Computational Investigation of Signaling Regimens using Proteomics Data

BY

MORTEN KÄLLBERG
B.Sc., University of Southern Denmark, 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioinformatics
in the Graduate College of the
University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:
Hui Lu, Chair and Advisor
Yang Dai
Michael Stroscio
Jung Hyun Min, Chemistry
Jinbo Xu, Toyota Technological Inst-Chicago

ACKNOWLEDGMENT

I would first like to thank my advisor, Professor Hui Lu, for his guidance, insightful advice, and constant encouragement throughout my graduate training. Dr. Lu's expertise in a broad range of topics has given me the opportunity to develop skills in numerous research areas during my tenure as a graduate student. In this process I was always afforded great freedom to pursue my own ideas and choose the path I found most interesting, for that I am grateful.

I would like to thank the faculty members of the Bioinformatics program Dr. Jie Liang and Dr. Yang Dai for creating a stimulating academic environment in which to conduct research. I thank Professors Jinbo Xu, Michael Stroscio and Jung Hyun-Min for taking time out of their busy schedules to serve on my thesis committee. In addition, I thank Professor Wonhwa Cho, and his lab members Young Chen and Ren sheng, with whom I have collaborated on the numerous projects investigating membrane-binding protein domains.

Further, I would like to thank current and former members of the Lu Lab, Matthew Carson, Georgi Genchev, Wang Xishu, Nitin Bhardwaj, Robert Langlois, Cong Liu and Wenyi Qin. Through the past five years they have provided for stimulating discussion on research questions and good company.

I would like to thank Lyndsi, who has been by my side through this entire experience, both when it was an adventure and an ordeal. Her loving support was what kept me sane throughout the process; her contributions are on every page.

ACKNOWLEDGMENT (Continued)

Lastly, I want to extend a special thanks to my sister Anne and my parents Ole and Anita whose constant encouragement and loving support throughout my entire education has allowed me to achieve this goal. I dedicate this thesis to them.

MK

TABLE OF CONTENTS

CHAPTER		PAGE
1	**INTRODUCTION**	1
1.1	Signal transduction and Proteomics	3
1.2	Signal transduction at the protein domain level	6
1.3	Motivation and Significance	8
1.4	Project Overview	10
2	**COMPUTATIONAL METHODS IN PROTEOMICS**	14
2.1	Molecular Dynamics	15
2.1.1	Force Field Functions	17
2.1.2	Boundary Conditions	19
2.1.3	Numerical Integration and Statistical Ensembles	20
2.2	Model System I: Mechanical Proteins	21
2.2.1	Steered Molecular Dynamics	23
2.2.2	Titin I27	25
2.2.2.1	Modeling the unfolding process	28
2.3	Machine Learning Methods	29
2.3.1	Supervised Classification	29
2.3.1.1	Support Vector Machines	30
2.3.1.2	Decision Trees	32
2.3.1.3	Alternating Decision Trees	33
2.3.1.4	Boosting, Bagging and Random Forest	34
2.3.2	Unsupervised Classification	36
2.3.2.1	K-means clustering	36
2.3.3	Evaluation of classifiers	37
2.3.3.1	Metrics	38
2.3.3.2	Plots	40
2.4	Model System II: Peripheral Membrane Targeting Domains	41
2.5	Protein Structure Modeling	44
2.5.1	Approaches to structure modeling	45
2.5.2	RaptorX: A server for protein structure prediction	47
3	**SCORING OF MASSSPECTRAL SEARCH RESULTS IN LARGE-SCALE PROTEOMICS STUDIES**	50
3.1	Introduction	50
3.2	Methods	55
3.2.1	Reference Dataset	55
3.2.2	Classification Algorithms	57

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

3.2.3 Evaluation Metrics . . . 59
3.2.4 Availability and requirements . . . 60
3.3 Results . . . 60
3.3.1 Feature Calculation . . . 61
3.3.2 Classifier Performance . . . 65
3.3.3 An Interpretable Model . . . 67
3.3.4 Extending the Peptide Prediction Protocol to Protein Prediction 73
3.4 Discussion and Conclusion . . . 74

4 GENOME-WIDE CHARACTERIZATION OF LIPID-BINDING IN SIGNAL TRANSDUCTION: A CASE STUDY OF THE PDZ DOMAIN FAMILY . . . 77
4.1 Introduction . . . 77
4.2 Methods . . . 79
4.2.1 Dataset . . . 79
4.2.2 Feature development . . . 81
4.2.2.1 Surface patch definition . . . 82
4.2.2.2 Mapping quantities onto the surface . . . 84
4.2.2.3 Sequence feature - functional classification matrix . . . 85
4.2.2.4 Classifier: SVM and AdaBoost on C4.5 . . . 86
4.2.2.5 Classifier evaluation . . . 87
4.2.2.6 Homology modeling . . . 88
4.3 Results . . . 89
4.3.1 Classification Model for Predicting Membrane-Binding . . . 91
4.3.2 Predictions for 2,000 PDZ Domains from 20 Different Species 98
4.3.3 Experimental Validation of Prediction . . . 98
4.3.4 Functional Classification . . . 101
4.4 Discussion and Conclusion . . . 108

5 LEARNING THE RULES OF MEMBRANE-BINDING . . . 113
5.1 Introduction . . . 113
5.2 Methods . . . 117
5.2.1 Dataset . . . 117
5.2.2 Classifiers and evaluation . . . 118
5.2.3 Features . . . 121
5.3 Results . . . 124
5.3.1 Overall classifier performance . . . 126
5.3.2 Knowledge-mining . . . 127
5.3.2.1 C1 model . . . 128
5.3.2.2 C2 model . . . 131
5.3.2.3 PH model . . . 132
5.4 Discussion and Conclusion . . . 135

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

6 MODELING THE UNFOLDING OF MECHANICAL PROTEINS WITH DISCRETE STATES 137
6.1 Introduction 137
6.2 Methods 144
6.2.1 Markov Chains 144
6.2.2 Defining the State Space 147
6.2.3 Evaluation of clustering algorithms 149
6.2.4 Ensuring Markovian Properties of the Model 150
6.3 Results 151
6.3.1 Diverse mechanical properties of I27 mutants predicted by SMD 151
6.3.2 A model for the unfolding of WT I27 157
6.3.2.1 Determining the clustering space 157
6.3.2.2 Ensuring Markovian properties 161
6.3.2.3 Unfolding networks at different forces 162
6.3.2.4 Key transition pathways 167
6.3.3 Explaining the effect of the Y9P mutant on the mechanical stability of I27 172
6.4 Discussion and Conclusion 175

7 CONCLUSION 179

VITA 182

CITED LITERATURE 184

LIST OF TABLES

TABLE		PAGE
I	Features used in the machine learning formulation of peptide-spectrum matching problem.	62
II	Validation metrics for a collection of classification procedures for peptide-spectrum matching problem.	66
III	Training set domains for PDZ classifier.	79
IV	Comparison of the performance of the SVM and ABC4.5 classifiers.	95
V	Experimental evaluation of our prediction for membrane binding of PDZ domains.	99
VI	Dataset statistics for the three domain families.	118
VII	Performance comparison of models for C1, C2, and PH domain families.	125
VIII	Simulation statistics for WT I27 and four mutants.	153
IX	Clustering space for different protein species.	159

LIST OF FIGURES

FIGURE		PAGE
1	A conceptual sketch of a signal transduction pathways resulting in gene transcription.	4
2	Layers of complexity in cellular function.	5
3	Conceptual illustration of force-field terms.	18
4	The setup of mechanical unfolding of I27 using SMD and the N-C-terminal extension as a function of time induced by forced unfolding.	25
5	Mechanical proteins.	26
6	An example of a ROC plot comparing three classifiers 1, 2, and 3. Classifier 1 represents the worst possible classifier doing no better than random for any FPR. Both classifier 2 and 3 dominate 1 over the entire range of FPR, with 2 having the highest AUC ROC.	41
7	An example of Membrane targeting domain interaction.	43
8	RaptorX job flow.	48
9	The principle steps of a tandem mass spectrometry experiment.	52
10	The Sequest rank distribution of correct database hits.	56
11	Receiver Operator Curves (ROC) and Precision/Recall Curves (PRC) for the peptide-spectrum matching problem.	68
12	Graphical representation of the alternating decision tree learned from the dataset *all* for the peptide-spectrum matching problem.	72
13	The three steps in determining surface patches.	83
14	Pseudocode for the patch growing procedure.	84
15	Plasma membrane localization of high-affinity PDZ domains.	90
16	The scores obtained from the RFC matrix for the PDZ domain PSD95.	92
17	Residue and cumulative scores obtained from the RFC matrix.	96
18	The prediction value for the SVM-1 μM K_d classifier as a function of the K_d value of domains	97
19	Membrane-Binding statistics for 2,000 PDZ domains found in 20 Species.	100
20	Functional Classification of Membrane-Binding PDZ Domains	106
21	Effects of Lipid Binding of Each Class of the PDZ Domain on its Peptide Binding.	107
22	Performance of a sequence based nearest neighbor classification procedure.	115
23	The ADtree model constructed for the C1 domain family.	128
24	Rules learned for the C1-family.	130
25	Rules learned for the C2-family.	131
26	Rules learned for the PH-family.	134
27	Conceptual depiction of a protein structure energy-landscape.	138

LIST OF FIGURES (Continued)

FIGURE **PAGE**

28 A transition network superimposed on the underlying energy landscape. 140
29 The construction of a Markov Chain Model of the protein unfolding energy landscape from a collection of SMD trajectories. 143
30 Overview of the procedure for constructing the MCM from a collection of SMD trajectories. 144
31 Pseudocode for SMD trajectory clustering procedure. 148
32 Overview of kinetic properties for five I27 mutant species. 152
33 Overview of break-time distributions from constant force-pulling of I27 mutants at five forces. 155
34 The average unfolding time for I27-WT and 4 mutants. 156
35 The adjusted Rand index 160
36 The three longest normal modes for the 250 and 500 pN 161
37 Unfolding network for wild-type I27 pulled at 250 pN. 165
38 A unified representation of I27 unfolding at a number of pulling forces. 166
39 The inter-strand drifting unfolding pathway. 168
40 The peeling of A strand unfolding pathway. 170
41 The unraveling from both ends unfolding pathway. 171
42 The support strand rearrangement unfolding pathway. 173
43 Y9P unfolding pathway. 175

LIST OF ABBREVIATIONS

ABC4.5	Adaboost C4.5
Acc	Accuracy
ADtree	Alternating Decision tree
AFM	Atomic Force Microscopy
AUC	Area Under Curve
CNF	Conditional Neural Fields
CRF	Conditional Random Fields
CV	Cross Validation
DAG	Diacylglycerol
ECM	Extracellular Matrix
GPCR	G-Protein Coupled Receptor
IMD	Interactive Molecular Dynamics
MCM	Markov Chain Model
MD	Molecular Dynamics
MS/MS	Tandem mass-spectrometry
MTD	Membrane Targeting Domain
MTT	Multiple Template Threading

LIST OF ABBREVIATIONS (Continued)

PDB	Protein Data Bank
PID	Protein Interaction Domain
PKC	Protein Kinase C
PLC	Phospholipase C
PM	Plasma Membrane
PS	Phosphatidylserine
PtdInsP	Phosphoinositides
RFC	Recursive Function Classification Matrix
RMBD	Reversible Membrane Binding Domain
ROC	Receiving Operator Characteristic
Sen	Sensitivity
SES	Solvent Exposed Surface
Spe	Specificity
SVM	Support Vector Machines
vdW	van der Waals

SUMMARY

We present computational methods addressing three key challenges in the quest to construct a more complete picture of protein signaling pathways, namely, confident identification of proteins in a sample, functional classification of large-scale proteomics data, and characterization of the dynamic conformational changes in protein structures.

First, we develop a probabilistic protocol for identification of short peptide fragments characterized by tandem mass-spectromety (MS/MS). A machine learning procedure for correctly matching peptides with mass spectra was constructed. Further, we demonstrated how the developed model can be represented as an interpretable tree of rules, thereby effectively removing the 'black-box' notion often associated with machine learning classifiers, making the underlying model clearer to end-users. Finally, using a probabilistic framework, a method for protein identification based on the peptide predictions was proposed and tested.

Second, a genome-wide functional classification protocol for identifying dual specificity membrane- and protein-binding domains was developed. Experimental characterization of 90 PDZ domains demonstrating that 40% had submicromolar membrane affinity was used for building a model utilized to predict the membrane binding properties of 2000 PDZ domains from 20 species. We demonstrate that reversible membrane binding is a key component in spatially regulation protein interaction networks and further propose a mechanistic classification of dual-specificity binding. As an extension to the PDZ domain models, we build a knowledge-mining procedure for learning the general mechanisms of membrane-binding, using C1, C2, and

PH domains as test-beds. We demonstrate how this method was able to uncover properties of each family known to be important in membrane-binding.

Last, we present a method for modeling the changes in single molecule dynamics induced by a signaling event as a discrete state Markov Chain model. Specifically, we use the partial unfolding of so-called mechanical proteins by way of steered molecular dynamics to demonstrate how the protein energy landscape is altered when different external mechanical forces are applied. By probing the protein structure with a range of forces, we show that the transitions pathways taking the protein structure from folded to partially unfolded vary significantly depending on the external input. The constructed model is instrumental in explaining experimental single molecule studies of the unfolding of the protein domain I27, as well as the changes in mechanical properties of a number of I27 mutant structures.

CHAPTER 1

INTRODUCTION

A fundamental principle of modern biology is that of *homeostasis*. The idea was first formulated by French physiologist Claude Bernard in 1865 as *milieu intérieur*, or the internal environment, referring to the dynamic changes in extracellular fluid composition ensuring a stable environment for tissue and organs in multicellular organisms. Today we more broadly define homeostasis as the ability of a biological system to regulate its internal environment such that properties like temperature, pH, and salt concentrations are kept stable.

Formally, homeostasis is the property enabling an organism to efficiently adapt to a wide range of conditions by responding properly to changes in its external environment. A homeostatic system is typically comprised of a *receptor* that monitors and responds to changes in the environment by sending a signal to an *effector* (typically organs or muscles) which seeks to counteract the deviation from acceptable levels of the property being monitored. Once the desired response has been achieved the signal is usually suppressed by *negative feedback*.

An example illustrating a homeostatic system is the regulation of blood-glucose levels. The blood concentration of glucose is to be kept at a dynamic equilibrium concentration at all times, a feat achieved through constant monitoring of glucose levels by the cells in the pancreas' Islets of Langerhans. To regulate deviations in sugar levels the pancreas releases the two counter-balancing hormones insulin and glucagon, the former promoting the uptake of glucose by muscle and liver tissue and the latter promoting the release of stored sugar back into the

bloodstream. The concentration of both hormones is regulated by negative feedback loops such that a reduction/increase in glucose levels will stop the release of insulin/glucagon.

A key discovery in understanding how the release of hormones and other messenger molecules act to regulate the behavior of the organism on the cellular level was made in the 1970s by Martin Rodbell. Studying the effects of glucagon on membrane receptors found in rat liver cells, Rodbell found that guanosine triphosphate caused the disassociation of glucagon from the membrane receptor, thereby stimulating the intracellular protein G and markedly changing the the metabolic activity of the cell (166). What Rodbell observed is an example of one of many modes of *signal transduction*.

Signal transduction is the process of mediating the message of an extracellular signal to elicit the appropriate intracellular response. A diverse set of molecular interactions is deployed in order to drive such information exchange, however, broadly described the process occurs in two stages. First, an extracellular signaling molecule, or *ligand*, binds to a transmembrane cell surface receptor such as G-protein-coupled receptors (GPCRs) or Receptor tyrosine kinase (165; 91). The binding of a ligand causes a transient change in the stability of the receptor protein, thereby inducing a conformation change in the structure of the receptor. Second, the structural change of the cell surface receptor can either cause the direct structural modification of one or more protein entities within the cell or indirectly affect key proteins through the release of so-called second-messenger molecules (such as Calcium or IP_3) into the cytosol. In either case the initial modification of key proteins in the cell will set of a chain of alterations in several interacting cytosolic proteins following receptor activation. The activation of such

signal transduction pathways provides the opportunity for a further fine-tuning of the cellular signaling response through signal amplification as well as the integration of information between a number of different signals known as cross-talk (172). Ultimately the activation of an effector protein will result in a cellular response such as gene expression, cell proliferation, or apoptosis. A conceptual depiction of a signaling transduction pathway is shown in Figure 1, here a ligand binds to an extracellular transmembrane receptor triggering a phosphorylation cascade, eventually activating a downstream effector target.

It is evident that to fully apprehend the complexity of living organisms, models of the signaling dynamics that enable coordination of basic cellular activities such as growth, tissue repair and immunity response are essential. Further, such models can form a basis from which errors in information integration causing systemic diseases like cancer, diabetes, and autoimmunity can be analyzed (93), thereby providing a starting point for rational selection of drug targets.

1.1 Signal transduction and Proteomics

Upon the completion of the human genome project, a complete library of all human genes became available to the scientific community (102; 195). This feat has often been described as 'unavailing the cellular blueprint.' The genome is, however, only a starting point (if a very powerful one) for understanding cellular dynamics, as it is the expressed manifestation of the genes that determines the function of a given cell.

It is now widely recognized that the genome is more accurately described as a parts list from which individual cells choose the subset of tools that fit their purpose in the organism. To understand the cell in its functional state at a given point in time, it is necessary to know which

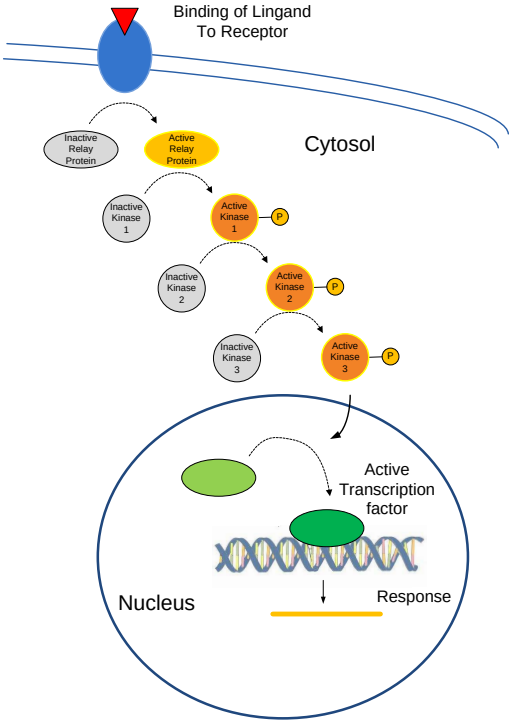


Figure 1: A conceptual sketch of a signal transduction pathways resulting in gene transcription. A ligand binds to an extracellular membrane receptor initiating a signaling cascade through the consecutive activation of protein kinases by phosphorylation. The end result is activation of a transcription factor, resulting in gene transcription.

genes have been transcribed and translated into a protein product. In fact, the full collection of mature proteins including all splice-variants and post-translational modifications (e.g., phosphorylation, glycosylation, and methylation) cannot be known from the static genome (47). In other words, the behavior of a protein population is dynamic, and layers of complexity not accessible by simply knowing the rates of synthesis of its individual constituents exist. Consequently, we need to directly study the expressed proteins to fully comprehend cell behavior.

Figure 2 illustrates the different levels of complexity in the cellular machinery. At the top level we find the genome, which provides the recipe for the protein entities making up the proteomic profile of the cell. A further level of detail can be obtained by examining the complex interaction networks formed by protein-protein associations. The highest level of detail is given by the study of single protein domain dynamics and how the activation of these domains occur through modification of the domain energy landscape.

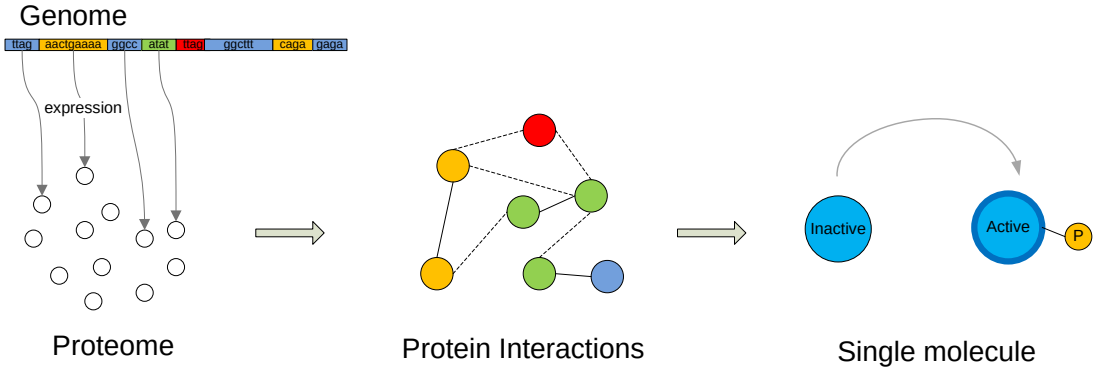


Figure 2: Layers of complexity in cellular function. The genome constitutes a parts list from which selected proteins are expressed to form the *proteome*. The dynamic properties of the proteome depend on the specific interactions between its constituents, which are ultimately guided by the specific functional state of each protein entity.

1.2 Signal transduction at the protein domain level

As alluded to above, the complexity of the proteome goes well beyond the sum of its parts. Substantial progress in understanding the networks of interacting partners that make up the pathways of cellular signal transduction has been made, giving rise to maps of so-called "interactomes" (213; 163). While such interactomes offer a broad overview of physical protein-protein contacts observed, deriving knowledge of the mechanisms behind the information flow requires an understanding of the distinct structural entities making up the proteins present in a given network.

The majority of genomic proteins are multi-domain units (more than 80% in multicellular organisms) consisting of several independently evolving sub-sequences, each displaying a unique fold and functional role within the context of the host-protein. To understand the role of the host-protein in a broader context, knowledge of each of its individual functional units is key. Traditionally, fold and biological function have been believed to display a one-to-one relationship with the fold uniquely dictating the functional role of the domain (215; 27). While this assertion holds true in some cases, it has been observed that even domains of highly similar fold can vary greatly in function, in fact only 38% of homologous catalytic domains sharing more than 60% sequence identity were found to be completely functionally identical, while 43% of these differ substantially in both substrate and co-factor specificity (167).

Ultimately we are faced with the question: What are the characteristics of protein domains that allow them to cooperate and form interactions with each other, giving rise to higher level behavior? It has become increasingly evident that both the native state and dynamic

shift in the protein domain structure induced by outside inputs are important factors. Thus, just like protein folding is understood in terms of an "energy landscape" with many meta-stable conformational states visited before the native state is reached, so too should a protein domain's ability to respond and transmit signals be considered a feature of its energy landscape (albeit one belonging to lower energy states than folding). To obtain a complete picture of how a protein domain carries out its function we need a detailed picture of the ways its energy landscape is modulated by other protein domains, small ligand and peptide molecules, and the covalent binding of functional groups (i.e. phosphorylation) (181).

It is, however, in general not feasible to exhaustively describe a high-dimensional energy landscape (76). Consequently, we must use methods for approximating key landscape features when exploring functional properties. A sufficiently accurate picture may be obtained by simply inspecting the protein sequence, deriving properties such as net charge, hydrophobicity score, and detecting deviations from expected amino-acid propensities (though in many instances sequence information is not sufficient to infer function as previously mentioned). The next level of energy landscape information can be found in the static structural data provided by x-ray crystallography experiments. In example, solving the structure of blood protein Hemoglobin made it readily apparent to researchers how the allosteric binding of oxygen occurred through cooperative modulation of the protein structure (155). Finally, the dynamics of the energy landscape can be explored by time-evolution simulation using molecular dynamics, and the distribution of various meta-states can be uncovered using Markov Chain Monte Carlo methods (178). Whichever level of detail is necessary for a specific application, it is clear that our ability

to understand and modulate signaling pathways ultimately relies on the development of computational protocols that accurately capture the key energy features of cellular communication.

1.3 Motivation and Significance

The study of the molecular interactions that make up signal transduction pathways are key in understanding the regulation of cellular function, a premise for rational design of treatment regimens for a number of pathological conditions. Recent decades have seen an explosion in experimental data from both large-scale proteomics studies and single molecule experiments (2). Large challenges do, however, still exist in integrating and unifying such data into generalized models that can be used for predicting the behavior of similar biological systems without conducting further experiments.

