.6). Our homology model adds insight to this observation.

In the model, the interaction of the helices equivalent to cathepsin L helices α1 and α2 is shifted (Figure 5.10a). This shift is mediated by the presence of an additional aromatic residue in falcipain-2, Phe_{186}. This residue is part of the hydrophobic core of aromatic residues that contributes to the helix interaction in cathepsin L and cathepsin K, normally mediated by two Trp residues on α1 and the Phe residue in the ERFNIN motif on α2. In falcipain-2, Phe_{186} provides additional stability, allowing α1 to shift

across α2 and eliminating the steric overlap between the prodomain residues and the cathepsin B occluding loop. In procathepsin L and procathepsin K, which do not inhibit cathepsin B, Phe_{186} is replaced by Arg_{38} (procathepsin L) and Arg_{41} (procathepsin K); arginine is a basic residue that interacts less favorably with the other hydrophobic residues. (Figure 5.10b).

A recent study indicated that a synthetic fifteen residue peptide (Leu_{155}-Ile_{169}) from a region of the falcipain-2 prodomain immediately upstream of conserved Phe residues (Phe_{165} and Phe_{168}) inhibited falcipain-2 (28). The authors proposed that this segment plays an important role in inhibition of falcipain-2. However, inhibition by the peptide was at much lower (10^4 times less) potency than inhibition by our prodomain constructs, which acted at sub-nanomolar concentrations. In our model, the Leu_{155}-Ile_{169} residues form the α1-helix. As noted, these residues represent an insertion relative to cathepsin L and cathepsin K. The α1 helix does not appear to actively inhibit falcipain-2, but rather appears to provide structural stability through an interaction with the α2 helix. It is thus likely that the full prodomain inhibits falcipain-2 differently from the small peptide studied recently (28), as for this peptide to come within the proximity of the falcipain-2 active site would require replacement of the α3 helix and a novel fold relative to other papain-family proteases.

Our work defines the minimum inhibitory region of the falcipain-2 prodomain. We show that several residues conserved across cathepsin L sub-family proteases are necessary for this inhibition, and present a structural model for the interaction of the falcipain-2 prodomain with both its own mature domain and that of other proteases. As natural inhibitors of parasite protease activity, propeptides present a promising basis for

design of small molecules to treat malaria.

5.8. Methods used to characterize falcipain-2 prodomain inhibition

5.8.1. Reagents

Benzyloxycarbonyl-Leu-Arg-7-amino-4 methyl coumarin (Z-Leu-Arg-AMC) and Z-Phe-Arg-AMC were from Peptides International. Restriction endonucleases and polymerases were from New England Biolabs. Oligonucleotides were synthesized at the Biomolecular Resource Center, University of California, San Francisco, and by Integrated DNA Technologies. The synthetic peptide was from AnaSpec. All other reagents were from Sigma-Aldrich or as mentioned in the text.

5.8.2. PCR and Sequencing

All DNA fragments were amplified from the pTOP-FP2 plasmid, which encodes the falcipain-2 gene (6). The sequence of each construct was confirmed by DNA sequencing at the Biomolecular Resource Center, University of California, San Francisco. Portions of the falcipain-2 gene were amplified using primers specific for each construct (Pandey 2009, Table S1).

5.8.3. Cloning, Expression, and Refolding of Different Prodomain Constructs

Amplified DNA fragments were digested with *Bam*HI and *Hin*dIII, ligated into digested plasmids (pRSET-B; Invitrogen) and used to transform AD (DE3) pLys *E. col*i (Invitrogen). Cells were induced with β-D-thio-galactopyranoside, and recombinant proteins were solubilized in 8 M urea, 20 mM Tris-Cl, pH 8.0 at room temperature for 60 min with gentle shaking. Insoluble material was separated by centrifugation at 27,000 g for 30 min at 4C. For the purification of the recombinant protein, the supernatant was

incubated with nickel-nitrilotriacetic acid resin (Ni-NTA; Qiagen) and purified under denaturing conditions, as previously described (6). Ni-NTA purified propeptides were bound to SP-sepharose columns (Amersham Bioscience) and eluted by a step-wise gradient of 0-1 M NaCl in 8 M urea, 20 mM Tris-Cl, pH 8.0. The denatured proteins were diluted 100-fold (final concentration 20 μg/ml) in 100 mM Tris-Cl, 1 mM EDTA, 250 mM L-arginine pH 9.0, refolded at 10–12C for 20 h, and concentrated using a 10 kDa cut-off membrane (Millipore) to 10 ml. Insoluble protein was removed using a 0.45 μm syringe filter (Millipore).