The motivation of this work is to bridge the gap between available experimental information and biological knowledge, allowing for *in silico* experiments to serve as a complement to, or a replacement of, more expensive wet lab methods. For example, consider the case where a collection of surface plasmon resonance experiments have been conducted to characterize the binding affinity of numerous protein domains for plasma membranes of varying composition. When faced with the task of determining the binding behavior of a newly discovered protein domain, by integrating the information from previous experimental work, the methods developed in this work allow us to predict membrane-binding properties of the new target without performing further experiments.

In addition, if we find that a prediction protocol generalizes the known data examples well (its predictions are accurate), analyzing the individual components of the model will provide

insight into the mechanisms governing the higher level behavior. Consider again the example of classifying membrane-binding protein domains. By inspecting the aspects of the model that lead us to classify a given protein domain as membrane-binding, we can suggest specific mechanisms important for membrane association. These insights can be used as a guide to experimentalists in forming hypotheses and designing experiments (for example, the identification of key protein residues can be used for suggesting mutation studies), thereby accelerating biological knowledge discovery.

The above described large-scale classification of protein domains by means of statistical protocols could potentially be of utility in identifying candidates for modifying network behavior. In the case of drug design, once a single regulation target has been chosen, designing a ligand that will provide the desired regulation does, however, require further knowledge of the specific structural domain components. To this end, the integrative view of the domain dynamics that can be obtained from sampling multiple molecular dynamics trajectories simulating specific interactions could provide valuable insights. A framework for representing molecular dynamics trajectories as a comprehensive model summarizing all major dynamic aspects of a protein domain is the focus of the last part of this thesis. The example used for the development here was so-called mechanical proteins, the method, should, however, be extendable to other systems.

In sum, the research described herein was carried out to address a number of major challenges in the study of signal transduction mechanisms using proteomics data. In this work we propose computational methods addressing three key challenges in the quest to construct a more complete picture of protein signaling pathways, namely, confident identification of pro-

teins in a sample, functional classification of large-scale proteomics data, and characterization of the dynamic conformational changes in protein structures.

1.4 Project Overview

The individual parts of this dissertation are tied together by the theme of uncovering signaling methods from proteomics data, with Chapter 2 providing an in-depth description of the computational methodologies used and biological systems studied. Even so, each chapter has been written to read as a self-contained account of the project presented, complete with background, discussion, and perspective. The chapters are organized as follows:

In Chapter 2 we review computational methods used for *in silico* studies of proteomics data. Specifically, molecular dynamics, machine learning, and protein structure modeling are discussed, the latter in the context of a protein structure prediction server, RaptorX, developed in conjunction with the research presented. Further, we review two model systems that are used for method development in the later chapters, namely *Reversibly membrane targeting domains* and *Mechanical protein domains*.

This chapter is in part based on the publications:

1. Genchev, **Källberg**, Gursoy, Mittal, Dubey, Perisic, Fang and Lu. *Mechanical Signaling on the Single Protein Level Studied Using Steered Molecular Dynamics.* Cell Biochemistry and Biophysics. 1085-91,95. 2009.
2. **Källberg**, Wang, Peng, Zhiang, Lu, Xu. *Template-based protein structure modeling using the RaptorX web server.* Nature Protocols. 2012. (Accepted)

In Chapter 3 we present a probabilistic protocol for identification of short peptide fragments characterized by tandem mass-spectromety (MS/MS). MS/MS provides a powerful platform for characterizing the proteomic profile of a cell or a tissue sample in a given state. It is, however, often difficult to validate the resulting protein list derived from such experiments. In this work a machine learning procedure for correctly matching peptides with mass spectra is developed. Further, we demonstrate how the developed model can be represented as an interpretable tree of rules, thereby effectively removing the 'black-box' notion often associated with machine learning classifiers, making the underlying model clearer to end-users. Finally, a method for extending the developed peptide identification protocol to give probabilistic estimates of the presence of a given protein in the sample is proposed and tested.

This chapter is in part based on the publication:

1. **Källberg** and Lu. *An improved machine learning protocol for the identification of correct Sequest search results.* BMC Bioinformatics. 11:591. 2010.

In Chapter 4 we present a machine learning protocol for genome-wide functional classification of dual-specificity membrane- and protein-binding domains. Emerging evidence indicates that membrane lipids regulate protein networking by directly interacting with protein interaction domains. Experimental characterization of 90 PDZ domains showed that 40% had submicromolar membrane affinity. Using a computational model built from these data, we predict the membrane binding properties of 2000 PDZ domains from 20 species,

showing that reversible membrane binding is a key component in the spatially regulation protein interaction networks. The accuracy of the prediction was experimentally validated for 26 PDZ domains.

This chapter is in part based on the publications:

1. **Källberg** and Lu. *Structural Feature Extraction Protocol for Classifying Reversible Membrane Binding Protein Domains.* Conf Proc IEEE Eng Med Biol Soc. 6735-8. 2009.
2. Chen/Sheng/**Källberg***, Silkov, Tun, Bhardwaj, Kurilova, Hall, Honig, Lu, Cho. *Genome-Wide Identification and Functional Annotation of Dual Specificity Protein- and Lipid-Binding Modules That Modulate Protein Interactions at the Membrane.* Molecular Cell. April 27., 2012. *Authors contributed equally to this work

In Chapter 5 we extend the machine learning protocol developed in chapter 4 to other protein domain families, specifically C1, C2, and PH domains. We present a machine learning protocol for determining membrane-targeting properties achieving 85-90% accuracy in separating binding and non-binding domains within families. Our model is based on features from both sequence and structure, thereby incorporation statistics obtained from the entire domain family and domain specific physical quantities such as surface electrostatics. By using the enriched rules in Alternating Decision tree classifiers we are able to determine the meaning of the assigned function labels in terms of biological mechanisms. The accuracy of the learned models and good agreement between the rules discovered

using the ADtree classifier and mechanisms reported in the literature, reflect the value of machine learning protocols in both prediction and biological knowledge discovery.

This chapter is in part based on the publication:

1. **Källberg** and Lu. *A structure based protocol for learning the family specific mechanisms of membrane binding domains.* BIOINFORMATICS. 2012. (Accepted)

In Chapter 6 we present a method for modeling the changes in single molecule dynamics induces by a signaling event as a discrete state Markov Chain model. Specifically, we use the partial unfolding of so-called mechanical proteins by ways of steered molecular dynamics to demonstrate how the protein energy landscape is altered when different external mechanical forces are applied. By probing the protein structure with a range of forces, we show that the transitions pathways taking the protein structure from folded to partially unfolded vary significantly depending on the external input. The constructed model is instrumental in explaining experimental single molecule studies of the unfolding of the protein domain I27, as well as the changes in mechanical properties of a number of I27 mutant structures.

CHAPTER 2

COMPUTATIONAL METHODS IN PROTEOMICS

In this chapter we review three computational methodologies used in the analysis of proteomics data. First, the principles of Molecular Dynamics (MD) simulation used for modeling the atomic level time-evolution of protein structures are reviewed, with specific focus on probing of non-equilibrium mechanical systems by way of Steered Molecular Dynamics (SMD). Second, we describe a collection of statistical methods for binary classification referred to as Machine Learning (ML) algorithms and their use in accurate function classification of protein domains. Third, we discuss the problem of protein structure modeling from a target amino-acid sequence. Our discussion of the three methodologies is not intended as an exhaustive account of any technique, but rather meant to introduce the tools as a means by which one can conduct *in silico* experiments exploring protein-driven signal transduction.

Two distinct biological signaling systems will serve as test-beds for the computational methods presented in this thesis. To better allow the reader to interpret the computational results obtained using our methods, a brief background review of each system is included following the relevant methodology section. The two systems that are used for this purpose are each characterized by their own unique role in signal transduction. *Model system I* is so-called mechanical protein domains that act as force-sensing messengers conducting signals through specific and reversible unfolding when subject to mechanical strain. *Model system II* consists of membrane-

targeting protein domains which partake in signaling networks by reversible translocation and binding to membrane surfaces.

2.1 Molecular Dynamics

Molecular dynamics (MD) is a computational simulation technique for modeling time-dependent physical quantities of a molecular system from the collective set of forces acting on each atom in the system at discrete time-points. Here we will cover the principles of MD as they apply to the modeling of biological macromolecules, specifically focusing on structures of protein domains obtained from X-ray crystallography experiments. Further, we limit our focus to describing the aspects of MD implemented in the simulation package NAMD (156; 83) used with the CHARM++ package for parallel computing, while noting that a number of other setups for carrying out MD simulations are available (43; 183).

Consider a protein structure consisting of N atoms with $\mathbf{r}^N = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ denoting the set of $3N$ atomic position coordinates, and m_i the mass of the i^{th} atom. By numerically solving the Newtonian equations of motion for every atom in the system we are able to determine the position and velocity of each atom as a function of time, and thusly track dynamic changes of the entire structure. Formally this notion can be expressed in the following set of differential equations:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{f}_i \tag{2.1}$$

$$\mathbf{f}_i = -\frac{\partial}{\partial \mathbf{r}_i} U \; i \in (1 \ldots N) \tag{2.2}$$

To solve this set of equations, the forces acting on each atom, $\mathbf{f_i}$, are determined from a potential energy function $U(\mathbf{r^N})$ dependent on the position of all atoms, thereby linking the motion of the atoms together. The potential energy function, typically referred to as a "force field," is most often decomposed into a number of additive terms each representing the effect of a specific type of atomic interaction. For our purposes the force field is made up of the following terms:

$$U = U^{VDW} + U^{coulomb} + U^{bond} + U^{angle} + U^{dihedral} \tag{2.3}$$

The terms in Equation 2.3 can be subdivide into bonded terms (U^{bond}, U^{angle}, and $U^{dihedral}$) and non-bonded terms (U^{VDW} and $U^{coulomb}$) accounting for interactions between atoms that share covalent bonds and non-bonded long to medium-range electrostatic interactions, respectively. Other so-called cross-terms representing complex interactions between the five base terms mentioned have been developed for special application. For our purposes the force field will, however, be limited to these basic interaction as they have been found to represent a good trade-off between computational complexity and accuracy for macromolecules (156).

As alluded to above, carrying out an MD simulation can be reduced to these three task:

1. Determining a force field that truthfully represents all atomic interaction while remaining computationally tractable in terms of complexity.
2. Numerically integrating the equations of motions for a fixed number of discrete time-step.
3. Using statistical mechanics methods for controlling macroscopic quantities such as temperature and pressure in the simulations setup.

The following subsections will address the details of each of these subtasks.

2.1.1 Force Field Functions

The three types of bonded interactions included in our force field refer to the stretching, bending, and torsional rotation of molecular bonds,

$$U^{bond} = \sum_{bond_i} k_i^{bond}(r_i - r_{0i})^2 \tag{2.4}$$

$$U^{angle} = \sum_{angle_i} k_i^{angle}(\theta_i - \theta_{0i})^2 \tag{2.5}$$

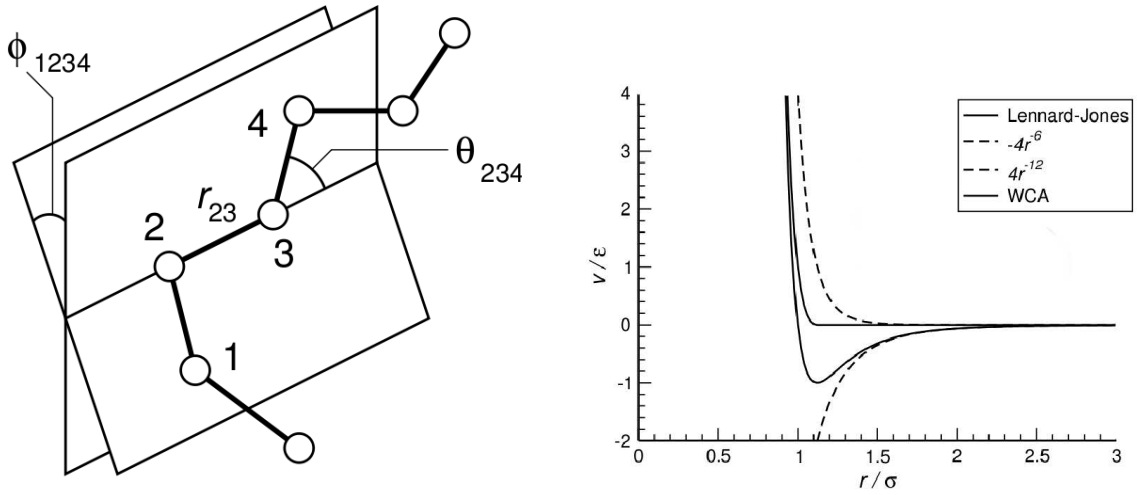
$$U^{dihedral} = \sum_{dihedral_i} k_i^{dihe}[1 + \cos(n_i\theta_i - \lambda_i)], n_i \neq 0 \tag{2.6}$$

where $bonds_i$ denotes all covalent bonds, $angles_i$ are all covalently linked 3-atom sets sharing a central vertex, and $dihedral_i$ are all atom pairs separated by precisely three covalent bonds with a central bond subject to a torsion angle. Figure 3(a) illustrates the three geometries for the atom-set {1,2,3,4}. The metrics r_{23}, θ_{234}, ϕ_{1234} do in this case correspond to the bond length, bend angle, and torsional angle, respectively, all of which have atom-specific equilibrium values denoted by r_{0i}, θ_{0i}, and λ_i.

The two remaining terms making up Equation 2.3 both describe interactions between non-bonded atom pairs:

$$U^{VDW}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.7}$$

$$U^{coulomb}(r) = \frac{Q_1 Q_2}{4\pi\epsilon_0 r} \tag{2.8}$$



(a) A six atom chain molecule illustrating the geometric terms needed to evaluate bonded potential functions. r_i is the inter-atomic distance between atoms 2 and 3; θ_i is the bend angle between atoms 2, 3, and 4; and ϕ_i is the torsion angle between atoms 1 and 4 around the rotation axis formed by atoms 2 and 3.

(b) Depiction of the different component making up the Leonnard-Jones potential for modeling VDW forces. The two distance attraction and repulsion component (r^{-6} and $-r^{-12}$) are shown as dashes lines along with the truncated WCA model of the potential.

Figure 3: Conceptual illustration of force-field terms.

U^{VDW} accounts for van der Waal's (vdW) forces as approximated by the 6-12 Leonard Jones potential, which is repulsive when interaction atoms are in close proximity and attractive at longer distances (illustrated in Figure 3(b)). Further, $U^{coulomb}$ represents long-range electrostatic interactions.

2.1.2 Boundary Conditions

Since simulations can only be carried out for a system of finite size, we have to deploy measures to avoid artifacts at the boundaries of the simulation cell. To this end so-called periodic boundary conditions are used by infinitely replicating the particles in the system cell by periodic translation in all spatial dimensions. In this setup, any particle that exits the system cell on one side reenters on the opposite side and is subject to the effects of all particles in all cell copies, thereby eliminating any cell boundary effects. It is important to note that while the effect of an infinite number of cells is present, in practical implementations the system need only be represented in one copy making the method computationally practical.

There are, however, computational limitations in using periodic boundary conditions as both VDW and electrostatic interactions exist between every pair of non-bonded atoms in all periodic cells. Since carrying out the a full computation of all VDW forces is intractable, these interactions are often truncated beyond a pre-specified cut-off distance. For long-range electrostatic interactions a cut-off scheme is, however, likely to introduce artifacts as the energy contribution of these terms drops off much slower than for VDW forces. To efficiently handle this problem particle-mesh Ewald (PME) summation is used for computing the full contribution of the electrostatic forces from all cells (156).

2.1.3 Numerical Integration and Statistical Ensembles

The simulation of large-scale biological systems often requires millions of integration time-steps, as the step-size in each iteration needs to be sufficiently small to adequately sample the fast oscillating modes of the molecular bonds and long enough to investigate global structural properties of biological interest. Often a highly accurate trajectory is less important than proper sampling of the phase space. Therefore when carrying out simulations where preservation of the particle count, energy, and volume of the system (the so-called NVE ensemble) is desired, one chooses an integrator based on its ability to model fundamental dynamic properties such as momentum, time-reversibility and energy.

To this end the Velocity-Verlet algorithm has proven useful. In this method, based on the position and velocity at time n, (r_n,v_n), and the forces, F_n, one can obtain (r_{n+1},v_{n+1}) by the following computations:

$$\begin{aligned}
\text{Half-kick} \quad & v_{n+1/2} = v_n + M^{-1}F_n \cdot \Delta t/2 \\
\text{Drift} \quad & r_{n+1} = r_n + v_{n+1/2}\Delta t \\
\text{Compute force} \quad & F_{n+1} = F(r_{n+1}) \\
\text{Half-kick} \quad & v_{n+1} = v_{n+1} = v_{n+1/2} + M^{-1}F_{n+1} \cdot \Delta t/2
\end{aligned}$$

In the above the vector M signifies the mass of all atoms in the system. Key advantages of the Velocity-Verlet scheme are the conservation of linear and angular moment and the fact that only one force evaluation is required per iteration. Further, for a fixed time-step the global method error grows in proportion to Δt^2 (156).

For certain applications we may wish to simulate other statistical ensembles, such as NVT (holding particle count, volume and temperature constant). This feat is achieved by modifying the Newtonian equations to generate the correct ensemble distribution by coupling the system to a thermal reservoir, thereby adjusting the forces by a factor proportionally to the kinetic energy of the system. To this end the stochastic Langevin equation is used to ensure the produced simulations adheres to the Boltzmann distribution for the canonical ensemble (NVT):

$$M\dot{v} = \mathbf{f}(\mathbf{r}) - \gamma v + \sqrt{\frac{2\gamma k_b T}{M}} G(t), \tag{2.9}$$

where M is the mass, F is the force, γ is a friction coefficient, k_b is the Boltzmann constant, v the velocity, T the temperature, and $G(t)$ is univariate Gaussian random process. The thermal coupling is thus achieved in part by adding a random fluctuation term (the later term) and a scaling term (γv) for appropriate adjustment of the forces.

2.2 Model System I: Mechanical Proteins

Cellular signaling mechanisms are driven by a controlled transfer of energy, whether through the buildup of a diffusion gradient, the conversion of chemical to electrical energy, or simply by driving a chemical reaction by means of the cellular energy currency ATP. One physical quantity in signal transduction that has gained attention within the last decade is that of

mechanical force. Mechanical force plays a crucial role in many physiological processes by regulating the reversible folding and binding of single protein domains (101; 85). Consequently, protein domains involved in these processes need to respond properly to mechanical strain in order to perform their function. Examples can be found in such diverse areas as stem cell differentiation (126) and the differentiation of myotubes (54). A comprehensive understanding of these processes on the molecular level will provide new insights into how a cell utilizes mechanical energy in transmission of signals.

Through the development of single molecule measurement techniques such as atomic force microscopy (AFM) (62; 162), optical tweezers (193), and surface force apparatus (107), mechanical behavior of a protein can be investigated on the single molecule level. Several single molecule studies have focused on proteins which are stretched and can withstand strain under physiological conditions (henceforth referred to as *mechanical proteins*). In example, mechanical proteins exist in muscle cells, on the extracellular matrix (ECM), on cellular surfaces, and in the cell nucleus. Specifically, muscle proteins such as titin control elastic behavior (113); ECM proteins, such as tenascin, signal through binding of cell receptors (140); membrane bound proteins such as cadherin, control cell adhesion through binding (51); lamins located in cell nucleus such as retinoblastoma protein regulate the cell proliferation and differentiation (84). In other words, proteins are responsible for numerous mechanical functions in the cell and a detailed understanding of how proteins function and interact under mechanical stress will provide unique insights into many areas of cellular function.

Experimental methods only partially reveal the underlying mechanics when studying mechanical proteins; for example, measurements from an AFM experiment will only provide the force at which the protein ruptures along with the extension length of the poly-peptide chain (133). Thus, to achieve a detailed understanding of the mechanical properties of protein structures one has to rely on computer modeling. Computer simulation and theoretical modeling of forced unfolding events have been addressed by a number of methods. Steered molecular dynamics (SMD) has been extensively applied in examining force induced protein unfolding events in the titin immunoglobulin domain and fibronectin type III domains, among others (118; 96). A so-called biasing potential, i.e., one that only exerts forces if a particle moves opposite a specified direction, was used in studying fibronectin unfolding (142; 143). In addition, several non-MD based methods are being actively pursued (such as a simplified description of protein unfolding simulations using both lattice and off-lattice models) (94; 95). A GO-potential has been used in studying titin I27 unfolding (39), and a generic contact potential has been applied to reveal details on ubiquitin unfolding (40). More recently, the Gaussian network model framework, otherwise used for studying protein stability in the native state, was extended to consider forced protein unfolding scenarios (56).

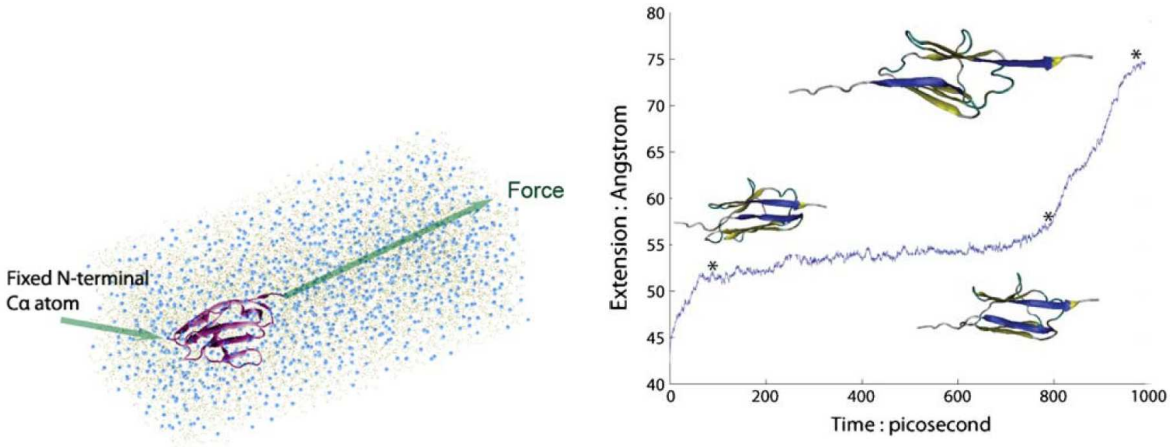
2.2.1 Steered Molecular Dynamics

SMD is intended to simulate the conformational changes occurring when a protein structure is exposed to mechanical strain. This is achieved by adding an external force to conventional force fields, thereby imitating the effect of mechanical strain on the protein domain. Many molecular dynamics packages, such as NAMD, CHARMM, GROMACS, and AMBER (31; 183)

provide a framework for performing SMD with various user specified force protocols. These tools are often combined with molecular graphics programs such as VMD (80), used for illustrating and analyzing the simulation results. Combining the NAMD package and VMD even allows for real time mechanical manipulation of the system of interest, by means of so-called interactive molecular dynamics (IMD) (72).

A typical SMD study is comprised of the following four steps: Solvation, energy minimization/equilibration, simulation, and data analysis. First, the PDB structure of the system of interest is placed in a box or sphere of solvent such as water as illustrated in Figure 4(a). In order to avoid artifacts arising from the introduction of solvent molecules, energy minimization and equilibration of the system is performed at room temperature to ensure that the system is in a stable, near native state prior to data collection. Subsequently, one can apply a force to the protein by fixing the position of one atom of the protein and adding a force to another atom. Typical force application protocols include constant velocity pulling and constant force pulling schemes. The resulting simulation trajectory and force recordings are analyzed by plotting characteristics such as (force; N-, C-terminal extension)-curves for constant velocity pulling or (time, N-, C-terminal extension)-curves for constant force pulling, as illustrated in Figure 4(b). In addition, key conformational changes such as hydrogen bond breakage and structural rearrangements can be monitored by animating the simulation trajectory.

While SMD has proven highly successful in reproducing and explaining experimental findings, the method does have certain potential limitations. For example, the current feasible computational and data storage capabilities create a practical limit of the simulation timescale



(a) Setup for pulling a protein structure at constant force using SMD. One end of the structure is fixed while a vectorial force is applied in a fixed direction.

(b) The N- C-terminal extension as a function of time in I27. The extension plot of mechanical proteins display a three-phase behavior each indicated by a "*" in the plot.

Figure 4: The setup of mechanical unfolding of I27 using SMD and the N- C-terminal extension as a function of time induced by forced unfolding.

that can be explored. In certain cases this requires application of a stretching force that is many-fold higher than the force applied in AFM experiments. Additionally, a recent long timescale (1 μs) folding study has suggested a bias in the potential force field towards misfolded states (64), meaning that results from very long simulations studying refolding pathways may be representative of a biased energy landscape.

2.2.2 Titin I27

Proteins undergo dynamical changes in response to mechanical forces under physiological conditions. It has been observed that the relative substructure placement, specific sequence segments, and the vectorial direction of the applied force with respect to the structure of

interest all play a role in the mechanical resistance of protein domains. A range of proteins have had their mechanical properties characterized, among these are Titin Ig domains (I27, I1,I32) (12; 69; 7) Fibronectin FnIII domains (141), Tenascin FnIII domains (137), Ubiquitin (38), Protein G (114), Top7 (29), Barstar (176), Spectrin (111), GFP (48) among many others. A key commonality between the majority of mechanical proteins is a collection of β-sheet at the core of the structure forming a network of hydrogen bonds preventing mechanical unfolding. Examples of the key hydrogen bond patches in Ubiquitin, I27 and Top7 are shown in Figure 5. Here we will use the Titin 27th Ig domain to illustrate the current understanding of mechanical resistance, henceforth referred to as I27.

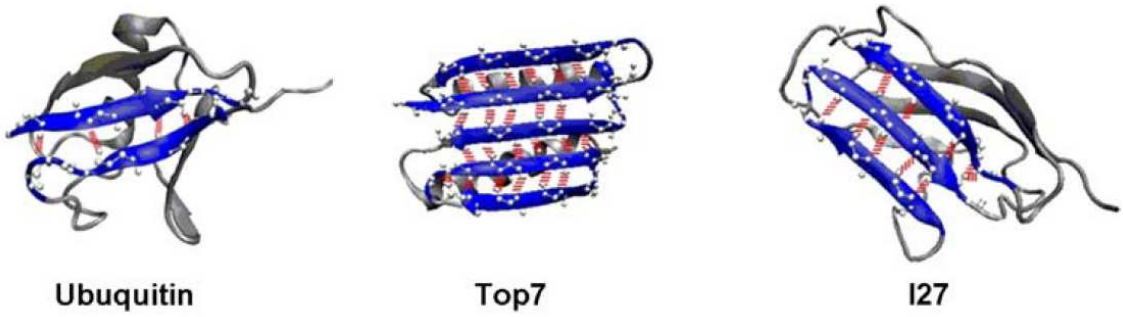


Figure 5: Hydrogen bonding network of mechanical proteins.

When a striated muscle fiber is stretched as a result of muscle contraction, a counter force develops, restoring the muscle fiber to its resting length. Titin, a major constituent of the

muscle sarcomere, is responsible for the passive elasticity of muscle. In addition Titin plays a key role in maintaining the integrity of the sarcomere against the shearing forces that arise during stretching. Titin is a multi-domain protein spanning half of the sarcomere from the Z disk to the M line. The I-band region of titin, responsible for the extensibility and passive tension developed during stretching, is composed of tandem repeats of immunoglobulin (Ig)-like domains and the small non-modular PEVK region. At the Z-disk of sarcomere, the N-terminal region of titin comprised of Iglike Z1 and Z2 domains interacts with ligand telethonin, thus anchoring titin to the Z-disc and preventing its shearing during muscle stretching. Protein engineering and single molecule AFM have revealed the mechanical components that form the elastic region of titin. While the PEVK region extends under weaker forces, it is the Ig domains that unfold reversibly under high force and protect the protein from rupturing (113).

The I27 domain consists of two anti-parallel β-sheets extending from the N- and C-terminal end of the structure, respectively. Applying two SMD protocols, one pulling with constant velocity and one with constant force pulling (119), revealed the presence of a significant energy barrier in this domain at an extension of $\approx$14 A (121) (measured as the distance between the N- and C-terminal residues), constituting the key energy barrier preventing unfolding. These results were consistent with AFM recordings, which revealed that each domain has a significant mechanical unfolding barrier (124).

From further SMD studies it was determined that unfolding occurs as a three-phase process: Phase I, the pre-burst extension, is signified by the breaking of H-bonds between β-strands A and B resulting in a stable unfolding intermediate. Phase II is the burst event where separation

of β-strands A' and G occurs. It is associated with the highest resistance force during unfolding due to the cooperative resistance offered by the hydrogen bonds formed between these two strands. In Phase III, the post-burst extension, little further resistance to extension is observed and all remaining hydrogen bonds are broken in a sequential manner (121). The SMD study of I27 revealed a key feature of proteins with strong mechanical resistance: A shearing topology connected by H-bond patches. Subsequent work revealed other distinct topologies displaying mechanical resistance, which confirmed the initial insight gained from the I27 experiments, namely that β-sheet folds do, in general, display stronger mechanical resistance than helix folds (120).

2.2.2.1 Modeling the unfolding process

Given the above establishment of a central hydrogen-bond patch as the main energy barrier preventing unfolding, the process has often been modeled as a simple two-state reaction with a single well defined activation energy as the rate limiting step. Based on the results by Bell (16), the force dependent unfolding rate, $k(F)$, can be described by $k(F) = A\exp(\frac{\Delta G - F\Delta x}{k_b T})$,with A being a pre-exponential factor, T the temperature, k_b the Boltzmann constant, ΔG the height of the free energy barrier of the reaction under the absence of force, Δx the physical distance from native to transition state along the reaction coordinate, and F the pulling force. From this relation we observe that force application serves to accelerate the unfolding process by lowering the free-energy barrier. Assuming no significant refolding rate under force application, $S(t)$

denotes the probability that a protein remains folded after time t, and can thus be determined using a first-order rate equation:

$$\frac{dS(t)}{dt} = -k(F)S(t) \tag{2.10}$$

From Equation 2.10 we have that $S(t)$ follows an exponential distribution. Recent work by Kuo *et al.* has, however, shown this distribution to be a poor fit to experimental data, indicating that the effect of force on the free-energy barrier is not well-modeled as a two-state process (98). There is thus a need for a detailed model of the unfolding process to capture the details of the energy landscape.

2.3 Machine Learning Methods

Machine learning collectively refers to a set of statistical procedures designed to approximate a given target function for classification purposes either from a set of pre-labeled training data (supervised learning) or from a set of uncategorized data (unsupervised learning).

2.3.1 Supervised Classification

Formally, the objective of a supervised classification protocol is to learn a function g from a set of labeled examples $(\mathbf{x}, y) \in S$ assumed to be identically and independently drawn from the same distribution D. In general we refer to examples as *instances* of the learning problem, with each instance being made up of a feature vector, $\mathbf{x}$, containing numerical values believed to reflect key properties of the instance and a label, y, signifying the category the instance belongs to. The objective is to determine a form and specific parameters of g such that the function maps $\mathbf{x} \in \mathbb{R}^n$ to values $y \in Y$ with the smallest possible error.

One of best characterized learning problems is that of binary classification. Binary classification refers to the special case where $y \in \{0, 1\}$ and $\mathbf{x}$ has a fixed length for all instances. A number of classification problems fall within this framework. For instance, consider an image recognition application where one is interested in determining whether a specific object is present in the image or not (e.g. a person or a bone fracture). In this case a human expert will determine the label values for the training data and feature vectors can be derived from the pixels of the image.

Since binary classification problems are very common, a number of algorithms have been devised for discriminating two classes based on a set of features given a predefined form of the function g (often referred to as the hypothesis space). The objective of all such methods is to reduce the probability of misclassification with respect to a specific choice of g and the distribution, D, from which the instances are drawn (though it is not necessary to know this distribution to minimize the error). Formally, $e(D, g)$ is given by:

$$e(D, g) = Pr_{(\mathbf{x},y),D}(g(\mathbf{x}) \neq y) \tag{2.11}$$

In the following sections we cover a number of binary classification algorithms, all of which are ultimately designed to minimize the expression in Equation 2.11.

2.3.1.1 Support Vector Machines

Support Vector Machines (SVM) classifier (44) is a so-called linearized learning scheme which seeks to determine a hyperplane or a set of hyperplanes in the feature-space (or a space of higher dimension than the feature-space) which can be used for classification or regression.

In the case of binary classification, the idea is to determine the hyperplane that provides the largest possible margin in separating two groups of data-points, as this hyperplane will be the least sensitive to noisy data and consequently provide the most consistent model.

Formally, given labeled training data-points $(\mathbf{x_i}, y_i)$, with $y_i \in \{-1, 1\}$ and $x_i \in \mathbb{R}^n$, SVM seeks to determine the parameters $\mathbf{w}$ and b for the plane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that provides the maximum margin between point for which $y_i = 1$ and $y_i = -1$, respectively. Assuming the training data is linearly separable this plane can be found by finding the parameters maximizing the distance between two hyperplanes parallel to $\mathbf{w} \cdot \mathbf{x} + b = 0$, namely $\mathbf{w} \cdot \mathbf{x} + b = 1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$. By use of geometry we find that this optimization problem can be solved by minimizing the Euclidiean norm $||w||$, subject to the constraints $\mathbf{w} \cdot \mathbf{x} + b \leq -1$ and $\mathbf{w} \cdot \mathbf{x} + b \geq -1$ for $y_i = -1$ and $y_i = 1$, respectively.

It is, however, rarely the case that experimental data is linearly separable, either due to noisy data or to the fact that the two classes of data-points are not separable by a hyperplane in the proposed feature-space. Two additions to the above outlined optimization problem are made to accommodate these challenges. First, rather than requiring all data to be classified with a margin of one, a so-called slack parameter, ξ, and a cost parameter C are introduced to allow for data falling on the 'wrong side' of the classification hyperplane. Further, by using so-called kernel functions, $\Phi(\mathbf{x})$, to map the feature vectors into an alternative vector-space, it is possible to construct classification hyperplanes in a higher-dimensional space, thereby introducing non-linear classifiers in the original feature space. Including these two features gives us the soft-margin formulation of the SVM classifier:

$$min_{w,b}\frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \tag{2.12}$$
$$\text{Subject to } y_n(w^T\Phi(\mathbf{x_n}) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

In practice, the dual form of Equation 2.12 is often used, as this formulation only requires one to define a closed-form expression for the dot-product of two feature vectors in a higher-dimensional space rather than explicitly calculating the single vector representation. There are a number of general families of kernel functions, including Gaussian, polynomial, and sigmoid, that have been shown to be of great utility for constructing general purpose classifiers. Determining the appropriate parameters for a specific kernel function given a learning problem is often done using a cross-validation procedure, through which the parameters that best tune a validation metric can be chosen (see Section 2.3.3 for more details).

2.3.1.2 Decision Trees

A decision tree is a classification model structured as a binary tree where each internal node denotes a split on a specific feature and each leaf node a classification. Algorithms for constructing decision trees, such as C4.5, all work by iteratively finding the best axis-parallel split on increasingly smaller subsets of the dataset in the feature space. This feat is achieved through a greedy search with respect to some loss function (also called an impurity measure)

indicating how much the homogeneity in terms of classification label is increased if we split the data set with respect to a specific feature.

Decision trees are simple interpretable models that are fast to construct. They do, however, suffer from a number of short-comings mostly stemming from the simple hypothesis space they work in. The fragmentation problem, in example, refers to the reduction in dataset size occurring when the tree depth gets large and the resulting tendency of over-fitting. Essentially, at some point there is too little data to reliably determine which feature to split on and what the threshold for the split should be. As we shall see in the later section on meta-classifiers, this tendency of over-fitting can to some extent be overcome through averaging over several models built on the same dataset.

2.3.1.3 Alternating Decision Trees

The Alternating Decision tree (ADtree) algorithm (65) is an alternative to the traditional decision tree in terms of mode of classification. In traditional decision trees each instance will traverse a specific path in the tree, with the leaf node of that path determining the final classification label. In the ADtree, on the other hand, each node is a voted stump classifier that adds a piece of evidence in terms of a real valued score towards the final classification decision; the final decision is then determined from the additive evidence of all stumps visited. The weight each rule has in determining the final classification label is established using the Adaboost algorithm discussed in Section 2.3.1.4.

The ADtree is thus not limited to the boolean logic characterizing decision trees, allowing for the representation of both dependent and independent rules in one tree model, a feat that

makes the ADtree a superior classifier to standard tree algorithms in many applications. Another advantage of the weighted voting scheme deployed by ADtrees is that they can often be represented with a limited number of nodes (when compared to decision trees), thereby making them more interpretable to human observers and thus suitable as a knowledge mining tool (123).

2.3.1.4 Boosting, Bagging and Random Forest

Boosting, Bagging, and Random Forest are not in themselves classification algorithms, but rather so-called meta-classifiers. A meta-classifier or *ensemble classifier* functions by combining classification results from a collection of models through a weighted voting scheme. There are several advantages to such strategies, especially for small datasets where a number of hypotheses in a given hypothesis space may appear to me near optimal due to over-fitting. By using an ensemble the average risk of choosing a inferior model is reduced. In addition, most often a model constitutes a local rather than a global optima in the hypothesis space, thus by using multiple starting points one is more likely to select a subset of models closer to the true unknown function that is sought. Finally, the correct classification model may not be representable by a single model in the chosen hypothesis space, but may be representable as a conjunction of models.

The Bagging algorithm (short for bootstrap aggregation) (24) creates an ensemble of models by iteratively creating a new model on a random sample of the original dataset. In each bootstrap iteration the original dataset is sampled with replacement to create a modified dataset of equal size. This type of ensemble construction is particularly advantageous when using

classifiers that are susceptible to noise as the base model, since any 'noise' in the original dataset will be averaged out by the sampling procedure. In bagging the the expected performance on unseen examples can be estimated using the 'out-of-bag' examples, defined as the error in classifying instances left out of the training dataset in each iteration, a feat that allows for efficient parameter tuning.

The Random Forest algorithm (25) is similar to bagging in that it creates an ensemble of classifiers by sampling the original dataset with replacement. The key difference between the two methods is, that unlike bagging, random forest not only samples instances in the dataset but also the feature space. The implication is that in any given iteration only a subset of the available features are used in learning a model. By sampling both the feature and instance space one is more likely to construct an ensemble of uncorrelated models. As with bagging the optimal size of the feature subsets to be used in each iteration can be estimate from the out-of-bag error.

The last, and arguably most successful, meta-classifier strategy is AdaBoost (short for Adaptive Boosting) (66). AdaBoost works by training a collection of *weak* classifiers each learned on a weighted distribution of the original dataset biased towards the instances wrongly classified in the previous iteration. The reweighing of each instance is based on Equation 2.13, where ϵ_m is the total error of the m^{th} iteration over the entire dataset. The final classification label of an instance is determined as a weighted sum over every classifier decision.

$$\alpha = \frac{1}{2} ln \frac{1 - \epsilon_m}{\epsilon_m} \tag{2.13}$$

2.3.2 Unsupervised Classification

In contrast to supervised learning procedure where a binary classification model is learned by training on labeled example data, unsupervised learning is concerned with the problem of discovering hidden structures in unlabeled data. Approaches that fall in the unsupervised learning category include clustering techniques and methods for feature extraction and dimensionality reduction (e.g. principle component analysis, single value decomposition, independent component analysis) (77). Here we will focus on k-means clustering as an example of unsupervised learning.

2.3.2.1 K-means clustering

Clustering refers to the task of partitioning a set of n data-points into subsets (called clusters) such that objects within the same cluster are more similar to each other than to objects in other clusters. In the case of the k-means partitioning the objective is to determine the best partitioning of a dataset into k subsets such that each instance belongs to the cluster with the nearest mean. While it is not computationally tractable to find a general partitioning algorithm that will guarantee an optimal solution (the problem has been shown to be NP-hard), there are heuristic procedures that will converge to a local optimum.

Formally, given a set of n instances, $x = \{x_1, x_2, \ldots x_n\}$, with each instance being a d-dimensional real-valued feature vector, we wish to partition the instance set into k subsets ($k < n$), $S = \{s_1.s_2, \ldots s_k\}$, such that the Within Cluster Sum of Squares (WCSS) is minimized:

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2 \tag{2.14}$$

The most common algorithm for minimizing the WCSS is Lloyd's Algorithm (122). Following an initialization step where k mean values $\{m_1, m_2, \ldots m_k\}$ are determined by randomly assigning data-points to clusters, the algorithm proceeds by alternating between an *assignment step* and an *mean update step*. In the assignment step new clusters are formed by allocating data-points to the cluster with the nearest mean-value as measured by a predefined distance metric (often Euclidean distance), in the update step the cluster means are updated with respect to the recent redistribution of data-points. These two steps are repeated for a fixed number of iterations or until the process converges (the cluster assignments no longer change).

2.3.3 Evaluation of classifiers

A key challenge when constructing a binary classification models is to evaluate how well the trained model can be expected to perform on unseen training data. In other words, to what extent can we expect the observed performance on training data to generalize when predicting the nature of new examples. Further, evaluation procedures are important in selecting the best model from a collection of models.

Model evaluation proceeds in three steps: Partitioning of the original dataset into training and test sets, calculation of evaluation metrics on the test set, plotting of evaluation metric correlations over the test set.

Various procedures exist for dividing the original dataset into a training and a test set, each appropriate for different dataset sizes. For large datasets the *Holdout* method, where 2/3 of

the dataset is used for training and 1/3 for evaluation is appropriate. If, however, the original dataset does not have a sufficient number of instances, the resulting test set from a hold out procedure may be too small and result in an overly pessimistic performance estimate. In this case *Cross-validation* is a more appropriate procedure. In n-fold cross-validation the dataset instances are distributed into n equally sized subsets. Subsequently n models are trained on a training set comprised of $n-1$ subsets and evaluated on the remaining subset. The extreme of this procedure is the *leave-one-out* evaluation where only one instance is left out in each cross-validation iteration. Most often, the leave-one-out method is too computationally expensive, thus $n \in \{10 \dots 20\}$ is commenly used.

2.3.3.1 Metrics

For binary classification there are four possible classification scenarios for an instance: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These metrics are often summarized in a so-called confusion matrix:

N	PPos	PNeg
NP	TP	FN
NN	FP	TN

The sum of the columns in this table gives the count of positively (Ppos) and negatively (PNeg) predicted instances, respectively. The sum of the rows indicate the actual counts for positive (NP) and negative (NN) instances in the training data. N indicates the total number of instances.

The most common question we would want answered is 'How likely is an instance to be predicted correctly?,' or in other words, what is the *Accuracy* of the model. Equation 2.15 defines $P(\hat{y} = y)$, where y is the true label of the instance and $\hat{y}$ is the predicted label, as the proportion of all instances predicted correctly. Similarly the error rate can be obtained as $1 - Accuracy$.

$$Accuracy = P(\hat{y} = y) \approx \frac{TP + TN}{N} \tag{2.15}$$

The accuracy measure does, however, not tell us anything about how prediction errors are distributed among the two prediction classes. Specifically, we want to know what the type I and type II error rates that can be expected from a given model are. *Sensitivity* as defined in Equation 2.16 estimates the conditional probability of a positive example being predicted as positive, and is approximated by the number of TP instances divided by the number of all known positive examples. The type II error of a model can be estimated as $1 - sensitivity$.

$$Sensitivity = P(\hat{y} = +|y = +) \approx \frac{TP}{NP} \tag{2.16}$$

In the same manner, we define *Specificity* as the probability that a negative instance will be predicted as negative. This probability is estimated as the proportion of correctly predicted negative examples divided by the total number of negative instances as indicated in Equation 2.17. The type I error of a model can be estimated as $1 - specificity$.

$$Specificity = P(\hat{y} = -|y = -) \approx \frac{TN}{NN} \tag{2.17}$$

2.3.3.2 Plots

The above presented metrics do, however, suffer from the limitation that they are sensitive to major skews in the class distribution of the dataset. This feat is overcome by so-called ranking metrics which summarize the expected model performance for all possible class distributions in one metric. One such metric is the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUC) is a measure of how good the confidence rated predictions of a classifier are. Consider the case where two models both misclassify a single instance as negative but with different confidence, say 0.8 and 0.41, respectively. Judging these two models by accuracy, we would deem their performance equal. Using ROC, however, we would say that the latter of the two performs the best. The logic behind this notion is that a classifier which is less confident in its incorrect predictions is more desirable.

Figure 6 illustrates the use of ROC in classifier evaluation. The true positive rate (sensitivity) is shown as a function of the false positive rate (1-specificity), and is determined by either sampling different class distribution from the dataset or ranking instances by confidence values and calculating the metrics at threshold cut-offs. The better a classifier performs the closer it will be to the top left-hand corner of the plot; a random classifier will produce a diagonal line from the bottom left-hand corner to the top right-hand corner. A classifier is said to be

strictly dominating another classifier if at every point of the curve it is to the left and above the competing classifier.

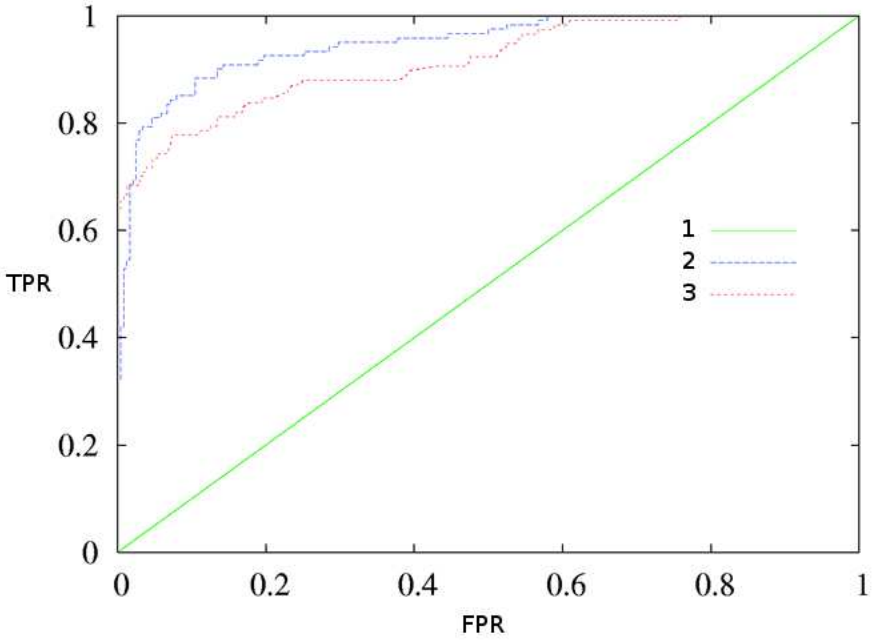


Figure 6: An example of a ROC plot comparing three classifiers 1, 2, and 3. Classifier 1 represents the worst possible classifier doing no better than random for any FPR. Both classifier 2 and 3 dominate 1 over the entire range of FPR, with 2 having the highest AUC ROC.

2.4 Model System II: Peripheral Membrane Targeting Domains

The majority of genomic proteins are multi-domain units (more than 80% in multicellular organisms) consisting of several independently evolving sub-sequences, each displaying a unique

fold and functional role within the context of the host-protein. To understand the role of the host-protein in a broader context, knowledge of each of its individual functional units is key. Traditionally, fold and biological function has been believed to display a one-to-one relationship with the fold uniquely dictating the functional role of the domain (215; 27). While this assertion holds true in some cases, it has been observed that even domains of highly similar fold can vary greatly in function, in fact only 38% of homologous catalytic domains sharing more than 60% sequence identity were found to be completely functionally identical, while 43% of these differ substantially in both substrate and co-factor specificity (167).

Numerous cases of protein folds illustrating the feat of high functional discrepancy in spite of structural similarity can be found among so-called Reversible Membrane Binding Domains (RMBDs). RMDBs exist in a wide-array of cytosolic proteins, and serve the task of translocating their host-protein to a membrane surface in response to a signal induced change in membrane composition (see Figure 7). Given the diversity and complexity in membrane structures it is not surprising that many membrane and lipid binding protein domains can be found in the eukaryote proteome (82), in fact RMDBs have been observed in such diverse domain families as C1 (33; 190; 210), C2 (33; 164; 134), PH (61; 109), FYVE (Fab1/YOTB/Vac1/EEA1) (186), PX (phox) (209), ENTH (Epsin N-terminal homology)(28), ANTH (AP180 N-terminal homology)(28), BAR (Bin/Amphiphysin/Rvs) (74; 188), FERM (Four point one-ezrin-radixin-moesin) (26), tubby (30), and recently PDZ domains (68).