5.8.4. Inhibition of Falcipain-2 by the Prodomain

Inhibitor kinetics were calculated as previously described (15). In brief, different concentrations of prodomain constructs (2– 50 nM) were pre-incubated with 2 nM falcipain-2 in 100 mM sodium acetate, 5 mM DTT, pH 5.5 for 10 min at room temperature. The substrate Z-Leu-Arg-AMC (10 μM) was added, and fluorescence (excitation 355 nm; emission 460 nm) was continuously measured for 20 min at room temperature with a Labsystems Fluroskan Ascent spectrofluorometer. Enzyme concentration was determined by titration with the irreversible inhibitor morpholine urea-phenylanine-homophenylanine fluoromethyl ketone. K_i values were determined by nonlinear regression analysis using PRISM (GraphPad Software).

5.8.5. Inhibition of other Proteases by the Falcipain-2 Prodomain

Substrates were Z-Leu-Arg-AMC (10 μM) for falcipain-2, falcipain-3, and cruzain; Z-Phe-Arg-AMC (10 μM) for cathepsin L and cathepsin K; Z-Arg-Arg-AMC (10 μM) for cathepsin B; Pro-Arg-AMC (10 μM) for cathepsin C; and FITC-casein (8 μg / μl) for the other studied proteases. For each reaction, 1 mg of purified falcipain-2 prodomain (or,

for controls, no prodomain) and 2-10 nM of each enzyme were incubated for 10 min in 350 ml of 100 mM sodium acetate, 5 mM DTT, pH 5.5 (for α-chymotrypsin and collagenase 10 mM Tris, pH 7.5), substrate was added, and substrate hydrolysis was monitored as described above or, for FITC-casein, as previously described (8).

5.8.6. Circular Dichroism

Experiments were performed on a Jasco J-175 spectropolarimeter. Signals were monitored between 195 and 300 nm in 20 mM sodium phosphate, pH 5.8 at 20C. Purified proteins were concentrated (200 μg/μl) using a 10-kDa cutoff Amicon ultraconcentrater (Millipore) and transferred to the phosphate buffer. All experiments were performed in a quartz cell of 1 cm path length (Hellma).

5.8.7. Falcipain-2 Modeling

Falcipain-2 residues 161–484, encompassing the full mature domain and the C-terminal region of the prodomain, were aligned with procathepsin L, procathepsin K, and procaricain, at sequence identities of 20–25% in the prodomain region. 100 homology models were built based on the crystallographic structures of these proteins as templates (PDB codes were 1CS8, 1BY8, and 1PCI, respectively) and the crystallographic structure of mature falcipain- 2 (1YVB), using the standard ‘automodel’ routine of MODELLER-9v4 (16). Models were evaluated with the Z-DOPE statistical potential (17) and the TSVMod protocol for predicting absolute model error (18). The model receiving the best Z-DOPE score was subjected to loop refinement of residues 15–20 (sequence NKQYNS), restraining the first 14 residues to a helical conformation, using the ‘loop’ routine of MODELLER-9v4 (19).

5.8.8. Cathepsin-B Modeling

The prodomain of falcipain-2 was modeled in complex with the crystallographic structure of mature cathepsin B. The same homology modeling and loop modeling procedures were performed as for falcipain-2, here based on the crystallographic structures of the prodomain regions of procathepsin L, procathepsin K, and procaricain, and the solved structure of procathepsin B (PDB code 3PBH), as templates. Structural alignments of procathepsin L and cathepsin B were performed with the SALIGN command of MODELLER-9v4 (20)..

Chapter 6. Resources associated with this dissertation

6.1. PCSS WebServer

The algorithm presented in Chapter 2 was converted into a publically available web server titled “Peptide classification based on sequence and structure” (PCSS). It is available at www.salilab.org/pcss. The server trains on a user-defined input set of positive and negative peptides to build an SVM model for scoring peptides to be evaluated, which can be uploaded by the user in a separate step. The model is based on the sequence and structure features described in the algorithm. The server also outputs the Receiver-Operator Characteristic curves to indicate the discriminatory ability of the model for the input dataset.

6.2. GrBah dataset of granzyme B substrates

We compiled a dataset of all experimentally verified Granzyme B substrates, describing the protein name and identifier, cleavage sequence and location, the type of experiment

used to define the site, and the publication that conducted the study. This dataset, which we refer to as GrBah, is available as supplemental material in (Barkan 2010).