The reversible binding of RMDBs to membranes (plasma-membranes in particular) serves to transiently compartmentalize the cytosolic space by clustering their host proteins at specific

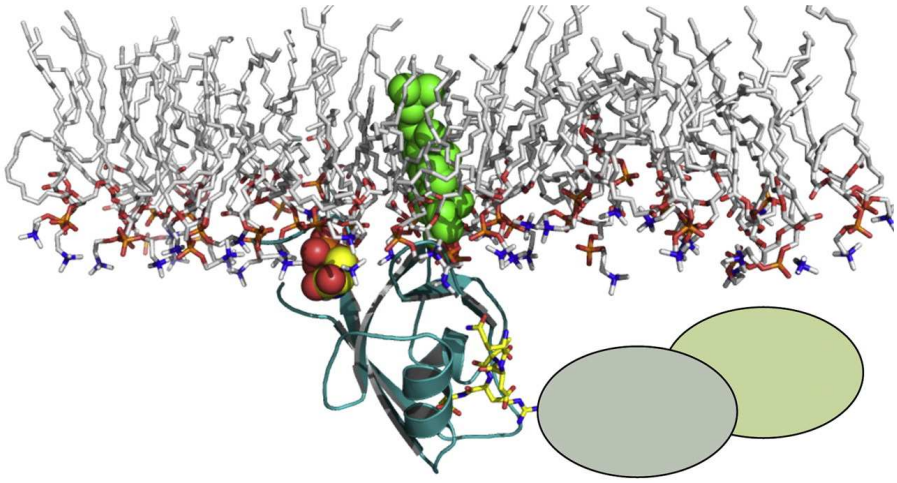


Figure 7: A PDZ (cyan) domain interacts with both membrane and other protein domains (represented by ovals) thereby regulating protein interactions networks at membrane surfaces.

membrane locations. This co-localization helps reduce the dimensionality of the space in which proteins interact by effectively increasing the local concentration of interacting parties, thereby allowing for greater efficiency and specificity in the signal transduction processes (127).

The recruitment to membranes can occur by both specific binding to lipid-head groups and by non-specific binding to membranes (36). The first examples of lipid-specific membrane binding were found in protein kinase C (PKC), a core component of several signaling pathways. PKC contains two C1 domains that bind phorbol esters, diacylglycerol (DAG), and membranes (158; 88) as well as a C2 domain binding Ca^{2+} (in the isoforms α,βI,βII,γ) thereby facilitating binding to acidic phospolipid membranes by making the overall electrostatic profile of the domain more positive. In general, a number of properties have been found to be of importance in

membrane binding (albeit not all to the same extent in all domain families), these include: The nonspecific electrostatic attraction between anionic membranes and cationic surface residues, association of hydrophobic surface residues with the membrane hydrocarbon core, and the specific interaction between key residues and lipid head-groups.

While experimental approaches have been successful in identifying novel RMBDs (33; 49), and efforts using FRET (42) and spin labeling (9) have been valuable in shedding light on the principles of lipid-binding, large scale identification and description of these domains remains labors-intensive and expensive. Computational protocols offer an efficient alternative to experimental identification procedures, allowing for rapid characterization of thousands of domains. Membrane binding properties are, however, difficult to predict, as they are not determined by well-defined sequence motifs or a specific structural composition. PH domains have, for instance, been found to span a large range of binding affinities though being very similar structurally (109; 180). For this reason a more sophisticated method for functional classification of RMBDs is necessary.

2.5 Protein Structure Modeling

Functional properties of a protein domain, such as enzymatic activity (5) or the ability to interact with other proteins (75), can often be derived from the approximate spatial arrangement of its amino acid chain in the folded state. Knowing the structure of a newly discovered protein is thus highly valuable in determining the role it plays in biological processes, and can serve as an important stepping-stone in generating hypotheses or suggesting experiments to further explore its nature. While the Protein Data Bank (PDB) (10) provides experimentally solved

structural data for an increasing number of protein domains, solving protein structures remains costly, time-consuming, and in certain instances, technically difficult. Consequently, the vast majority of protein sequences available in public databases do not have a solved structure at this point in time. More than ≈10 million unique protein sequences have been deposited, while only a little more than 70,000 have had their structures solved. To bridge this gap, a wide array of computational protocols for protein secondary and tertiary structure prediction from a target sequence are continuously being developed.

2.5.1 Approaches to structure modeling

Computational structure prediction methods can in principle be divided into two categories, template-based and template-free modeling, with some composite protocols combining aspects of both. Methods in the former group include Comparative Modeling methods (125), which, given a target sequence, identify evolutionarily-related templates with solved structure by sequence or sequence-profile comparison (e.g., BLAST and HHpred (191)), and construct structure models based on the scaffold provided by these templates. Alternative methods build on the observation that known protein structures appear to be comprised of a limited set of stable folds. It is thus often found that evolutionarily distant or unrelated protein sequences share common structural elements, a feat utilized by threading methods, such as MUSTER (206), SPARKS and RAPTOR (207; 208). It has been demonstrated that in some cases incorporating structural information to match the query sequence to potential templates enables similarity in fold to be detected despite lack of explicit evolutionary relationship.

Template-based modeling can generate useful approximate models for a large number of sequences with relative ease if close templates are available. Current methods do, however, become unreliable when there are no homologs with solved structures in PDB or when templates under consideration are distant homologs (6). Template-free methods offer an alternative for modeling such difficult cases. Pure *ab initio* methods (117; 179; 205) aim at building a 3D model using only primary structure information, the successful application of such methods is, however, limited to short target sequences (<120 residues) at present. In addition, a number of semi-ab initio approaches exist that assemble short structural fragments or use statistical information to spatially restrain the building of a model structure. Finally, so-called composite-methods, which combine subsets of the previously mentioned approaches, have been very successful in recent Critical Assessment of Protein Structure Prediction (CASP) competitions, most notably the TASSER methodology developed by Skolnick and Zhang (214).

While all of the aforementioned methods have made significant contributions to the field of structure prediction, it remains challenging to accurately predict the structure of a target sequence with a sparse sequence profile with no close homologs in the PDB. It has been estimated that 76% of the 4.2 million models deposited in MODBASE (157), a database repository for theoretical structure models, are built from remote homologs. Thus any improvement in structure prediction methods addressing these cases will have a significant impact on the utility of such theoretical models, and our ability to assign functional properties based on common fold patterns.

2.5.2 RaptorX: A server for protein structure prediction

For many of the computational protocols presented in this work, constructing a structure model from a protein sequence is a key step. In this final section we present a public web-server, RaptorX, developed in conjunction with other research projects, to automate the practical steps necessary for constructing a structure model of a target sequence.

As a reference, Figure 8 outlines the work-flow of the three modeling tasks users can accomplish using the RaptorX server, namely tertiary structure prediction, secondary structure prediction and custom alignment. Each task is decomposed into a number of timed conceptual steps with the logical flow from one step to the next indicated by the connecting arrows.

Template-based modeling critically depends on the quality of the target-template alignment. Previously, programs such as RAPTOR have been successful in efficiently optimizing the general protein threading scoring function and are among the best structure prediction protocols available as demonstrated at previous CASP evaluations (207). RAPTOR and other state-of-the-art threading programs are, however, limited by their linear scoring function, which cannot accurately represent any correlation that may exist among the features used for assessing alignment quality (for instance, secondary structure and sequence profile are known to be correlated). Further, the application of structural information in the alignment process does not take into consideration the level of similarity between target and template. Using structural information when modeling a target with a high-similarity template might introduce noise, while structural information becomes relatively more important when modeling a challenging target with sparse sequence profile.

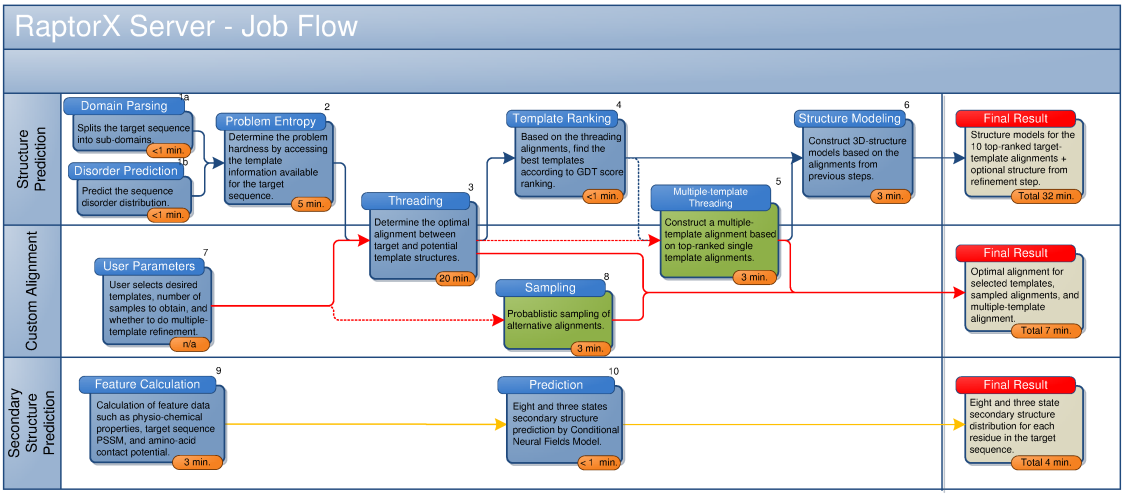


Figure 8: RaptorX job flow. For each step details on the computation and approximate completion time for a 200 residue target sequence are given (for step 3 the indicated time is for a full template library scan). Blue boxes indicate mandatory steps, green optional step, and grey resulting output. The blue, red, and yellow directed paths indicate the flow for structure prediction, custom alignment jobs, and secondary structure prediction, respectively. Dashed/solid paths indicate the following step to be optional/required.

To better address cases where no close template exists, a number of novel modeling strategies are introduced in the new software RaptorX (153), taking a completely different approach than that used in RAPTOR. First, a profile-entropy scoring method, which takes into consideration the number of non-redundant homologs available for the target sequence and template structure, is used to assess the quality of information content in sequence profiles (151), thereby allowing us to optimize the modeling strategy specifically to the target. Second, a Conditional Random Fields (CRF) for integrating a variety of biological signals in a non-linear threading score function not previously used by any threading software(150) is introduced. Finally, we have implemented a multiple-template threading (MTT) procedure (152), enabling the use of multiple templates to model a single target sequence. Unlike other MTT methods which mainly increase the alignment coverage, our MTT method can partially correct errors in pairwise alignments by exploiting inter-template similarity and thus improve the final model quality. To supplement structure prediction, RaptorX also provides domain parsing of long protein sequences and disorder prediction to help users interpret secondary and tertiary structure prediction results.

Aside from structure modeling, RaptorX server can be used to obtain custom pairwise target-template alignments and to generate an arbitrary number (<1000) of alternative pairwise alignments through probabilistic sampling as well as to generate single-target-multiple-template alignments. Further, RaptorX also provides a Conditional Neural Fields (CNF) based prediction protocol for determining the 3-state or 8-state secondary structure distribution for each residue in a target protein.

CHAPTER 3

SCORING OF MASSSPECTRAL SEARCH RESULTS IN LARGE-SCALE PROTEOMICS STUDIES

3.1 Introduction

The analysis of composite protein mixtures by use of mass spectrometry techniques has become a standard methodology for characterizing the proteomic profile of a cell or tissue sample (2). Mass spectral data has proven valuable in addressing complex problems such as the reconstruction of metabolic pathways (8; 73) and protein-protein interaction networks (71; 78), and is of great utility in applications spanning from the quantification of bacterial proteomes (115) to the investigation of infectious states in soybeans (198).

The basic principles of tandem mass spectrometry (MS/MS) (1) experiment for protein identification are illustrated in Figure 9. First, a protein sample is extracted from the cell culture of interest and digested using one of a number of site specific digestion enzymes, such as Trypsin. This process results in a mixture of short peptide fragments (typically on the order of eight to ten residues) stemming from all proteins present in the sample. Before conducting the mass-spectrometry analysis, a rough separation of protein fragments is typically done using HPLC to group peptides of similar mass together.

In the mass spectrometer the peptides are ionized so that they will carry one or more charges before being sent into the first mass analyzer. Here the mass of the individual peptides present

in the sample is determined and based on the resulting mass spectrum. A second process then iterates through the unique spectral peaks to characterize each of the peptide fragments. A unique "spectral fingerprint," by which the sequence of a fragment can be determined, is obtained by breaking the peptide into its amino acid components and generating a second mass spectrum from this fragmentation. The end result of the experimental procedure is a large (often > 20000) collection of mass spectra each corresponding to a peptide fragment in the original sample. To efficiently use the MS/MS technique in large scale protein characterization studies, robust and consistent data analysis procedures are required in order to confidently identify the originating peptide of a given spectrum and ultimately its parent protein. To this end, the combination of spectral data and the vast amount of genomic sequence information available in public databases has proven extremely rewarding. Algorithms such as Sequest (53), Mascot (154), and X!Tandem (45) (amongst others (168; 135)) can correlate thousands of mass spectra with theoretically derived peak lists from database peptide sequences, thus effectively automating the interpretation of experimental data. For the above mentioned algorithms, the result of a single spectrum searched against a database typically consists of a set of highly correlated peptide sequences along with a correlation score and a number of additional metrics intended for validation of the specific peptide-spectrum match.

There is, however, often no direct interpretation of these scores in terms of statistical significance (161), therefore simply ranking well-correlated peptides by metrics provided from the initial database search procedures and selecting a cut-off for filtering true matches from false ones is not desirable. Depending on the choice of threshold such a procedure will either be too

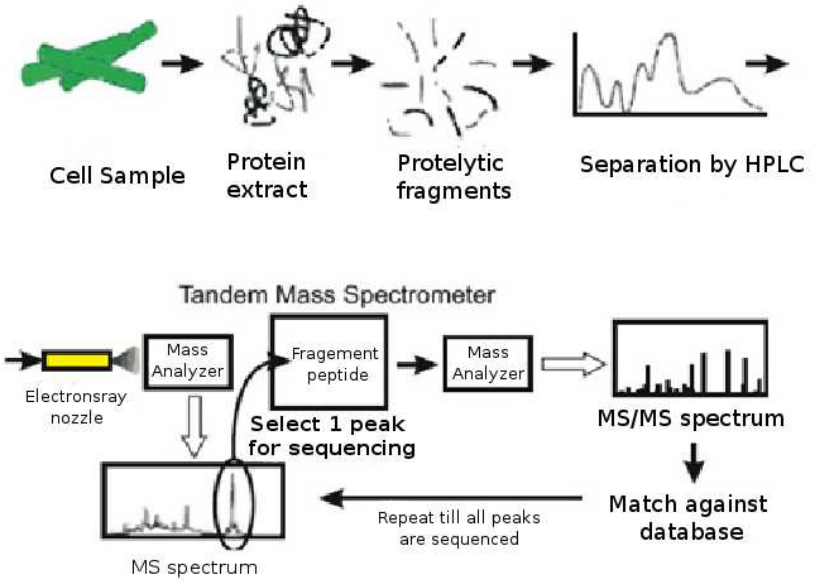


Figure 9: The principle steps of a tandem mass spectrometry experiment.

conservative or yield a high rate of false-positives (185). On the other hand, manual validation of the large amount of data produced by MudPIT style (200) experiments would be time consuming and out-of-tune with the high-throughput experimental work-flow characterizing the field at present. Thus, to ensure an effective production pipeline, a fully automated method for confident validation of the results produced by the above mentioned search algorithms is essential.

A number of procedures for validating peptide-spectrum matches have been suggested, either as direct extensions of the Sequest or Mascot algorithms or as supplementary post-processing

tools (182; 131; 169; 55). Our focus here will be on the analysis of the search results produced by the Sequest algorithm (53), and how to efficiently improve the number of true peptide-spectrum matches identified at a controlled false positive rate.

Currently, the most widely used tool for evaluating Sequest search results is the PeptideProphet methodology developed by Keller *et al.* (90; 89). By use of an empirically determined probabilistic mixture model based on the fitting of assumed distributions of various metrics (believed to reflect the reliability of the spectrum-peptide match) the search results are evaluated. The procedure returns a probability estimate of a peptide being present given the database search results. While giving much higher sensitivity measures than simple threshold based methods, this approach does suffer from two short-comings: First, there is no theoretical work supporting the assumptions made regarding the distributions used to fit the features utilized. Second, the model may not be easily extendable when potentially discriminatory information from novel types of data become available.

Machine learning provides an attractive platform for addressing the above concerns since no prior assumptions about the distribution of the individual features have to be made. In addition, the flexibility in feature handling of most machine learning algorithms makes further improvement of predictive power and robustness straight forward as new information becomes available. In recent years a number of bioinformatics problems have been addressed using machine learning (104), for example, the prediction of protein-DNA interactions (14; 15; 106) and protein-membrane interactions (16). Likewise, previous works have used machine learning methods for identifying true peptide-spectrum matches through different formulations of the

problem. Anderson *et al.* (4) were the first to apply such procedures to mass spectral data in their study of Support Vector Machines (SVM) classification of Sequest search results from ion-trap data. Razumovskaya *et al.* conducted a similar study demonstrating how a neural network could improve the filtering of Sequest search results to be superior to simple threshold-based procedures. In a study using ion-trap data, Elias *et al.* demonstrates how the identification of peptide-spectrum matches can be improved through probabilistic modeling of fragment intensities observed in the spectrum at hand (52). Ulintz *et al.* (194) developed an approach using tree-based ensemble algorithms and demonstrated that these were superior to the SVM protocol used in previous studies. A recent study has further demonstrated how physiochemical properties of the peptide in question can provide discriminatory power between true and false matches without using database search engine scores (59). Finally, it has been demonstrated that so-called consensus approaches in which the combination of information from several different database search schemes can provide additional discriminatory power.

From the above review it is clear that a variety of supervised classification regimens, using many different sources of information, have been tested thus far. Here we present a work that improve on three separate aspects on the above mentioned machine learning procedures for identifying correct peptide-spectrum matches from the Sequest database search procedure.

First, our classifier performs 6% better as measured by the area under the ROC compared with results by Ulintz *et al.* (194) using the same dataset. The improvement is achieved by introducing a number of global dataset features that take into consideration factors such as the total number of peptide-spectrum matches belonging to a given protein and the percentage of

potentially observable peptide sequences from a given protein actually appearing in the search result.

Second, by using the Alternating Decision Tree (ADTree) (65) classification algorithm we are able to represent the developed model as a tree with a limited number of nodes, thereby rendering the model interpretable to humans. While this trade does not add anything in terms of predictive power, interpretability of the model makes the procedure clearer to experimentalists and allows us to compare the prediction rules to expert rule-of-thumb, giving an empirical validation of such rules.

Third, we build a straight-forward probabilistic procedure for extending the machine learning identification of the peptide-spectrum matches into the protein prediction problem (i.e. identifying the proteins contained in the initial sample) by converting the classification scores into true probability estimates by means of logistic calibration. The latter of the two problems is often of most interest to experimentalists, as one is interested in knowing the probability of a protein being in the sample, not simply which peptide fragments were confidently identified.

3.2 Methods

3.2.1 Reference Dataset

Our method was tested using a publicly available MALDI MS/MS dataset obtained from a sample of 246 known proteins (58) published on ProteomeCommons.org (57). The peaklists were searched using the Sequest algorithm (53) on the IPI human FASTA database ver. 3.14 (for comparison with results reported by Ulintz *et. al* (194)) with the post-translational modification methylation, oxidation, and phosphorylation. Comparison with the PeptideProphet

(90; 89) validation results of the Sequest output was done using output from ver. 4.1. The PeptideProphet ROC curve and evaluation metrics reported below were obtained from this output.

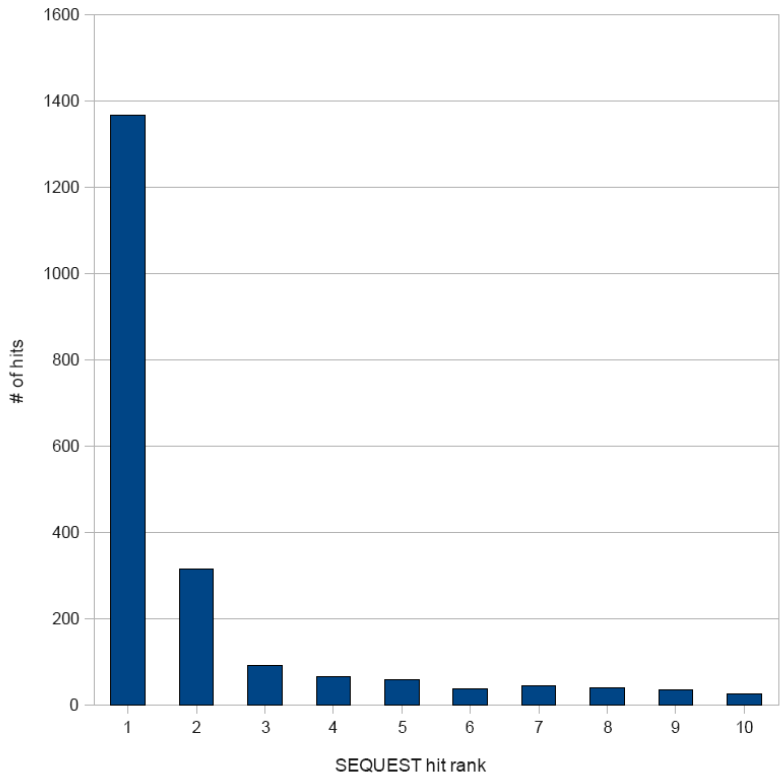


Figure 10: The Sequest rank distribution of correct database hits.

To correctly evaluate our approach the original dataset was split into two, one for method validation and one for training the machine learning protocols. Each of these datasets consists of

a total 43,348 examples of which 2,035 are correct peptide-spectrum matches. In contrast to the comparison works by Ulintz *et al.*, (194) all 10 top-ranked tentative peptide matches from each spectrum searched are included in the training and testing set. Including all potential matches is important, as 34% of the true matches have been found not ranked first as illustrated in Figure 10. Furthermore, only including the top one or top five ranked matches will exclude some potentially difficult to classify instances that may add valuable information for identifying novel proteins.

3.2.2 Classification Algorithms

Models were constructed using four different binary classification procedures, namely AdaBoost (66) applied to C4.5 (159) and Willow tree (103), Random Forest (25) applied to C4.5, and Alternating Decision tree (65) (in the following denoted ABC4.5, ABWillow, RFC4.5, and ADtree, respectively). All algorithms used in this study are supervised classifier, a model does thus need to be trained on a labeled training dataset (training mode) and can thereafter be used to predict new examples without further parameter tuning (prediction mode). Casting the problem in a binary classification framework, we refer to each peptide-spectrum match as an instance (in the dataset), with the i^{th} instance consisting of a feature vector $x_i \in [1 \times n]$ and a label $y_i \in \{0, 1\}$, with n denoting the feature count. All algorithms described construct a function, $g(x)$, that minimizes the empirical risk of misclassifying an instance, under the assumption that all instances are drawn with respect to the same (unknown) probability distribution. In the following we limit ourselves to describing conceptual details of the utilized algorithms, referring the reader to cited works for technical details.

C4.5 and Willow tree are both decision trees algorithms iteratively growing a classifier tree by finding splits of the dataset with respect to the feature value which results in the greatest gain in Shannon entropy (a function used to quantify how homogeneous the instances reaching a certain leaf node in a tree classifier are with respect to instance label). The procedure halts when all instances in a leaf node are of the same class or a pre-defined stopping criterion has been reached.

We apply two so-called meta-classifier techniques to the above mentioned tree algorithms, namely AdaBoost (66) and Random Forest (25). Both work by training a collection of decision trees over iteratively modified versions of the original training set and combining the prediction power of these models into one superior ensemble-classifier. The AdaBoost procedure iteratively updates importance weights for the dataset instances for each tree model constructed during training. The distribution of weights is changed such that higher weight is given to instances misclassified in the previous iteration. The final classification of an instance is made by the majority vote on classes returned by the tree collection. In the case of Random Forest each tree is trained on a bootstrap sample of the available instance and each node split only considers a number m of the available features (where $m << n$). The final class label of an instance is assigned by taking the mode of the class labels returned by the constructed tree set.

The ADtree algorithm also utilized the AdaBoost technique, but unlike ABC4.5 and AB-Willow it has the advantage of producing models that are easily represented as a tree with a limited number of nodes (less than 20). This property is achieved by constructing a tree that is a conjunction of rules which all contribute real-valued evidence toward a given instance being

classified as either true or false. Unlike traditional tree models the classification of instances by ADtree is thus not determined by a single path traversed in the tree, but rather by the additive score of a collection of paths. The ADtree is graphically represented with two types of nodes: Elliptical *prediction nodes* and rectangular *splitter nodes* (see Figure 2 for an example). Each splitter node is associated with a value indicating the rule condition: If the feature represented by the node is less than or equal to the condition value for a given instance, the prediction path will go through the left child node, otherwise the path will go through the right child node. The final classification score produced by the tree is found by summing the values from all the prediction nodes reached by the instance, with the root node being the precondition of the classifier. If the summed score is greater than zero, the instance is classified as true.

In addition to providing a classification label, the tree score of an instance (the margin score) is a measure of confidence in the classification label, a feature that makes it possible to convert these into true probability estimates. To this end we use Logistic calibration (67), providing a one-to-one mapping between the marginal score and a probability estimate.

3.2.3 Evaluation Metrics

All instances will be classified into one of the following categories: True Positive (TP), False Positive (FP), True Negative (TN), or False Negative (FN). By determining the count of instances in each category, the following quality metrics can be estimated:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Sensitivity &= \frac{TP}{TP+FN} \end{aligned}$$

$$Specificity = \frac{TN}{TN+FP}$$
$$Net.prediction = \frac{Sensitivity + Specificity}{2}$$

Additionally, the area under the curve (AUC) of the receiving operator characteristic (ROC) is used to have a metric that is unbiased towards the class distribution of the dataset. The ROC is defined as the (1-specificity, sensitivity)-curve, with each point corresponding to a specific threshold for class separation. An AUC value of 1 corresponds to an error-free performance over the entire range of thresholds, whereas a random classifier achieves an AUC value of 0.5. In addition to the AUC measure we use Precession-Recall Curve (PRC) (precision = $\frac{TP+TN}{TP+FP}$ and recall=sensitivity) to judge whether a classifier is truly superior to another, as it has been shown that domination in ROC-space does not always result in superiority in PRC-space (46).

3.2.4 Availability and requirements

The MALDI TOFTOF dataset used for constructing and validating the procedure is publicly available on ProteomeCommons.org `https://proteomecommons.org/dataset.jsp?i=71683` (57; 58). The code for constructing the models presented is freely available as part of our in-house machine learning workbench, MALIBU (103), available at http://proteomics.bioengr.uic.edu /malibu/. MALIBU is used for both training and validation of the classifiers. All algorithm parameter tuning was done with standard settings for the MALIBU package(103).

3.3 Results

The developed machine learning protocol for identification of true peptide-spectrum matches was constructed in three steps: calculation of features for representation of each instance in the

dataset; construction of classification models based on the annotated instances; and evaluation and interpretation of the resulting models. In the following, we present the details of each step and describe a method for extending the developed protocol into a probabilistic protein identification method.

3.3.1 Feature Calculation

A summary of the features utilized in this work can be seen in Table I. We divide the features into three groups reflecting how they are derived. The *Sequest* group contains features that can be obtained from the output of the Sequest algorithm, such as the correlation score ($Xcoor$) between the theoretically calculated and experimentally obtained spectra and the difference between parent ion mass and database peptide mass, ($deltaMH$) amongst others (Sp, $SpRank$, $deltaCn$, $ionfrac$). As these values are well characterized elsewhere (90; 53) we will not go into further detail here.

The *Published* group contains features that have been used in previously published results on classifier construction for the problem at hand. The computations needed to derive these features are self-explanatory given the description in Table I. We will refer the reader to the study by Ulintz *et. al.* for further details on computation and the underlying intuition leading to the inclusion of these features (194).

The *Novel* group consists of features not previously included in other machine learning formulation of this classification problem, and includes features for quality assessment of the spectral data as well as probability measures specifying the likelihood of observing the entire dataset. Six novel features are calculated and their rationale is described below.

Group	Name	Meaning	Origin
SEQUEST	Xcoor	Rank score from the SEQUEST search.	SEQUEST
	deltaMH	Difference between mass of parent ion and identified peptide mass.	SEQUEST
	deltCn	Difference between Xcoor of the highest ranked peptide and the peptide in question	SEQUEST
	SP score	Preliminary score of peptide in search procedure	SEQUEST
	SP rank	Initial rank of peptide based on SP-score	SEQUEST
	Ion fraction	Percentage of ions in the mass spectra that could be correlated with the spectrum	SEQUEST
Published	Number of tryptic	Number of tryptic cleavage sites in the peptide targets (NTT)	Calculated
	Peptide lenght	Residue count of the peptide	Calculated
	Summed Intesity	Sum of peak intensities in the spectra	Calculated
	Mobil proton factor (MPF)	Measure of the proton mobility in peptide	Calculated
	C-terminal Residue	Amino acid residue at c-terminal (Arg=1, Lys=2, Other=3)	Calculated
	Mass-window peptides	# of DB peptides within prespecified mass-window mass-window of the parent ion	Calculated
	Proline count	# of Pro residues in the peptide	Calculated
	Arginine count	# of Arg residues in the peptide	Calculated
Novel	Intensity Mean	The mean of the peak intensities	Calculated
	Intensity Std.	Std. of the peak intensities	Calculated
	Intensity bins	The distribution of intensities in 20%-bins	Calculated
	Protein Hit Count (PHC)	Probability score of observing x number of peptides from parent protein	Calculated
	Potential Coverage Ratio	The potential sequence coverage	Calculated
	PTM percentage	The percentage of possible PTMs found in a peptide	Calculated

TABLE I: Features used in the machine learning formulation. For each individual feature we give a brief description and indicate whether the feature was obtained from the output of the SEQUEST algorithm or calculated from the identified peptide, the mass spectrum, or database statistics. The features have been divided up into three subgroups *SEQUEST*, *previous*, and *novel*, denoting those features that can be derived directly from the SEQUEST algorithm output, those used in previous studies of the identification problem, and those introduced in this work, respectively. A detailed explanation of the features can be found in the main text.

Intuitively, one would be more confident in identifying a borderline peptide-spectrum match as being true if other peptides from the same parent protein are observed in the search result. In other words, given prior knowledge, one would favor specific peptide-spectrum matches over others with similar correlation values, due to our overall knowledge of the search result. This intuition leads to the implementation of two novel features, namely the *Protein Hit count* (PHC) and the *Potential Coverage Ratio* (PCR).

We formulate the PHC as the following probability: Given a database containing a certain number of observable peptide D (with respect to the mass limitations of the instrument used for analysis, the digestion enzyme utilized, and the post-translational modifications specified in the database search) and a search result containing P samples from this database, we want to calculate the probability that k or fewer observations of a given protein would be made by randomly sampling from this database. For each peptide stemming from a protein that has been matched k times in a search we will specify the PHC by the binomial distribution, where n is the number of potentially observable peptides from this protein:

$$PHC = \sum_{i=1}^{k} \begin{pmatrix} \mathrm{D} \\ \\ \mathrm{P} \end{pmatrix} (\frac{n}{D})^i (1 - \frac{n}{D})^{P-i}$$

The above probability is estimated using a Poisson distribution and is reported in negative log-space in order to avoid numerical artifacts. Notice that since both the database size and the number of spectra are included in the calculation of the above term, any learning algorithm trained on a specific training set with given a database and a collection of spectra should work

equally well on datasets obtained from a different database size searched with a different number of spectra.

One concern that may be raised when utilizing information from the parent protein, as is the case with PHC, rather than the peptide-spectrum match itself, is how such features will handle the fact that some peptides can be mapped to several parent proteins due to the existence of orthologs and homologs in the database searched. One should, however, recall that a spectrum and a peptide fragment match provided by the Sequest protocol is always linkable to the specific parent protein that gave rise to the theoretical peptide-fragment matched to the spectra. Consequently there is never any doubt which parent protein the specific peptide-fragment should be counted towards. In fact, in instances where two proteins (one present in the sample and one not present) have a certain degree of sequence similarity the PHC may actually help weed out false-positive hits from distant homologs, as hits from such homologs will have a lower PHC (and PCR) than hits from the protein actually present in the sample.

The PCR is simply defined as the percentage of residues belonging to observable peptide fragments that are observed in the set of peptide-spectrum matches from the Sequest search. Further, we include the PTM percentage, which denotes the percentage of potential post-translational modifications (given the current search settings) included by Sequest to obtain the present correlation scores. The logic behind including the PTM percentage is as follows: PTMs are often functional modifiers of proteins. The need for the hypothetical inclusion of a high-percentage of the potential PTMs in a short peptide fragment in order to get a good correlation with the spectra at hand could indicate that the match is not a true-positive as it

seems unlikely to have a large number of functional modifiers close together in a relative short peptide-fragment.

Previous works (11; 192) have shown that an automated quality assessment of the spectral data can help validate peptide-spectrum matches by sorting out low quality spectra. The simplest features incorporating this notion are *Intensity mean* and *Peak count*, which specify the average intensity of all peaks in the raw spectrum and the total number of peaks, respectively. Both of these values are often used in human assessment of spectral quality (92) and have discriminatory power in sorting out spectra of poor quality (192).

3.3.2 Classifier Performance

We compare the performance of a collection of classification algorithms using datasets including different subsets of features. One set includes the *Sequest* and *Published* feature-groups from Table I and another one includes all features, referred hereinafter as the *S+P* dataset and *All* dataset, respectively. Each dataset is divided into a training set for classifier construction and parameter tuning (by means of cross-validation), and a distinct test dataset for evaluating the classifier performance. We choose to evaluate our method using a test set rather than by using cross-validation on the training set to ensure that dependencies between features from different instances within the dataset do not inflate the performance metrics (this concern is particularly relevant for the PHC feature).

Table II shows the performance of a number of classifiers on the *S+P* and *All* datasets. The high ratio between negative and positive instances in the datasets means that accuracies correlate strongly with prediction performance on negative cases. Consequently, the accuracy

Feature groups	Algorithm	Accuracy	Sensitivity	Specificity	AUC ROC	Net pred.
All	ABWillow	0.97505	0.56504	**0.99385**	0.96379	0.77945
	ABC4.5	0.97361	0.58815	0.99269	0.94821	0.79042
	RFC4.5	0.97276	0.57212	0.99259	0.87901	0.78235
	ADtree	**0.97688**	**0.7248**	0.98988	**0.96923**	**0.86118**
SEQUEST	ABWillow	0.96951	0.48762	0.99336	0.90723	0.74050
	ABC4.5	**0.97258**	0.57018	0.99250	0.907084	0.78139
Published	RFC4.5	0.97228	**0.58961**	0.99122	**0.912744**	**0.79042**
	ADtree	0.96925	0.48762	**0.99310**	0.90604	0.74032
-	PeptideProphet	0.9688	0.54	0.99	-	0.765

TABLE II: Validation metrics for a collection of machine learning algorithm runs over testsets containing feature from the groups denoted in the feature table

and specificity metrics, which for both datasets are well above 98%, are not instructive for comparing the performance. Better comparison can be made with Net Prediction and AUC, as they are insensitive to skews in class distribution. Gauging these metrics, it is clear that the novel features introduced in this work provide added discriminatory power between true and false instances. The best performance is achieved by the ADtree algorithm with the *All* dataset, giving a 6% higher AUC ROC than the best performing algorithm in an *S+P* dataset. When comparing the performance of the same algorithm on the two dataset, we observe that 3 out of 4 algorithms perform better on the *All* than on the *S+P* dataset, a fact that is also clearly illustrated in the ROC curves in Figure 11 (left). Here we observe that the ADtree and ABwillow algorithms applied to *All* dataset outperform all other classifiers over the entire range of False Positive Rates (FPR), whereas the ABC4.5 on the *All* dataset falls somewhere in between these two and the results from classifiers trained on the *S+P* dataset. In addition, all

classifiers trained on the *All* dataset perform better than the PeptideProphet procedure over the entire FPR range. When comparing the classifiers trained on the *S+P* feature collection to the PeptideProphet result, the picture is not as clear. As can be seen on the enlargement in Figure 1 (left), the machine learning algorithms do in general (regardless of feature set) perform better than PeptideProphet at lower FPRs, while PeptideProphet gives better sensitivity at higher FPRs (Note, the high FPR range is rarely used in real applications). We also note that the results obtained on the *S+P* dataset containing the same features as utilized by Ulintz *et al.* closely match the result reported on a preliminary version of mass spectral data used in this study (194). The PRC depicted in Figure 11 (right) offers an alternative view of classifier performance. The plot does not allow for judgment of which algorithm does better on a specific dataset, as all show strengths and weaknesses at different recall values. It is, however, clear that all algorithms trained on the *All* dataset do better than the ones trained on the *S+P* dataset, confirming the discriminatory power of the new features introduced in this work.

3.3.3 An Interpretable Model

As observed above, the ADTree algorithm is among the strongest performers on the dataset incorporating all features, rivaled only by ABWillow tree. In comparison with machine learning algorithms such as SVM, the ADTree algorithm provides the advantage of being represented as a collection of user interpretable rules. Figure 12 shows a graphical representation of the ADTree model learned from the *All* dataset (see **Methods** for how to interpret the tree).

The base-rules in the tree are numbered in accordance with their order of discovery (the number indicated in parenthesis after each feature name), which can be interpreted as the rule

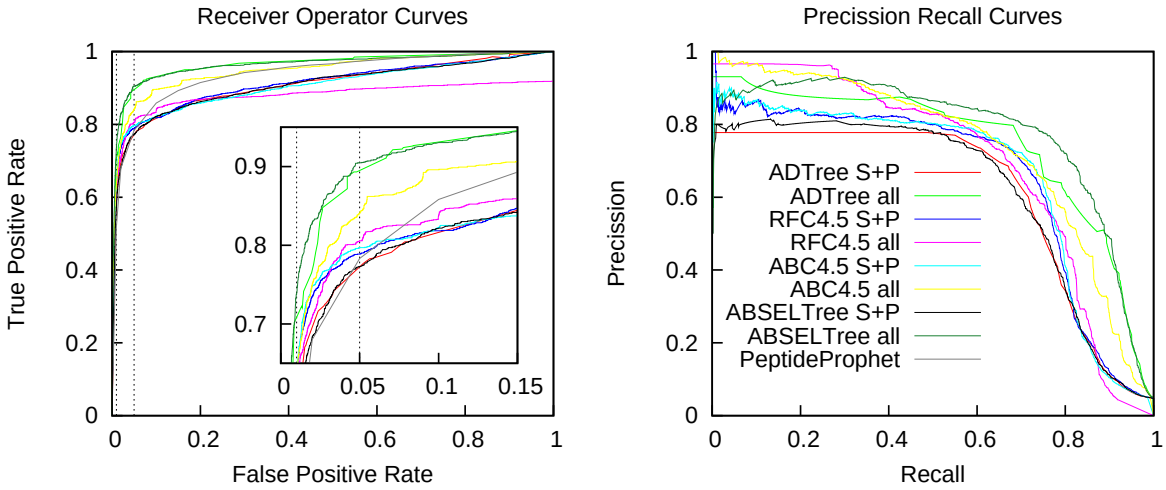


Figure 11: *Receiver Operator Curves (ROC)* (*left*) and *Precision/Recall Curves (PRC)* (*right*). Classifiers trained with the novel set of features have the suffix *all*, otherwise the suffix *S+P* is used (this does not apply to the curve for PeptideProphet shown in the ROC plot). The ROC shows how the TPR varies with the FPR, indicating what percentage of true hits one can expect to obtain at a given false-positive-rate. The PRC given an alternate view of the classification depicting the precision as a function of the recall (note PeptideProphet results only shown in ROC).

importance or predictive power of the feature (65). Given that interpretation, surprisingly, the PHC appears to have the strongest discriminatory power amongst true and false instances in that a $PHC > 15.9$ adds significant weight towards a positive prediction (the final faith of an instance satisfying this rule is of course also based on the other base rules involving PHC). Thus the learned model suggests that a higher than expected number of peptides from one protein in the Sequest search result, is indicative of these peptide-spectrum matches being true hits. The second and third base rules discovered are cut-offs for the *XCorr* score and *deltaCn*, two of

the main attributes of the Sequest algorithm used for judging how well the theoretical peptide spectrum correlates with the experimentally obtained spectrum. (It should, of course, be noted that except for rules with the root node as parent, the prediction bias of a rule should always be seen in context of its parent node(s)).

To better interpret the possible paths traversed by a dataset instance, subsets of base-rules have been highlighted in color in Figure 2. We will now examine these paths more closely to see how the classifier is able to discover meaningful knowledge, while at the same time providing high accuracy classification results.

The blue path is made up only one feature, namely NTT. If the peptide has at most one missed cleavage side, this provides evidence toward the hit being positive, though an instance could still ultimately be classified as a false hit. If we examine the red path we see that a $PHC < 15.9$ is negative evidence towards the hit being true, as it is unlikely that we would observe only a few peptides from a protein that is indeed present in the sample. If the peptide-spectrum match does, however, have a strong correlation score ($XCorr > 2.45$) this effect is reversed, giving the path a net positive score. Interestingly, an $XCorr$ score lower than the 2.45 threshold does not add significant evidence toward the match being false. Thus high $XCorr$ scores add evidence towards an instance being true, while scores below constitute a borderline region where other factors determine the faith of the instance.

The fact that the PHC feature only comes into play when the $XCorr$ score is below as certain threshold is an important model feature, as the PHC score might otherwise "hurt" the classification of proteins with few "mass specable" peptide-fragments. The PHC is, in other

words, not used unless the quality metrics correlating the spectrum and the proposed peptide do not provide sufficient evidence to conclusively determine whether the peptide-spectrum match is correct. In situations where there are only one or two "mass specable" peptides from a protein one would want the quality metrics of matches to be highly confident when using them to identify the parent protein, the strategy learned by the model is thus reasonable when handling such instances.

A related mechanism is observed when following the green path, here $deltaCn$ values of at least 0.05 add evidence toward the instance being a true hit. The following $XCorr$ filter shows that correlation values below 1.71 are strong evidence towards the instance being negative, values above this threshold do not add evidence towards the instance being positive. The yellow path does not add any new features to the classifier, but simply acts as a further filter on the PHC feature, constructing intervals with increasing summed evidence towards the instance being positive. The purple path, on the other hand, adds two new features. If the instances following this path has a $deltaCn < 0.05$, and at the same time an $IonFrac$ value of less than 20.8%, there is substantial evidence towards the instance being false, whereas higher IonFrac values are indicative of a true instance when combined with a low $deltaCn$. In other words, small differences in the mass of parent ion of the mass spectrum and the theoretical mass of the peptide that it has been matched with is a strong indicator of a true hit only if a certain fraction of the spectral peaks are accounted for by that specific peptide.

We observe that none of the features intended to address the issue of spectral data quality were found to be instrumental in significantly improving the classification accuracy for the

ADtree model. This is somewhat surprising, since well above 85% of spectra were considered to be of poor quality in studies addressing the problem of identifying such cases (11). Thus, one would expect that a feature identifying such spectra would provide certain discriminatory power. One possible explanation for this observation is that these cases are already covered by other rules from the ADtree, thus including the spectral quality feature to the model would not add additional predictive power. For instance, one might reason that cases with inferior spectral quality will only give rise to database hits with low $XCorr$ score, which would render these cases false hits due to this feature.

The rules discovered above using the ADTree agree well with expert criteria previously used as conservative estimates for identifying hits that would be true with high probability. Washburn *et al.* (200) did, for instance, settle on the following conjunction of rules as criteria for correct hits: $XCorr > 2.2$, $deltaCn < 0.1$ and the peptide has to be fully tryptic (meaning $NTT = 0$). The classifier developed is comprised of rules with similar cut-off values for the features used by experts, but does also utilize novel rules when making predictions, identifying true instances that would otherwise have been missed. Take for instance the $XCorr$ cut-off: We found that values above 2.45 provide strong evidence towards an instance being a correct match. If the value, on the other hand, is below this cut-off we did not find it to be significant evidence toward the hit not being correct unless the value fell below 1.71, providing room for a number of borderline instances that can be correctly classified using the additional features in the model.

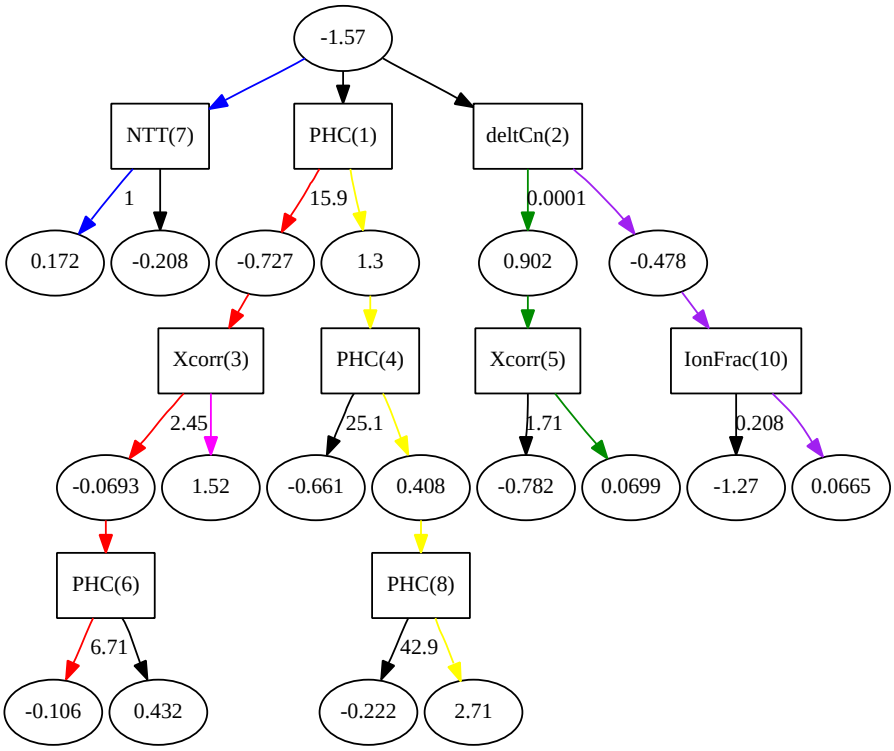


Figure 12: Graphical representation of the alternating decision tree learned from the dataset *all*. Prediction nodes are represented by ellipses and splitter nodes by rectangles. Each splitter node is associated with a real valued number indicating the rule condition, meaning: If the feature represented by the node is less than or equal to the condition value the prediction path will go through the left child node, otherwise the path will go through the right child node. The numbers behind the feature names in the prediction nodes indicate the order in which the different base rules were discovered, this ordering can to some extend indicate the relative importance of the base rules. A detailed explanation on how to interpret the ADTree is given in the main text along with a discussion of the colored paths outlined.

3.3.4 Extending the Peptide Prediction Protocol to Protein Prediction

The ultimate goal of MS/MS experiments is not necessarily the confident identification of peptides, but rather determining a probability measure for the presence of their parent proteins in the sample analyzed. One software application addressing this issue is the ProteinProphet software (136) by Nesvizhkii *et al.*, which identifies a minimal set of proteins accounting for the observed peptides by use of the expectation-maximization algorithm. Following the formulation in this work we show a straight-forward way of extending our peptide identification protocol to a protein identification protocol.

A conservative estimate of the probability, P, that a given protein is present (meaning that at least one of the peptide matched by the database search from this protein is correct) is given by

$$P = 1 - \prod_i (1 - \max_j p(+|D_i^j)) \tag{3.1}$$

where the product index i is over all distinct peptides from this protein, the index j is over all matches obtained for one specific peptide and $p(+|D_i^j)$ denotes the probability of the j^{th} identification of peptide i being a true match. We take the maximum overall identification from all identical peptides, as these should not be considered independent. Note further that this formulation theoretically allows for a specific peptide to be considered as evidence for two distinct proteins.

By use of logistic calibration we convert the classification scores obtained from the ADtree algorithm into probability estimates of a given peptide-spectrum match being correct (or in

other words we estimate $p(+|D_i^j)$). Combining these estimates with (Equation 3.1) we can calculate probability estimates for each protein that has at least one peptide identified in the database search actually being present. Using this relatively simple extension of the classification framework we are able to identify 87% of the proteins present in the sample at false positive rate 5%. In comparison, using the probability estimates from PeptideProphet only achieved an identification rate of 85%. This is not surprising as we previously saw that the ADtree procedure identifies more correct peptide-spectrum matches than PeptideProphet.