6.3. Predicted protease cleavage sites

Proteome-wide predictions of Granzyme B and caspase cleavage sites generated in Chapter 2 are available at www.salilab.org/pcss

6.4. Atomic Domino module

The peptide docking method, which incorporates the atomic DOMINO procedure described in Chapter 3, is available as a module as part of the Integrated Modeling Platform (IMP; http://www.integrativemodeling.org/).

6.5. Mass spectrometry datasets

All experimentally generated results described in Chapter 4 are available as supplemental data in their respective publications.

6.6. Host Pathogen predictions

The ModTie algorithm for predicting large-scale host-pathogen interactions is available at http://pibase.janelia.org/modtie/. Predictions made as part of the study described in Chapter 5 are available at http://salilab.org/hostpathogen/. The sequence-based algorithm for making predictions described in the same chapter is available upon request.

6.7. Falcipain 2 model

The comparative models generated as part of the Falcipain-2 study in Chapter 5 are available upon request.

Chapter 7. References

1. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(1): p. 13-20.
2. Petsalaki, E. and R.B. Russell, *Peptide-mediated interactions in biological systems: new discoveries and applications.* Curr Opin Biotechnol, 2008. **19**(4): p. 344-50.
3. Stein, A., R. Mosca, and P. Aloy, *Three-dimensional modeling of protein interactions and complexes is going 'omics.* Current opinion in structural biology, 2011. **21**(2): p. 200-208.
4. Mas, J.M., et al., *Protein similarities beyond disulphide bridge topology.* Journal of molecular biology, 1998. **284**(3): p. 541-548.
5. Stein, A. and P. Aloy, *Contextual specificity in peptide-mediated protein interactions.* PLoS ONE, 2008. **3**(7): p. e2524.
6. Dyson, H.J. and P.E. Wright, *Intrinsically unstructured proteins and their functions.* Nat Rev Mol Cell Biol, 2005. **6**(3): p. 197-208.
7. London, N., D. Movshovitz-Attias, and O. Schueler-Furman, *The structural basis of peptide-protein binding strategies.* Structure, 2010. **18**(2): p. 188-99.
8. Gough, N.R., *Science's signal transduction knowledge environment: the connections maps database.* Annals of the New York Academy of Sciences, 2002. **971**: p. 585-587.
9. Gould, C.M., et al., *ELM: the status of the 2010 eukaryotic linear motif resource.* Nucleic acids research, 2009. **38**(Database): p. D167-D180.
10. Haynes, C., et al., *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.* PLoS computational biology, 2006. **2**(8): p. e100.
11. Russell, R.B. and T.J. Gibson, *A careful disorderliness in the proteome: sites for interaction and targets for future therapies.* FEBS letters, 2008. **582**(8): p. 1271-1275.
12. Alber, F., et al., *Determining the architectures of macromolecular assemblies.* 2007. **450**(7170): p. 683-694.
13. Ceol, A., et al., *DOMINO: a database of domain-peptide interactions.* Nucleic acids research, 2007. **35**(Database): p. D557-D560.
14. Vanhee, P., et al., *PepX: a structural database of non-redundant protein-peptide complexes.* Nucleic acids research, 2010. **38**(Database issue): p. D545-51.
15. Ahmad, M., W. Gu, and V. Helms, *Mechanism of fast peptide recognition by SH3 domains.* Angew Chem Int Ed Engl, 2008. **47**(40): p. 7626-30.
16. Jemth, P. and S. Gianni, *PDZ domains: folding and binding.* Biochemistry, 2007. **46**(30): p. 8701-8.
17. Tonikian, R., et al., *A Specificity Map for the PDZ Domain Family.* PLoS Biol, 2008.
18. Harris, B.Z. and W.A. Lim, *Mechanism and role of PDZ domains in signaling complex assembly.* Journal of cell science, 2001. **114**(Pt 18): p. 3219-3231.
19. Fagerberg, T., J.C. Cerottini, and O. Michielin, *Structural prediction of peptides bound to MHC class I.* J Mol Biol, 2006. **356**(2): p. 521-46.

20. Vanhee, P., et al., *Protein-peptide interactions adopt the same structural motifs as monomeric protein folds.* Structure, 2009. **17**(8): p. 1128-36.
21. Davey, N.E., G. Trave, and T.J. Gibson, *How viruses hijack cell regulation.* Trends in biochemical sciences, 2011. **36**(3): p. 159-169.
22. Ross, C.A. and M.A. Poirier, *Protein aggregation and neurodegenerative disease.* Nature Medicine, 2004. **10 Suppl**: p. S10-7.
23. Alber, F., et al., *The molecular architecture of the nuclear pore complex.* Nature, 2007. **450**(7170): p. 695-701.