3.4 Discussion and Conclusion

Supervised machine learning provides an attractive platform for examining the peptide prediction problem since no prior assumptions of the distribution of the utilized features have to be made when constructing the model. This is in contrast to generative/unsupervised models such as the PeptideProphet procedure, that assumes specific distributions (in this case Gaussian and Gama distributions) when classifying matches. While it has been shown that the assumption regarding a specific data distribution is reasonable (90; 136) in certain instances of the identification problem there is no general evidence or theoretical framework supporting this claim for all types of instrument or data. As conveyed in this work another attractive property of the supervised machine learning framework is the relative ease with which the developed models can be extended with novel features in order to improve predictive power. It thusly becomes possible to construct a tailored peptide identification framework for specific experimental procedures and equipment choices, thereby providing stronger guarantees on the control of error rates than would be possible with a generic setup. One drawback of the supervised learning

approach is of course the need to construct a training dataset from a known protein sample to do the initial parameter tuning of the model and determine the performance metrics. However, once the training is done the trained model will perform equally well on large scale and sparse datasets, since one does not have to be concerned with having too little data to properly estimate model parameters.

Since large-scale proteomics studies are often concerned with characterizing the proteomic make-up of the cell in a number of states, a reliable probabilistic measure for the presence of a given protein is essential. Above we demonstrated how predictions from the ADtree model (or any other supervised learning algorithm providing marginal classification scores) combined with logistic regression can be used in a simple probabilistic framework to give a high protein identification rate at a low FPR.

In sum, we have improved on previously published machine learning procedures for identification of correct peptide-spectrum matches by introducing novel features adding to the predictive power of all the tested algorithms. Furthermore, we have introduced the ADtree procedure into the problem domain, constructing an interpretable model that correlates well with previously published rules addressing the classification problem at hand. Finally, we show how the protein prediction problem can be addressed within the presented framework.

In this work we demonstrate how a generic classification model for MS/MS data obtained by use of the MALDI ionization can be constructed. In future work, we intend to extend the classification framework to take advantage of experiment specific parameters (ionization method,

instrument type, pre-processing steps of the sample) creating models tailored specifically to the instrumental set-up used to obtain the spectral data.

CHAPTER 4

GENOME-WIDE CHARACTERIZATION OF LIPID-BINDING IN SIGNAL TRANSDUCTION: A CASE STUDY OF THE PDZ DOMAIN FAMILY

In Chapter 3 we presented a method for probabilistic identification of the protein entities making up proteomes from mass-spectrometry data. In this chapter we will turn our attention to functional classification of the domain components making up the proteins in the proteome. Specifically we present a computational protocol for genome-wide identification of protein interaction properties with membrane.

The work presented in this chapter was done in collaboration with Yong Chen and Ren Sheng who carried out all wet lab experiments under the supervision of Professor Wonhwa Cho, while all computational protocols were developed by Morten Källberg. Here we will mainly focus on the details of the computational protocols and conclusions reached from theoretical calculations while referring the reader to the original publication for details on experimental procedures (32).

4.1 Introduction

Regulation of cellular processes, such as cell signaling, is driven by a large range of protein-protein interactions which are in many instances mediated by modular protein-interaction domains (PIDs), such as SH2, SH3, PDZ, and WW domains (18; 148; 149). Cellular membranes constitute a unique local environment for protein-protein interactions and therefore serve as the

main sites for protein complexes and networks (23; 35). Accumulating evidence suggests that membrane lipids are key in protein complex formation or networking through direct interactions with signaling proteins (35; 203). Membrane recruitment of cellular proteins is mediated by lipid-binding domains or motifs through selective lipid interaction or non-specific interact with the anionic membrane surface (36; 49; 110). Consequently, it is generally believed that the interaction of proteins at the membrane involve the coordinated effort of distinct lipid-binding domains (or motifs) and PIDs in the same molecules (110; 145). Novel studies do, however, indicate that PIDs, such as PDZ domain (60; 217) and PTB domain (160; 216), can interact directly with membrane lipids and facilitate both protein-protein and protein-lipid communication. In addition, it has been reported that some lipid-binding domains, such as the PH domain (211) and the PX domain (108), can interact with proteins as well as lipids. This observation indicated that PIDs and lipid-binding domains could possibly act as dual-specificity lipid- and protein-binding modules which are key in protein networking. To test this hypothesis, we have developed new experimental and bioinformatics tools to identify and describe dual-specificity PIDs on a genomic scale and applied these tools to the study of PDZ domains.

The PDZ domains is ≈90 amino acid long modular PID which interacts with a 5-12 residue C-terminal sequence of its target protein(s) (60; 177). The domain family was first found in three unrelated proteins, postsynaptic density 95 (PSD95), disc large 1 (DLG1), and zonular occludens 1 (ZO1), and has since been found in a large number of proteins. A SMART search (174) returns 148 human proteins with more than 500 different PDZ domains, thereby making them one of the most common PIDs in vertebrates. Most PDZ domain host-proteins display

multiple PDZ copies, and can thus be considered prototype scaffold proteins that reversibly interact with multiple binding partners to coordinate signaling complex formation facilitating networking (60; 177). It has been demonstrated that PDZ domains can interact with negatively charged model membranes and that, in certain instances, this PDZ-membrane interaction is important for the cellular function of their host proteins (130; 144; 204; 217). It is, however, still unknown if lipid binding is a general property of PDZ domains, and if they can act as dual-specificity modules as part of a biological system. It should thus be clear that the PDZ domain family is a good candidate for a first study of the genome-wide identification and characterization of dual-specificity PIDs.

4.2 Methods

4.2.1 Dataset

Table III summarized the key properties of the dataset used for training the PDZ classification protocol. Specific details on the construction of the dataset are given in Section 4.3

TABLE III: The binding affinity for the 70 experimentally tested PDZ domains used for training the classification model. *Start/Stop* denotes the first and last amino-acid belonging to the domain within the host protein, *Domain number* (#) indicated the domain location in the sequence relative to other PDZ domain, K_d is the mean ±SD binding affinity determined by SPR, *Selectivity* indicates specific lipid selectivity, Wu *et al.* indicated binding results reported in (204), *Structure* column denotes the PDB-identifier for the structure data used in feature calculation.

Gene	**#**	K_d**/nM**	**Start/stop**	**Selectivity**	**Org.**	**Wu** *et al.*	**Structure**
NHERF-1	1	20 ±1	14-91	low	rabit		-
DVL2	1	33 ±3	267-352	$PI(4,5)P_2$	mouse		-
DVL1	1	45 ±6	245-337		human		-
DVL3	1	50±5	243-335	$PI(4,5)P_2$	human		-

Continued on Next Page

Table III – Continued

Gene	**#**	K_d**/nM**	**Start/stop**	**Selectivity**	**Org.**	**Wu** *et al.*	**Structure**
Tamalin	1	90 ±8	100-186	low	mouse		-
SAP102	3	140 ±5	404 -482	$PI(4,5)P_2$ $PI(3,4,5)P_3$	rat	No binding	-
LNX1	4	180 ±40	638-721	low	mouse		-
PDZK2	3	280 ±50	263-343	low	mouse		-
MAGI-1	5	290 ±10	998-1091	low	human		-
PDZ-GEF	1	290 ±32	385-470	low	human		-
β2-syntrophin	1	320 ±80	115-195	low	human		2vrf
PDZK2	2	320 ±32	151-255	low	mouse		-
nNos	1	340 ±10	17-96	low	human		-
PSD95	3	390 ±30	313-391	low	rat	No binding	1tq3
INADL	6	480 ±190	1068=1160	low	human		2ehr
Chapsyn110	3	510 ±50	421-499	low	rat		-
γ2-syntrophin	1	530 ±140	73-153	low	mouse		-
Harmonin	1	600 ±70	87-165	low	mouse		-
MAGI-3	5	610 ±190	1021-1100	low	human		-
SAP97	3	620 ±70	465-543	$PI(3,4)P_2$	rat	No binding	2i0i
LNX2	1	670 ±100	232-314	low	mouse	No binding	-
MAGI-2	3	750 ±170	605-683	low	human		1ujv
α1-syntrophin	1	860 ±70	81-161	low	mouse	Binding	1z86
MAGI-2	5	900 ±170	920-1007	low	human		1uew
PSD95	2	930 ±120	160-244	low	rat	No binding	1qlc
PDZ-PhoGEF	1	950 ±110	47-120	low	human		2dls
LNX1	1	960 ±120	278-360	low	mouse		-
ZO-1 PDZ-2	2	980 ±200	186-261	$PI(3,4)P_2$ $PI(4,5)P_2$ $PI(3,4,5)P_3$	mouse		
INADL	5	1070 ±110	686-772		human		2d92
β1-syntrophin	1	1440 ±180	538-613		human	Binding	-
Rhophilin-1	1	1440 ±160	111-191		mouse		-
Syntenin1	1	2200 ±250	100-195		human		
SAP102	1	4980 ±870	149-233		rat		-
PSD95	1	-	65-149		rat	No binding	1iu2
MAGI-2	2	-	426-492		human		1ueq
MAGI-2	4	-	778-859		human		1uep
SAP97	1	-	224-308		rat		1zok
Spinophilin	1	-	496-581		rat		2g5m
Neurabin	1	-	505-590		rat		2fn5

Continued on Next Page

Table III – Continued

Gene	**#**	K_d**/nM**	**Start/stop**	**Selectivity**	**Org.**	**Wu** *et al.*	**Structure**
NHERF-2	2	-	150-230		human		2he4
NHERF-2	1	-	11-88		human		2ocs
SAP97	2	-	318-402		rat		2awu
CAL	1	-	288-368		human		2dc2
PDZK1	3	-	243-320		mouse		2d90
PDZK1	1	-	9-87		mouse		2edz
MAGI-1	1	-	295-401		human		2ysd
MAGI-1	3	-	643-720		human		3bpu
PTPN3	1	-	510-595		mouse		-
MALS-1	1	-	108-187		human		-
E6TP1	1	-	953-1025		human		-
LNX1	3	-	508-591		mouse		-
LNX1	2	-	385-465		mouse		-
Densin-180	1	-	1403-1493		rat		-
MAGI-2	1	-	17-98		human		-
MALS-3	1	-	93-172		mouse		-
LNX2	2	-	338-418		mouse		-
MAGI-1	2	-	472-554		human		-
PDZK2	4	-	394-472		mouse		-
Harmonin	2	-	211-289		mouse		-
MAGI-3	1	-	410-476		human		-
MAGI-3	3	-	726-807		human	No binding	-
MAGI-3	4	-	851-935		human	No binding	-
PDZK1	2	-	128-215		mouse		-
PDZK2	1	-	85-177		rat		-
MUPP1	6	-	996-1077		mouse		-
MUPP1	7	-	1139-1231		mouse	No binding	-
MUPP1	8	-	1338-1421		mouse	No binding	-
MUPP1	12	-	1847-1933		mouse	No binding	-
MUPP1	13	-	1972-2055		mouse	No binding	-
MAGI-3	2	-	578-641		human		-

4.2.2 Feature development

Previous analysis of membrane binding mechanisms suggested that certain common physical properties enable the domain targeting of membranes. In particular, properties such as the

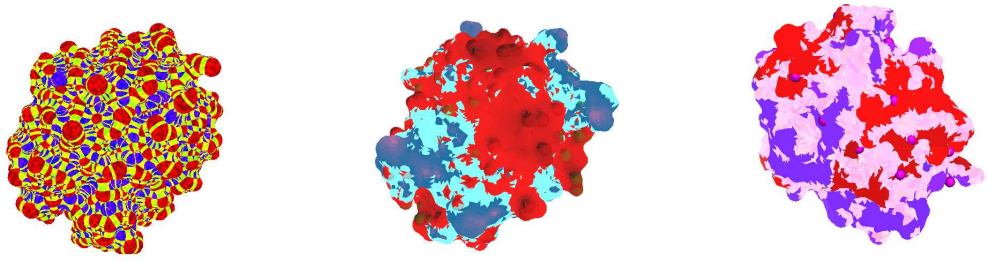
non-specific electrostatic attraction between anionic membranes and cationic surface residues, association of hydrophobic surface residues with the membrane hydrocarbon core, and the specific interaction between key residues and lipid head-groups have been found to be of major importance (35; 36; 82). In the following we will create a method for quantification of these properties into a vector of numerical feature values, for use in the construction of machine-learning protocols.

Two distinct sets of features are developed: A set representing structural information and one that is solely based on sequence statistics. The definition of each is outlined in the sections below.

4.2.2.1 Surface patch definition

The characteristic properties of the structure are captured by identifying continuous regions of the solvent exposed surface, so-called surface patches, defined by physical or chemical quantities (electrostatic potential, hydrophobicity etc.) common to the this specific area. The steps of patch growing detailed below are outlined in Figure 13. The basic idea is as follows: First, the surface is defined as a collection of neighboring triangles (see 13(a)), second, a numerical representation of the quantity of interest is associated with each triangle (see 13(b)), and finally the patches that are most highly correlated with the function of the structure are defined (see 13(c)).

By using the definition of solvent-excluded surface (SES) in (170), the topological boundary defined by the Van der Waals radius of the atoms in the structure of interest is determined by use of the MSMS algorithm developed by Sanner (170). The final SES is expressed by a



(a) The solvent-excluded surface as calculated by the MSMS (Maximal Speed Molecular Surface) algorithm, the coloring of the triangles representing the surface represent the number of probe contact made.

(b) Mapping of the electrostatic potential onto the surface, dark-red regions indicate higly negative potential values while dark-blue regions indicate highly positive potential values.

(c) The patches grown on the surface with parameter $C = 60$, red regions indicate patches with negative potential value, purple regions indicate patches with positive surface value, and pink regions indicate location with mixed potential values.

Figure 13: The three steps in determining surface patches for a certain quantity for protein structure PDB-id 1a53, here illustrated with the electrostatic potential of the structure.

triangulation procedure and thus results in a collection of neighboring triangles representing the molecular surface.

For now let's assume that each triangle on the surface is associated with a numerical value corresponding to the quantity that forms the basis for patch growing. We will denote this value for a triangle t by $t.val$ and the distance between the centroids of triangles t_1 and t_2 by $dist(t_1, t_2)$. Furthermore, $t.neigh$ will denote the neighbor triangles of t, meaning those that share an edge with t, and $t.included$ will be a boolean flag indicating whether a given triangle has been included in a patch. The collection of patches is then found by repeating the following recursive procedure until all surface triangle have been included in a patch: Choose a random triangle that has not yet been included in a patch and extend the patch from here according

```
GROW-PATCH(Seed triangle T):
 for t1 in T.neigh:
   if NOT t1.included AND
   |t1.val - T.val|/dist(T,t1)<C:
       Add t1 to the current patch
       t1.included = TRUE
       GROW-PATCH(t1)
```

Figure 14: Pseudocode for the patch growing procedure.

to the procedure outlined in the code below, repeat until all triangles have been included in a patch (see Figure 14).

The constant C in the `GROW-PATCH`-method is used to determine if a patch should be extended in a given direction. An appropriate value for C needs to be set for each patch type of interest. For this application the C-value was fixed manually by simple visual inspection of the patches on the molecular surface. A value of $C = 50$ was chosen.

4.2.2.2 Mapping quantities onto the surface

In order to do the patch-growing we need to assign values from the quantity of interest to each triangle on the SES. Here we give examples of how this can be done for both spatial and residue/atom based data.

(1) The electrostatic potential of a structure can be calculated by solving the Poisson-Boltzmann (PB) equation numerically using a finite difference scheme as implemented in APBS (6). The spatial potential values are mapped onto the surface by taking a weighted average of

the 8 discrete data points closest to the point 1 A from the triangle surface in the direction of its normal vector. (2) Hydrophobicity values are assigned to the surface based on the Kyte-Doolittle value of the amino acid that gave rise to the triangle of interest (99). (3) Hydrogen-bonding is mapped to the surface by determining if an atom is capable of forming a hydrogen bond, indicated by setting $t.val = 1$.

We include the following properties of the five largest positive electrostatic patches found as features: Size, maximum/minimum potential value, and average potential value. Cumulative size of the two largest, three largest, four largest, and five largest patches are included as features. The hydrophobic patches are derived using the procedure specified above, and features similar to those derived from the electrostatic patches were used in classification. Further, the surface propensity of the 20 amino-acids is found by identifying the residues having at least 30% solvent exposure as defined by the DSSP (86) procedure are included.

4.2.2.3 Sequence feature - functional classification matrix

In addition to the patch-growing procedure we also construct a feature solely based on statistics from a collection of domain sequences. We developed an approach obtaining a score for each sequence by its similarity to other sequences in the dataset, this is done using a recursive functional classification (RFC) matrix inspired by Park *et. al* (146). A multiple sequence alignment of all domain sequences (both labeled and unlabeled) is created using a Hidden Markov Model-profile (for the PDZ domain the PFAM model PF00595 is used) as this procedure has been found to give a better alignment of structural elements than classical alignment methods. Based on the alignment we can calculate the probability of observing amino

acid a at location i in the alignment. Denoting the probability for binding and non-binding cases by $P_{a,i,+}$ and $P_{a,i,-}$, respectively, each entry in the the RCF matrix is given by:

$$RCF_{a,i} = \log\left(\frac{P_{a,i,+}}{P_{a,i,-}}\right)$$

Thus a positive/negative entry in the matrix indicates that the presence of amino acid a at location i is evidence towards the domain being membrane binding/non-binding. We can summarize the evidence for a giving domain sequence S as being binding in the following score:

$$\text{RCF-score(S)} = \sum_{s_i \in S} RCF_{s_i,i}$$

When the RCF-score for a sequence is calculated, we do not include that sequence in the calculation of the scoring matrix, furthermore a pseudo-count strategy is used starting out with a distribution reflecting the overall amino-acids propensity in unrelated proteins.

4.2.2.4 Classifier: SVM and AdaBoost on C4.5

The Support Vector Machine (SVM) methodology first proposed by Vapnik (44) facilitates the derivation of a classification hyperplane (a hyperplane separating positive and negative cases) for non-linear problems by working in a vector-space of higher dimension than that of the original feature space (using the so-called kernel-trick) (44). In this work, we tested Gaussian, Sigmoid, and polynomial kernel functions and found that the Gaussian gave the best results for this application. Thus, we exclusively used SVM with the Gaussian kernel function. The other learning procedure used in this work is the C4.5 decision tree algorithm developed by

Quinlan (159) combined with the AdaBoost technique for improving the overall classification power of a collection of weak classifiers (a classifier with performance just slightly better than random guessing) (173) as first proposed by Freund and Shapire (66). The combination of these techniques is referred to as ABC4.5. A single decision tree was constructed through a greedy procedure which iteratively finds splits of the dataset with respect to the feature value that results in the greatest information gain, as defined by Shannon entropy. The Adaboost algorithm iteratively constructs a collection of decision trees with each tree being learned on a different weighting of the instances in the original dataset. After each learning cycle the weight distribution on the dataset was updated in such a manner that higher weight is given to instances misclassified in the previous iteration. The final classification of an instance was made by the majority vote of the tree collection. We used our in-house machine learning workbench MALIBU for the construction and validation of models, giving a uniform interface for comparison and analysis of their performance (103).

4.2.2.5 Classifier evaluation

We measured the performance of the constructed classification models using the following metrics: Accuracy (Acc) defined as the ratio of true prediction to the total number of prediction, Sensitivity (Sen) defined as the percentage that a true example is classified as true, Specificity (Spe) defined as the percentage that a negative example is classified as negative. The classification result of an instance in a binary classification can fall into four categories: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). Using these

counts the three metrics are approximated by $Acc = TP + TN/(TP + TN + FP + FN)$, $Spe = TN/(TN + FP)$, and $Sen = TP/(TP + FN)$.

Additionally, we used the area under the curve of the receiving operator characteristic curve (AUC), a metric having the advantage of being insensitive to the class distribution of the dataset. The AUC is defined as the area under the (1-specificity, sensitivity)-curve, with each point corresponding to a specific threshold for class separation. The larger the AUC (between 0 and 1) is, the better the prediction method. A classifier having an AUC-value 1 performs perfectly over the entire range of threshold values, whereas a random classifier will only achieve an AUC-value 0.5. To provide a benchmark for the expected performance on unseen data we performed several rounds of training and evaluation of each classifier using n-fold Cross Validation (n-CV). In n-CV, we randomly divided the original dataset into n equally sized bins, each classifier was then trained n times using n-1 of subsets. The omitted subset in each round was used for estimating the evaluation metrics of interest, the average of which was thus based on evaluation over all instances.

4.2.2.6 Homology modeling

The developed method relies on features calculated from both sequence and structure, there is, however, only a limited number of experimentally determined structures for PDZ domains available. We therefore have to rely on homology modeling methods in determining feature values for the majority of the domains annotated in this work. All domains used could be modeled based on a template with which it has at least 35% sequence similarity, thus the

resulting structures are believed to be reliable representations of the domains. For modeling we used the software Modeler version 9v4 maintained by Sali lab (63).

To ensure that the above assumption is correct we chose a set of five domains with known structure, created novel structures based on homology modeling, calculated features and classified the domains. For all cases we found that both features and classification values corresponded well (values vary less than 3% and classification labels were correct in all cases) to those of the experimentally determined structures, lending credibility to the method.

4.3 Results

A published study measured the binding of 74 PDZ domains to negatively charged vesicles using vesicle pelleting assay (204). While this work indicated the affinity of PDZ domains for lipids, the qualitative nature of the data makes the systematic analysis of membrane-binding properties of PDZ domains and of the interplay between their membrane and protein interactions difficult. Consequently, it was necessary to collect a highly curated database of a sufficient size for statistical and systematic analysis. All reported membrane-binding PDZ domains interact with negatively charged membranes with low or no lipid head-group specificity (130; 144; 204; 217). Further, most of PDZ domain proteins interact with protein partners that are associated with the PM (60; 177) of which the cytosolic layer is highly negative due to the presence of phosphatidylserine (PS) and phosphatidylinositiol-4,5-bisphosphate (PtdIns(4,5)P_2) (36; 129). Given this observation we choose the vesicles having a lipid composition mimicking that of the inner PM (i.e., PM-mimetic vesicles) (36) as a model membrane and rigorously determined the K_d values for 70 monomeric PDZ domains from 35 different mammalian proteins

by surface plasmon resonance (SPR) analysis (34). We primarily choose uncharacterized PDZ domains (i.e., 51) in this work but do also revisit some previously characterized PDZ domains (i.e., 19) (204) in order to compare the results from the two competing experimental methods.

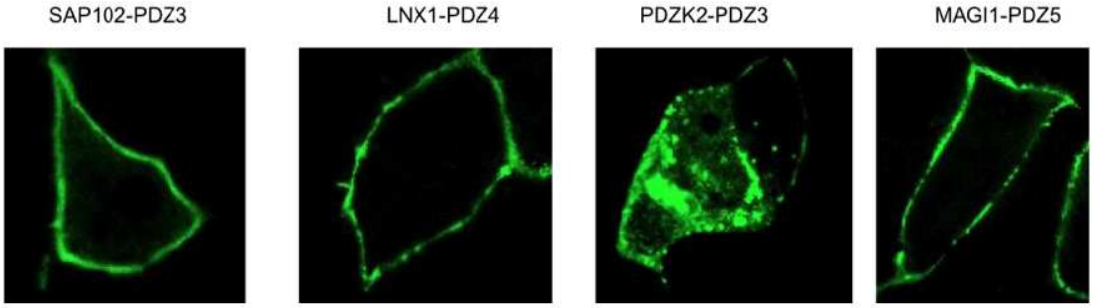


Figure 15: Plasma membrane localization of high-affinity PDZ domains. Those PDZ domains showing the highest affinity for PM-mimetic vesicles were transiently expressed in HEK293 cells as C-terminal EGFP-tagged proteins. Experimental data collected by Yong Chen.

As shown in Table III, 28 out of 70 tested PDZ domains (≈40%) display submicromolar K_ds for the PM vesicles with the highest affinity being in the 10^{-8}M range, thus being similar to that observed in canonical lipid-binding domains (36). Synenin-1 PDZ domains (218) and the second PDZ domain of ZO-1 (130), both of which have been found to have significant physiological membrane affinity, show 1-3 μM K_ds under our experimental conditions (see Table III). These results indicate that membrane binding is a more common feature of PDZ domains than have previously been assumed and that it could potentially be important in cellular re-localization

(see Figure 15) and function. Table III further display a significant discrepancy between our measurements and those of previous works (204). Specifically, 5 (of 17) PDZ domains that had been indicated to be non-membrane binding were shown to bind PM-mimetic vesicles with K_d = 140-930 nM. For those PDZ domains with affinity measures in the submicromolar range for the PM vesicles, we also measured the selectivity for phosphoinositides (PtdInsP), the majority of PDZ domains did, however, not display measurable PtdInsP affinity.

4.3.1 Classification Model for Predicting Membrane-Binding

The presence of a large fraction of membrane-binding PDZ domains in our data set facilitated construction of a high-accuracy prediction model for other PDZ domains. The PDZ domain does in general display a large degree of sequence similarity (60; 177). Our data does, however, show that sequence similarity between any two PDZ domains does not nessacerily translate into similar membrane-binding properties, thus making it a poor indicator for classification and prediction purposes. This is in contrast to other lipid-binding domains, such as the FYVE domain, for which a good correlation between sequence similarity and relative membrane affinity was observed (21). Given this observation it was necessary to construct a more sophisticated model based on quantification of physical and chemical characteristics of the domains.

We developed a machine learning-based prediction method for membrane-binding domains that uses a numerical vector representation obtained from primary and tertiary structures of proteins as input features and use a number of machine learning classifiers (16). To apply this method to our current task of discriminating membrane-binding properties among highly

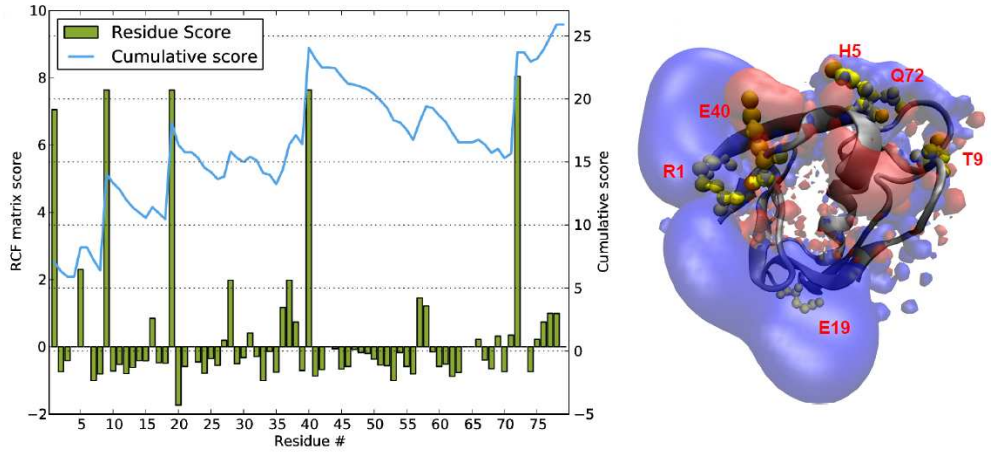


Figure 16: *(left)* The scores obtained from the RFC matrix for the PDZ domain PSD95. The green bars indicate the score for each residue in the domain while the light-blue line indicate the cumulative score across the domain sequence. *(right)* Structural representation of PDZ domain PSD95 with electrostatic isosurfaces at +/- 3 e/Kt marked in blue and red, respectively. The residues with the highest RCF score are marked in yellow.

homologous PDZ domains, we incorporated residue-specific features derived from the domain sequence data in addition to the protein-level features from structural data. Protein-level features enabling a domain to interact with membranes include nonspecific electrostatic attraction between anionic membranes and basic protein residues (132), association of hydrophobic protein residues with the membrane hydrocarbon core (184), and hydrogen bonds between key protein residues and lipid head groups (36). To incorporate residue-specific features, we determined the score of each residue and the cumulative score for a segment around it (146) by calculating the recursive functional classification (RFC) matrix. This statistical scoring approach helps identify residues that are more likely to be observed at certain positions in membrane-binding

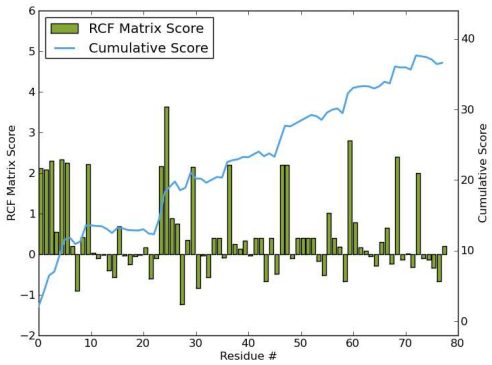
PDZ domains than in non-binding PDZ domains. Figure 16 depicts an example of this scoring procedure, displaying the score of each residue and the cumulative score for the neighboring segment of the rat PSD95-PDZ3 domain. Specific residues (i.e., R1, H5, T9, E19, D37, L38, S39, E40, and Q72) show strong correlation with membrane binding of the domain. Interestingly, these residues are not exclusively located in the electrostatically positive region, that is most often the site of binding to anionic membranes, but in positively and negatively charged areas. This pattern has also been observed in other membrane-binding PDZ domains. Those residues located in the negative region could play different roles, some residues (e.g., E and D) may form specific hydrogen-bonds with lipid-headgroups, a phenomenon also observe in PH domains (50), while others could have an indirect role in orienting the domain relative to the membrane. It should be noted that the identity and the relative contribution of membrane-binding residues vary significantly among similar PDZ domains. This is demonstrated in residue and cumulative scoring plots for three different PDZ domains, SAP102-PDZ3, rhophilin 2-PDZ, and tamalin-PDZ (see Figure 17). A few high-scoring residues make a predominant contribution for SAP102-PDZ3, whereas many residues contribute relatively evenly to membrane binding for tamalin and rhophilin 2-PDZ domains.

We optimized the classification method for the prediction of membrane-binding PDZ domains. To develop a binary classification method, one needs to define positive and negative cases. Since PDZ domains have a wide range of continuous K_d values, it was necessary to choose a specific K_d value as a threshold for physiologically significant membrane binding. In general, it is not straightforward to predict the cellular membrane binding of a particular pro-

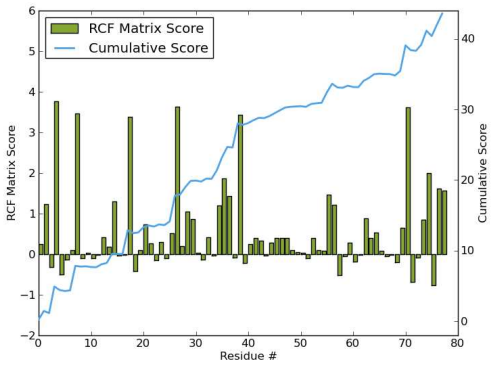
tein from its K_d value for a model membrane. It is because membrane binding of a protein is different from chemical binding of two species with well-defined binding sites (201) and because it is technically challenging to accurately determine the cellular lipid concentrations (212). We have thus taken a combinatorial approach of determining the relative membrane affinity (i.e., in terms of relative K_d) of a family of proteins by the SPR analysis and then measuring their cellular membrane-binding properties to estimate the threshold K_d value for their cellular membrane binding (21; 36). Since syntenin1-PDZ and ZO1-PDZ2, whose membrane affinity is physiologically significant (130; 217), have 1-3 μM K_ds, we set the threshold K_d of PDZ domains to 1 μM. This threshold value divided our SPR-tested PDZ domains into 28 binding cases and 42 non-binding cases. Lowering the cutoff K_d value to 0.5 μM would reduce the positive cases to 15. For evaluation of prediction, we tested two machine learning algorithms that have proven successful in diverse classification applications (13; 16), i.e., the kernel-based support vector machine (SVM) methodology and the decision tree algorithm C4.5 combined with the boosting algorithm AdaBoost (referred to as ABC4.5). Table IV summarizes the results from these algorithms with 10-fold crossvalidations and with different feature sets. The prediction was more accurate when structural and sequence features are used in combination rather than independently. Between the two algorithms, the SVM did well on both the 0.5 and 1 μM K_d cutoff data sets (see also Figure 18). Also, SVM algorithm achieved better accuracy (94%) with balanced sensitivity and selectivity with 1 μM K_d-cutoff. We thus decided to use SVM with all features and $K_d = 1$ μM as a threshold for the genome-wide prediction of membrane-binding activity of PDZ domains.

Algorithm/dataset	Validation	Accuracy	Sensitivity	Specificity	AUC ROC
All features					
SVM-500k$_d$	10-CV	0.925373	0.846154	0.97561	**0.954972**
SVM-1000k$_d$	10-CV	**0.940299**	0.846154	**1**	0.954034
ABC4.5-500k$_d$	10-CV	0.895522	**0.923077**	0.878049	0.920263
ABC4.5-1000k$_d$	10-CV	**0.940299**	**0.923077**	0.95122	0.952627
Structure features					
SVM-1000k$_d$	10-CV	0.893939	0.33333	0.982456	0.793372
ABC4.5-1000k$_d$	10-CV	0.818182	0.22222	0.912281	0.730994
Sequence features					
RCF matrix score	10-CV	0.823529	0.766667	0.832727	0.858182

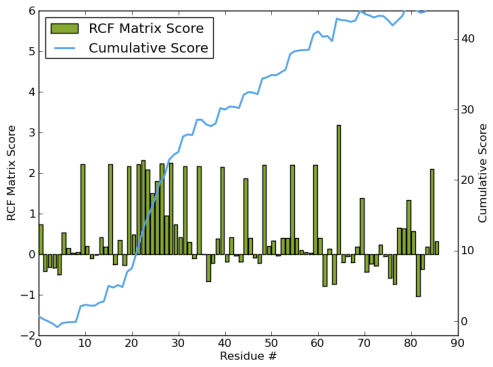
TABLE IV: Comparison of the performance of the SVM and ABC4.5 classifiers on a number of different datasets. The first column given the algorithm used and the K$_d$-cut-off used when separating positive and negative instances in the dataset. We have quantified the the classifier performance on three groups of features to examine their relative importance, for the sequence features there is only a single feature value available, here a simple cut-off strategy to find the best value was used. For each metric the value for the best performing classifier(s) has been highlighted in bold-face font.



(a) Rhophilin2-PDZ



(b) SAP102-PDZ3



(c) Tamalin-PDZ

Figure 17: Residue and cumulative scores obtained from the RFC matrix for (A) SAP102- PDZ3, (B) rhophilin2-PDZ, and (C) tamalin-PDZ. Notice that a few high scoring residues make predominant contribution to the highly positive summed score for SAP102-PDZ3 whereas there is no single motif determining binding for tamalin and rhopholin2 PDZ domains.

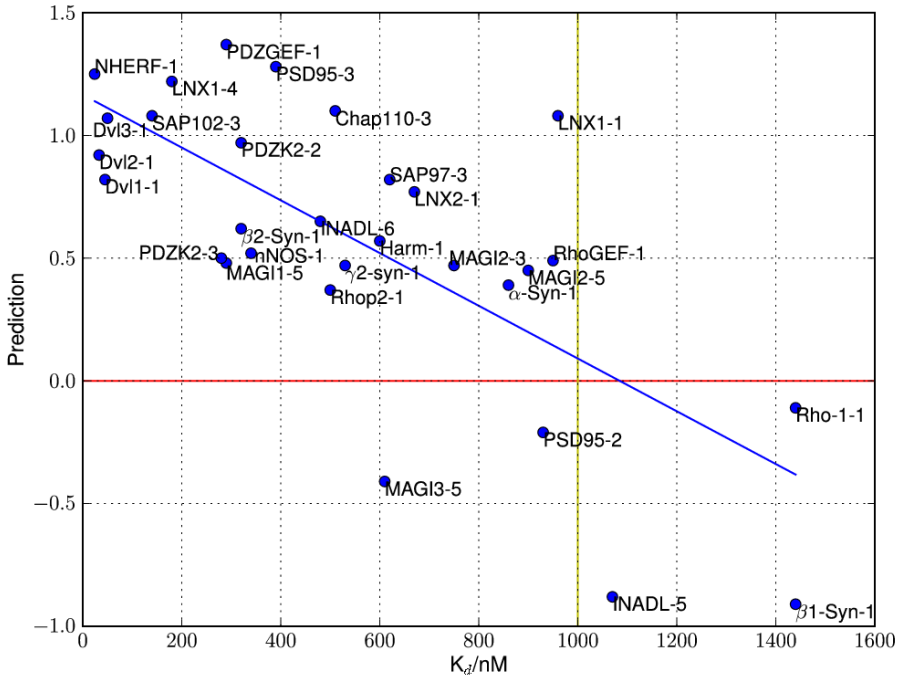


Figure 18: The prediction value for the SVM-1 μM K_d classifier as a function of the K_d value of domains. The yellow line denotes the K_d-cut-off for binding versus non-binding and the red line denotes the classification threshold. The blue regression line demonstrates the good correlation between the prediction and the Kd values. The nomenclature used is "gene name"-"PDZ domain occurrence in the gene" (e.g., SAP97-3 refers to the third PDZ domain in the SAP97 gene).

4.3.2 Predictions for 2,000 PDZ Domains from 20 Different Species

By utilizing the learned protocol, we predicted the membrane-binding properties of all 2,000 PDZ domains found in 20 different species. Since we used both structural and sequence features, domains included in our prediction are from all sequences for which reliable homology models could be generated. As seen in Figure 19, 30% of PDZ domains are predicted to have submicromolar membrane-binding affinity, although some degree of variation is found among species. It thus seems evident again that membrane binding is a common property among PDZ domains. The complete collection of the PDZ domains annotated in this study can be found in our online resource MeTaDoR (Membrane Targeting Domains Resource) (http://metador.bioengr.uic.edu/) (17). Several options for searching the collection are given, among them host protein-name, organism, and binding annotation. There is also an option to classify the domains with variable threshold K_d values. For each domain, the host protein and the domain location in the host protein are given, along with relevant links to public databases.

4.3.3 Experimental Validation of Prediction

To further validate our prediction model, we choose 25 PDZ domains out of the collection of 2,000 predictions and experimentally determine their membrane binding using SPR analysis. Similar to the first screening of PDZ domains, we primarily focus on uncharacterized PDZ domains for validation while adding a few PDZ domains previously characterized (204). Table V correlates experimental measurements, with the prediction values obtained from the 1 μM K_d cut-off. All the binding cases were classified correctly, while three nonbinding cases

Domain	**Start**	**Stop**	#	**Organism**	K_d/**nM**	**Prediction**
C2PA (Q9DC04)	185	271	1	Mouse	-	-0.20
Chapsyn110 (Q63622)	98	182	1	Rat	-	-1.70
Chapsyn110 (Q63622)	193	277	2	Rat	510 ±50	0.17
GRIP (P97879)	252	333	3	Rat	-	-0.30
GRIP (P97879)	471	557	4	Rat	-	-0.30
GRIP (P97879)	572	654	5	Rat	-	0.00
GRIP (P97879)	672	751	6	Rat	-	-0.70
InaD	17	103	1	Drosophila	-	-0.30
MUPP1 (Q8VBX6)	1614	1697	10	Mouse	-	-0.20
PAPIN (Q9QZR8)	85	177	1	Rat	-	-0.30
PAR3 (Q9Z340)	271	359	1	Rat	-	**0.05**
PAR3 (Q9Z340)	590	681	3	Rat	-	-0.06
PTPN3 (P26045)*	510	595	1	Human	450 ±100	0.60
PTPN13 (Q64512)	1084	1167	1	Mouse	-	-0.10
SAP102 (Q62936)*	244	328	2	Rat	-	-0.50
Shank1 (Q9WV48)	663	754	1	Rat	-	-0.40
ZO-1 (P39447)	23	107	1	Mouse	-	-0.70
ZO-1 (P39447)	186	261	2	Mouse	590 ±130	0.40
ZO-1 (P39447)	421	502	3	Mouse	-	-0.70
ZO-2 (Q9Z0U1)	10	94	1	Mouse	-	-0.90
ZO-2 (Q9Z0U1)	287	365	2	Mouse	1200 ±400	**0.10**
ZO-2 (Q9Z0U1)	489	570	3	Mouse	-	-0.40
ZO-3 (Q9QXY1)	11	90	1	Mouse	-	-0.20
ZO-3 (Q9QXY1)	187	261	2	Mouse	-	**0.10**
ZO-3 (Q9QXY1)	370	448	3	Mouse	-	-0.20

TABLE V: Experimental evaluation of our prediction for membrane binding of PDZ domains. The prediction values were calculated using the 1 μM K_d-SVM model. Positive values indicate membrane binding whereas negative values indicate non-binding. The further away from zero, the more confident the prediction is. Prediction values for three mis-classified cases were shown in bold.

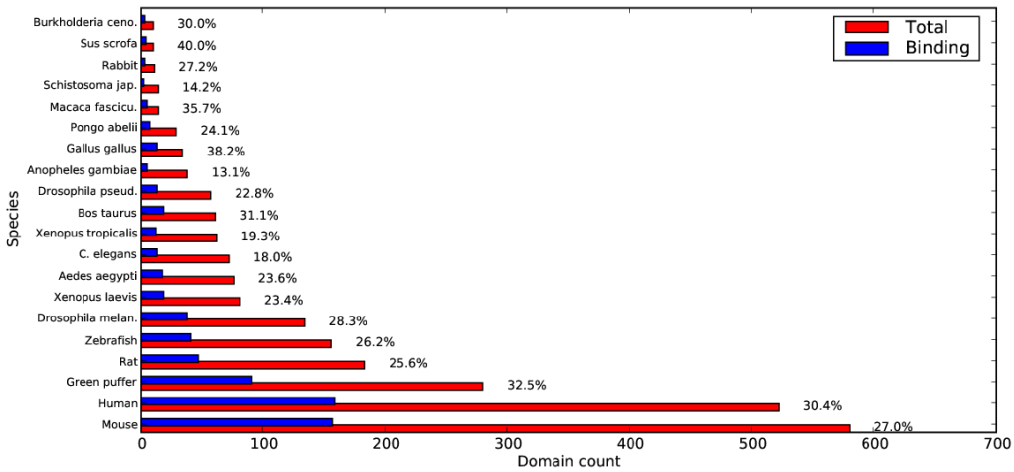


Figure 19: Membrane-Binding statistics for 2,000 PDZ domains found in 20 Species. Predicted percentage of membrane-binding PDZ domains is shown for each species. SVM classifier was used for prediction with all features included and $K_d = 1$ μM as a threshold.

were classified as binding ones. The overall accuracy on the test set was thus 90%, which is comparable to the cross-validation accuracy observed. It is worth noting, that three misclassified cases for nonbinding PDZ domains (i.e., ZO-2-PDZ2, PAR3-PDZ1, and ZO-3-PDZ2) are all border, low-confidence domains having prediction values 0.1, with 0 being the cut-off score distinguishing binding and nonbinding domains. In particular, ZO-2-PDZ2 was predicted show membrane binding, while the experimental K_d value (i.e., 1.2 ± 0.4 μM) is only slightly above the 1 μM Kd threshold. Collectively, this evaluation demonstrates the accuracy and reliability of our prediction. The selection of 70 domains used for the initial database and 25 domains used for evaluation did not bias the outcome of our prediction: i.e., when PDZ domains in the two

groups were interchanged, essentially the same results were obtained in terms of classification and prediction accuracy.

4.3.4 Functional Classification

To systematically analyze the location of membrane-binding sites and the interplay between membrane and protein-binding sites, we determine electrostatic potential at the solvent exposed surface for all mammalian PDZ domains from either experimentally solved structures or homology modeled structures. This analysis indicated that the vast majority of PDZ domains display at least one prominent surface cationic patch which may serve as an interaction site with negatively charged lipids. Depending on the location of the cationic patch relative to the known peptide-binding site, we could categorize the domains into two classes. Class A PDZ domains which have a cationic patch (or two largest patches, in the case that two or more patches are found), having little overlap with the peptide-binding site, and class B PDZ domains having the positive region close to the peptide-binding pocket. We inferred that this structural distinction had functional implications due to the main positive patch in each PDZ domain likely being the lipid-binding site. We tested this assertion by determining the placement of lipid-binding sites and the correlation of lipid and protein binding for domains found to belong to each class.

As a representative of class A PDZ domains, we chose SAP102-PDZ3 which displays a prominent cationic patch (R449, R459, and R484) at the side opposing the peptide-binding pocket. Further, this cationic patch also forms a pocket in the surface, making it a potential site for specific binding of a lipid head-group. Molecular docking was used to determine the most energetically favorable complex containing both the lipid-head group and the interacting peptide.

Figure 20(a) illustrates the lowest energy complex, clearly indicating that peptide and lipid-head group binding occur in two distinct non-overlapping locations. This model agrees well with our SPR analysis that confirmed this PDZ domain has definite selectivity for PtdIns(4,5)P_2 and PtdIns(3,4,5)P_3 over other PtdInsPs, shows PtdIns(4,5)P_2 dependency in membrane binding, and binds soluble inositol 1,4,5-trisphosphate (Ins(1,4,5)P_3). In addition, mutations of cationic residues in the groove (e.g., R449E) significantly reduced affinity for PtdIns(4,5)P_2-containing vesicles. Finally, of the tested mutations none were found to decrease binding to the C-terminal peptide of stargazin, which is a known interaction partner of SAP-102. Thus, this PDZ domain has a clearly defined binding pocket for PtdIns(4,5)P2 and PtdIns(3,4,5)P_3 that is distant from the peptide-binding pocket, it can thus bind a PtdIns(4,5)P_2 (or PtdIns(3,4,5)P_3) and a protein molecule at the same time (experimental results not shown). From the above analysis we propose class A PDZ domains having structurally distinct and functionally orthogonal lipid- and protein-binding sites. This assertion is further supported by functionally independent lipid and peptide-binding sites observed for two additional members of the class A family, PICK1-PDZ (144) and NHERF1-PDZ1 (R.S., Y.C., H.Y. Gee, P.J. Lee, H.R. Melowic, E. Stec, N.R. Blatner, M.P. Tun, M.K., T.K. Fujiwara, H.L., A. Kusumi, M.G. Lee,and W.C., unpublished data).

To directly determine the interplay between lipid and peptide binding of class A PDZ domains, we quantified the binding between SAP102-PDZ3 and the N-fluorescein-labeled stargazin peptide in the presence and absence of PM vesicles (notice that they contain 1% PtdIns(4,5)P_2 for which SAP102-PDZ3 shows selectivity) by fluorescence anisotropy measurements. As shown

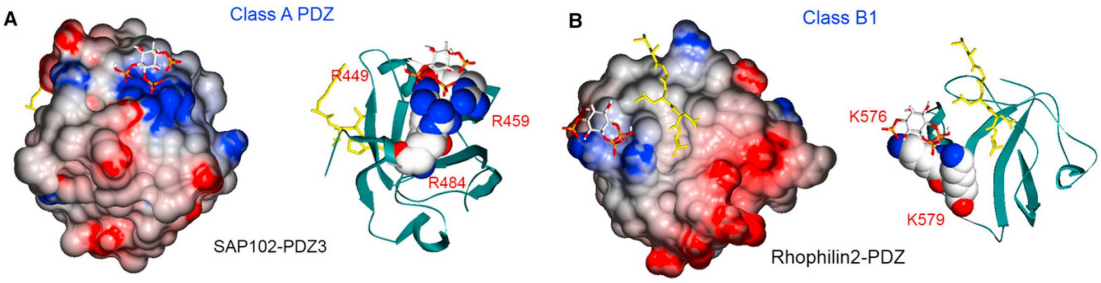
in Figure 21A, the presence of PM vesicles a very small effect on the peptide binding of SAP102-PDZ3. As the majority of PDZ domain were vesicle bound in the present experimental setup, the reuslt is consistent with SAP102-PDZ3 simultaneously being able to bind both membrane and a protein molecules. Further, the presence of PM vesicles had no influence on the affinity of SAP102-PDZ3 to other peptides, from which can be inferred that lipid-binding in class A domains is not directly modulating protein specificity.

Class B PDZ domains are characterized by a positive patch in the area of the α2A helix which constitutes the one side of the peptide-binding pocket. In this group we find domains which have been reported to in part having overlapping (218) or mutually exclusive (130) lipid- and peptide-binding sites. Since the peptide and lipid-binding modes of PDZ domains can vary significantly, however, it is difficult to predict the degree of functional overlap between the two sites based solely on structural examination. We therefore selected two members (rhophilin 2-PDZ and tamalin-PDZ) of this family and investigate the localization of their lipid-binding sites and range of their overlap with respective peptide-binding sites again using molecular docking. Rhophilin 2-PDZ interacts non-specifically with negative lipid surfaces, including PtdInsPs. It has two cationic residues (K576 and K579) placed similarly relative to the α2A close to the C-terminal end which may be involved in anionic lipid binding. Doing a double mutation of K576 and K579 (i.e., K576A/K579A or K576E/K579E) greatly reduced the affinity of Rhophilin 2-PDZ for PM vesicles, indicating that these residues may be specifically involved in binding. Interestingly, these mutants bind the C-terminal peptide of ErbB2 as well as the wild-type (WT), suggesting that the lipid-binding site does not overlap with the peptide-binding site

(experimental results not shown). By sampling a collection of PDZ domain-binding peptides, we identified this peptide as the best binding partner for the rhophilin 2-PDZ. Figure 20(b) show the results from the molecular docking procedure which clearly confirm that rhophilin 2-PDZ can bind an anionic lipid head group and a peptide at the same time. To determine if there is indeed a functional separation of lipid- and peptide-binding sites of rhophilin 2-PDZ, we determine the affinity of rhophilin 2-PDZ for N-fluorescein-labeled peptides both when PM vesicles are present and absent using fluorescence anisotropy analysis (Figure 21B). It was found that the presence of PM vesicles doubled the observed affinity of rhophilin 2-PDZ for the ErbB2 peptide while modestly (i.e., <1.8-fold) decreasing the affinity for other peptides. We thusly conclude that neighboring lipid- and peptide-binding sites of rhophilin 2-PDZ can bind to their respective partners independently, but unlike what was observed for type A PDZ domains, lipid-binding could potentially fine-tune the specificity of protein binding. Corporative binding of neighboring lipid- and peptide-binding sites was also seen in Dvl2-PDZ (R.S.,Y.C., H.Y. Gee, P.J. Lee, H.R. Melowic, E. Stec, N.R. Blatner,M.P. Tun, M.K., T.K. Fujiwara, H.L., A. Kusumi, M.G. Lee, and W.C., unpublished data). We choose to label this mechanism of class B PDZ domains as class B_1.

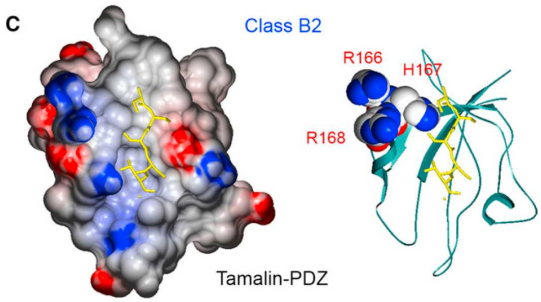
The crystal structure of tamalin-PDZ showed that two phosphate ions are bound to the N-terminal end of the α2A, suggesting that three cationic residues, R166, H167, and R168, are involved in anionic lipid binding (189). We determined that tamalin-PDZ does not display PtdInsP specificity. Mutation of any of the cationic residues to A or E reduced its binding to the PM-mimetic (or PtdInsP-containing) membranes, indicating the this residue positions are

instrumental in nonspecific negatively charged lipid binding. In contrast to rhophilin 2, these mutants did greatly diminish binding activity to the C-terminal peptide of mGluR5, a known interaction partner, which leads us to conlude that there is a significant overlap between the two biding sites. The molecular docking confirmed this interpretation of experimental results by showing a clear overlap between the lipid- and peptide-binding pocket (see Figure 20(c)), in that the most energetically favorable placement of the peptide and lipid-headgroup interfere with each other. (For this reason only the peptide is shown in the final model). The functional overlap between the two binding sites is further verified by the finding that the presence of PM vesicles significantly reduced the binding of tamalin-PDZ to the N-terminal fluorescein-labeled mGluR5 peptide (Figure 21C) and other peptides. To distinguish these PDZ domains from rhophilin-like class B_1 PDZ domains, we designate them class B_2 PDZ domains. Class A PDZ domains are characterized by lipid- and peptide-binding sites that are topologically distinct and have different function. For class B PDZ domains which have neighboring lipid- and peptide-binding sites, functional analysis is, however, needed to pinpoint the role of lipid and peptide binding.



(a) The energy-minimized model structure of the SAP102-PDZ3-peptide(RTTPV)-Ins(1,4,5)P_3 ternary complex. Electrostatic surface potential (left) and the ribbon diagram (right) show the separation of the peptide- and lipid-binding sites. In the ribbon diagram, R449, R459, and R484 that form the cationic groove are shown in space-filling representation and labeled.

(b) The energy-minimized model structure of the Rhophilin 2-PDZ-peptide(EYLGLDVPV)-Ins(1,4,5)P_3 ternary complex. Electrostatic surface potential (left) and the ribbon diagram(right) show the proximity of the peptide- and lipid-binding sites. In the ribbon diagram, K576 and K579 in the α2A helix that are involved in lipid binding are shown in space-filling representation and labeled.



(c) The energy-minimized model structure of tamalin-PDZ-peptide(IRDYTQSSSSL) binary complex. Because of severe steric clash, an Ins(1,4,5)P_3 molecule could not be docked on the PDZ-peptide complex. In the ribbon diagram, R166, H167, and R168 in the α2A helix that constitute the lipidbinding site are shown in space-filling representation and labeled.

Figure 20: Functional Classification of Membrane-Binding PDZ Domains.

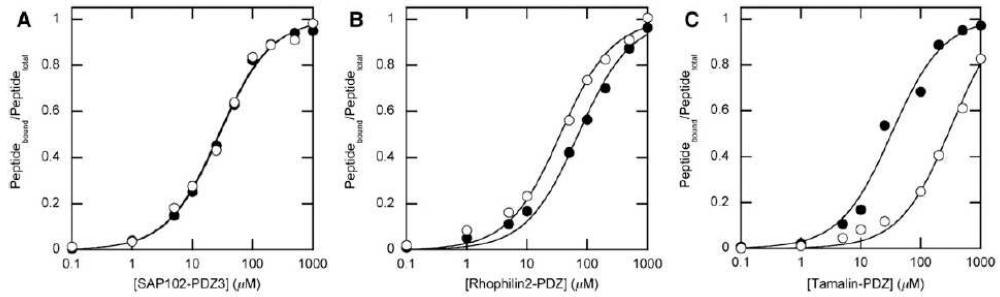


Figure 21: Effects of Lipid Binding of Each Class of the PDZ Domain on Its Peptide Binding.(A) Binding of class A SAP102-PDZ3 to F-Ahx-RTTPV in the absence (filled symbols) and presence (open symbols) of 150 μM PM-mimetic vesicles. (B) Binding of class B1 rhophilin 2-PDZ to F-Ahx-EYLGLDVPV in the absence (filled symbols) and presence (open symbols) of 150 μM PM-mimetic vesicles. (C) Binding of class B_2 tamalin-PDZ to F-Ahx-IRDYTQSSSSL in the absence (filled symbols) and presence (open symbols) of 150 μM PM-mimetic vesicles. The peptide concentration was 5 nM. Notice that for the class A and B_1 PDZ domains, vesicles have a modest to no effect on peptide binding, whereas for the class B2 PDZ domain, vesicles greatly interfere with the peptide binding. Experimental results by Chen and Sheng.

4.4 Discussion and Conclusion

This study describes genome-wide identification, characterization, and classification of membrane binding PDZ domains. Experimental characterization of 95 PDZ domains confirms that membrane binding is a common property of PDZ domains. Experimental measurements show that PDZ domains display a wide continuous range (i.e., 20 nM to >10 μM) of affinity for PM vesicles, making it difficult to arbitrarily distinguish membrane-binding domains from non-binding ones. We developed a flexible and robust binary classification strategy in which a threshold or cut-off K_d value is arbitrarily set and the domains are then divided into those with higher affinity (binding) and those with lower affinity (non-binding). In our online resource users are given the option to set the threshold K_d value, and one can thus predict the membrane-binding activity of PDZ domains in different affinity ranges. Our new classification and prediction protocols represent an advancement in bioinformatics computation, because they allow accurate prediction of membrane-binding proteins from a group of proteins with high sequence and structural similarity. We anticipate that a similar methodology can be utilized in the prediction of any other membrane-binding PIDs that acting as dual-specificity protein- and lipid-binding modules.

A recent study indicated that at least 80% of mouse PDZ domains have protein (or peptide)-binding activity (187). Given that 30% of mouse PDZ domains have sub-micromolar affinity for the PM membrane, the probability that a mouse PDZ domain is a lipid- and protein-binding dual-specificity module is >24%. Also, if a PDZ domain is found to bind membranes, the probability that it can also bind proteins is >90%. These are conservative estimates, and actual

numbers might well be higher for PDZ domains from mice and other species. Thus, it is safe to state that almost all lipid-binding PDZ domains are dual-specificity modules. To gain insight into the evolution of lipid- and protein-binding activities of PDZ domains, we constructed a dendrogram depicting the evolutionary relationship of a collection of PDZ domains. The dendrogram shows that the binding to specific protein classes is somewhat preserved in the tree, whereas membrane-binding properties vary even for evolutionary closely related PDZ domains (e.g., Dvl and SAP97/PSD95/Chapsyn110 clusters).

Also, most PDZ domains have peptide-binding activity, and the location of their peptide-binding pockets is highly conserved (177) in contrast to membrane-binding activity which displays a wide range of affinities and for which the locations of their lipid-binding sites are highly variable. Taken together these findings indicate that dual-specificity lipid- and protein-binding PDZ domains evolved from protein-binding ancestor PDZ domains through convergent evolution. Adding lipid-binding activity to the functionality of PDZ domains essentially allows for an extra layer of regulation on the critical cellular functions of their host proteins.

Through electrostatic potential calculation of membrane-binding PDZ domains we were able to determine two types of cationic surface patches, separating these domains into two groups. Mutational and functional analysis confirmed that class A PDZ domains have structurally distinct lipid and protein-binding sites. Thus, these PDZ domains can serve as dual-specificity lipid- and protein-binding modules that mediate both membrane binding and protein interaction simultaneously. Also, many class A PDZ domains have a fairly well-defined surface groove in the cationic region, indicating that these sites may play a role in lipid head group selectivity

(Table III) in contrast to most class B PDZ domains which do not display significant lipid selectivity. While our results show that lipid binding of class A PDZ domains does not change peptide affinity directly, under physiological conditions, however, lipid binding should enhance affinity and specificity for their protein partners due to reduction in dimensionality (35; 128). Thus, their dual specificity should be pivotal for the formation and regulation of membrane-associated protein networking.

Essentially all class B PDZ domains show significant cationic patches in or close to the α2A helix forming one wall of the peptide-binding site. It was observed that class B_1 PDZ domains have cationic patches confined near the C-terminal end of the 2A helix, while class B_2 PDZ domains mainly have cationic patches in the N-terminal end of the helix. Since a protein typically enters the pocket from the N-terminal end of the α2A helix to place its C-terminus near the carboxylate-binding loop, it may be that lipid binding at the N-terminal end of the α2A helix will prevent peptide binding by spatially restricting the entry way of the peptide into the pocket. Class B_1 PDZ domains are similar to class A PDZ domains in the sense that both act as dual-specificity modules enabling protein networking through membrane-association by coincident binding. In contrast to class A PDZ domains, lipid binding may promote peptide binding specificity of class B1 PDZ domains, possibly through a structural rearrangement of the peptide-binding pocket. Since lipid and peptide binding are mutually exclusive and compete with each other for class B_2 PDZ domains, lipids may act as a control mechanisms for the accessibility of the protein-binding pocket to potential binding peptides.

The key role of lipid binding is to spatially restrict its host-protein at the membrane. Consequently, lipid binding of PDZ domains mainly serves to control trans-location to membranes and the activity of their host proteins. For most reported class A and class B_2 PDZ domains, their lipid-binding activity appears to be important for the cellular localization of their host proteins. Experimental results from Rhophilin 2 containing a class B_1 PDZ domain indicate that this is a common feature of all dual-specificity PDZ domains. It is, however, often not straightforward to determine the correlation between membrane affinity and cellular membrane trans-location, because membrane-binding is also affected by interactions with membrane proteins or membrane-associated proteins. Take for example the yeast scaffold protein Ste5 for which lipid binding was found to modulate the behavior of the protein at the PM to a much greater extend than the trans-location to membrane in the first place (203). Similarly, the interaction with specific lipids of a group of PDZ domains may induced greater changes in dynamic properties of the parent proteins at the membrane rather than driving the actual membrane localization. In either case the prediction of PM affinity presented here will still serve as a reliable indicator of the likelihood of a PDZ domain to interact with any cell membrane, due to the fact that electrostatic interaction between the PDZ structure and the lipid-surface serves as the main factor in binding to any cytosolic membrane. It should also be pointed out that certain PDZ domains may have the ability to hetero- or homodimerize under physiological conditions (60; 177), which may modulate their effective membrane affinity through an additive crowd effect.

In sum, the presented data clearly supports the notion that many PIDs interacting at membrane surfaces are dual-specificity lipid- and protein-interaction modules. Further, lipid binding of different classes of PDZ domains regulates the cellular function and their host proteins by different mechanisms. It is thus becoming evident that the interpretation of complex data studying protein-protein interaction and networking must take into account the membrane-binding properties of PIDs.

The computational modeling of experimental measurements of PDZ domains will hopefully be a valuable resource for other groups studying the protein networking properties of the PDZ domain family. In addition, the methods developed in this chapter will hopefully be instrumental in functional characterization of novel PDZ domains as well PIDs from other families.

CHAPTER 5

LEARNING THE RULES OF MEMBRANE-BINDING

5.1 Introduction

Signal transduction networks formed by specific protein-protein and protein-lipid interactions are a primary means by which the cell transmits information from its external environment to intracellular recipients. One vehicle driving the intracellular signal transduction speed beyond that of simple diffusion is the selective and reversible binding of so-called *peripheral proteins* to membrane surfaces within the cell (82; 147). By redistributing cytosolic proteins to membranes in response to the onset of signaling events a *de facto* compartmentalization of the cellular space takes place, allowing for greater proximity among communicating parties, thereby facilitating interaction (81). The importance of this mode of signal transduction is underlined by the fact that more than 10% of human protein kinases contain at least one lipid-binding module (82). The ability to identify and understand peripheral proteins and the physical factors causing their co-localization at membranes is thus pivotal in uncovering the dynamics governing signaling regimens.

Peripheral proteins are most commonly scaffold proteins containing one or more domains that associate with lipid-head groups, thereby anchoring the entire protein structure near the lipid surface (36; 49; 147). An increasing number of ubiquitous and structurally distinct domains have been found to display lipid binding properties, collectively referred to

as membrane-targeting domains (MTDs). MTDs have been identified in the following families: C1 (33; 190; 210), C2 (33; 164; 134), PH (61; 109), FYVE (Fab1/YOTB/Vac1/EEA1) (186), PX (phox) (209), ENTH (Epsin N-terminal homology)(28), and recently PDZ domains (32). Despite their highly similar intra-family folds, not all domains in these families possess membrane-targeting properties. In fact, a diverse array of overlapping intra-family functions exist, spanning from protein-protein interaction to structural support and potentially enzymatic activity (82).

Numerous experimental techniques have been used to identify novel MTDs (33; 49) revealing details on binding mechanisms and orientation (42; 9). Genome-scale identification and characterization of MTDs does, however, remain labor-intensive and expensive. To this end *in silico* protocols offers a high-throughput complement to wet-lab methods, allowing for rapid characterization of thousands of domains. Membrane binding properties are inherently difficult to predict, as they are often not determined by well-defined sequence motifs or a specific structural composition. PDZ domains were, for instance, found to have highly diverging membrane binding behavior in spite of high sequence similarity (32), and PH domains span a large range of binding affinities despite being structurally very similar (109; 180).

In previous works from our lab machine learning, protocols for distinguishing MTDs from a general body of cytosolic protein domains known to have no membrane binding activity were constructed using Support Vector Machines (SVM) (16) and later on extended using other classifiers (105). By representing each domain as a numerical vector of feature-values derived from structural data, a classification model achieving 90% accuracy in separating binding and

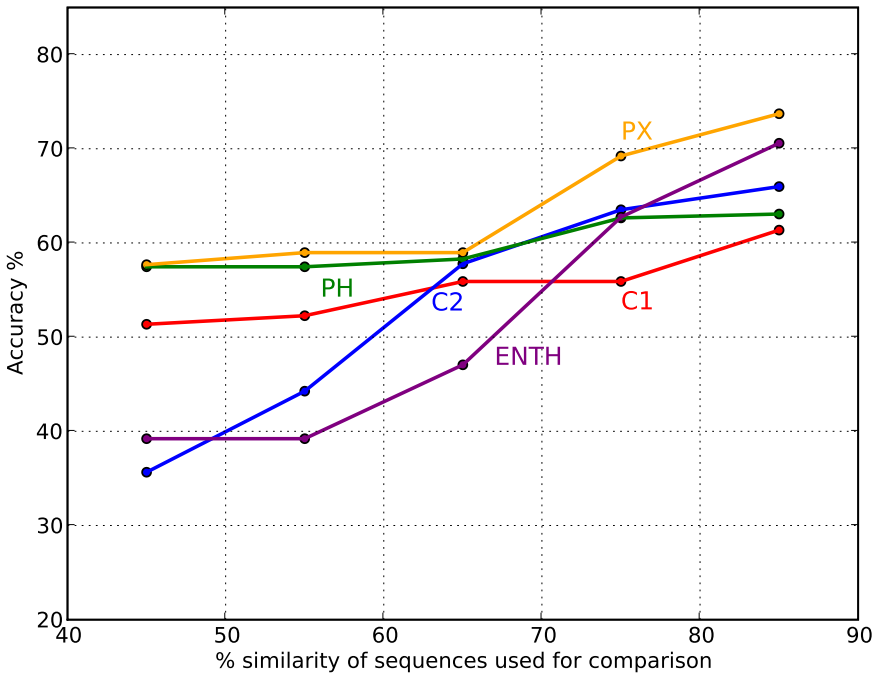


Figure 22: Performance of a sequence based nearest neighbor classification procedure. Accuracy measures for classification of membrane binding properties for five domain families using consensus of the three nearest neighbors for each domain. The accuracy for each family is depicted at varying levels of maximum sequence similarity allowed between instances in dataset for each domain family.

non-binding domains was constructed. There are, however, two issues to be addressed regarding this model. First, while performing well when separating MTDs from cytosolic protein domains of unrelated fold, the model does not provide similar performance in separating binding and non-binding domains within any specific family. As we will demonstrate, intra-family classification is in fact a very hard problem as even highly similar domains display different membrane binding

properties. Second, the constructed SVM model does, to a great extent, function as a "black-box" classifier giving little insight as to how the different calculated features play together in producing the final classification of a domain's binding properties.

In this work we construct a series of classification models for separating membrane binding domains from domains with other activity within families. Our focus is on C1, C2, and PH domain families, as domains from these three families have been found to be key players in a number of signaling pathways. We are, however, not merely interested in constructing models for classification, as such models are of limited utility in explaining the predicted behavior in a manner that leads to experimentally testable hypotheses. Rather, we want to provide both a confident assessment of a given domain's binding behavior and a body of biological evidence supporting the classification label. The goal is, in other words, to go from data mining to knowledge mining, revealing the specific mechanisms responsible for observable higher level behavior. To this end we take advantage of a new ensemble based classifier, namely the Alternating Decision tree (ADtree) algorithm (65). The ADtree relates to other classification tree algorithms like CART and C4.5 (159) by quantifying the relationship between features as a combination of rules each representing a binary decision on a feature. The ADtree is based on the boosting technique, but is at the same time a tree structure representable as a conjunction of rules all contributing a real-valued additive evidence towards classification. The final classification decision is thus determined by a committee voting scheme based on the real values evidence presented by each rule traversed in the tree by a given domain. This scheme

makes representation of the classifier as a spare and easily interpretable tree structure possible, a feat that has made it the preferred tree classifier in a number of studies (70; 116).

The paper is organized as follows: First we give the intuition behind the features used to represent the individual domains in a form suitable for constructing machine learning protocols. We then construct classification models based on SVM and ADtree to separate intra-family binding and non-binding domains. Finally, we analyze the individual rules utilized in determining membrane targeting behavior in the three domain families in terms of experimentally known binding mechanisms.

5.2 Methods

5.2.1 Dataset

Special care was taken when selecting the positive and negative examples in the datasets utilized, as both the instance groups come from the same domain family (17). After reducing the sequence identity to 70% using CD-HIT (79) of the full set of known domains, a total of 303 sequences were left. Each of these instances was then examined manually, and classified as positive (binding) and negative (non-binding) based on their functions, sub-cellular location and similarity with other sequences. The final statistic for the three datasets used for training are given in Table VI. As a reference the total number of annotated domains for each family in PFAM is also provided, to underline the ubiquitous nature of all three families. The 70% cut-off was chosen since sequence similarity at this level does not result in conservation of membrane-binding properties, as illustrated in the Results section.

Only a subset of the domains in the constructed datasets has an experimentally solved structure available. For the remaining cases we construct homology models using RaptorX for modeling (153).

TABLE VI: Dataset statistics for the three domain families.

Domain	Binding	Non-binding	MaxSimilarity	PFAM
C1	33	22	70 %	1536
C2	63	27	70 %	4666
PH	70	88	70 %	4125

5.2.2 Classifiers and evaluation

Models were constructed using two binary classification procedures, namely Alternating Decision tree (65) and Support Vector Machines (SVM) (44). Both are supervised classifier, for which a model is trained on a labeled training dataset (training mode) and thereafter applied to predict new examples without further parameter tuning (prediction mode). Casting the problem in a binary classification framework, we refer to each protein domain as an instance, with the i^{th} instance consisting of a feature vector $x_i \in [1 \times n]$ and a label $y_i \in \{0, 1\}$, with n denoting the feature count. Both algorithms described construct a function, $g(x)$, that minimizes the empirical risk of misclassifying an instance, under the assumption that all instances are drawn with respect to the same (unknown) probability distribution. In the following we limit ourselves

to describing conceptual details of the utilized algorithms, referring the reader to cited works for technical details.

The SVM methodology facilitates the derivation of a classification hyperplane (a hyperplane separating positive and negative cases) for non-linear problems by working in a vector-space of higher dimension than that of the original feature space (using the so-called "kernel-trick"). The separating hyperplane, $wx+b$, can be found by numerically solving the following quadratic optimization problem:

$$\min_{w,\xi_i,b} \frac{1}{2} w \cdot w + C \sum_i \xi_i \qquad \text{Subject to}$$

$$y_i(\phi(x_i) \cdot w + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Where C and ξ_i are cost parameters associated with misclassification and $\phi(x_i)$ is a non-linear mapping function. Rewriting the above problem in the dual form, the kernel-trick, specifying a function giving the inner-product of two vectors in a higher dimensional vector-space, can then be applied and an optimal separating hyperplane can be found in this new vector-space. In this work we tested Gaussian, Sigmoid, and polynomial kernel function and found the Gaussian to give the best results, thus whenever SVM results are reported it refers to SVM using a Gaussian kernel function.

The ADtree utilizes the boosting methodology (65) in the same manner as other successful classification schemes such Adaboost C4.5 (159), but has the advantage of producing models that are easily representable as a tree with a limited number of nodes (often fewer than 20), without sacrificing predictive power. This is achieved by constructing a tree that is a conjunction of rules which all provide an additive evidence toward a given instance being classified as positive or negative, depending on the evaluation of the rules (True or False). In addition to providing the classification label, the tree score of an instance (the margin score) can be interpreted as a measure of confidence in the classification label. Unlike traditional tree models obtained from algorithms such as C4.5, the classification of instances by ADtree is thus not determined by a single path traversed in the tree, but rather by a collection of paths. The tree is made up of two types of nodes *prediction nodes*, represented by ellipses, and *splitter nodes*, represented by rectangles. Each splitter node is associated with a real valued number indicating the rule condition: If the feature represented by the node is less than or equal to the condition value for a given instance, the prediction path will go through the left child node, otherwise the path will go through the right child node. The final classification score produced by the tree is found by summing the values from all the prediction nodes reached by the instance, with the root node being the precondition of the classifier. If the summed score is greater than zero, the instance is classified as positive, otherwise, as negative.

We use our in-house machine learning workbench MALIBU for the construction and validation of models, giving a uniform interface for comparison and analysis of their performance (105).

We measure the performance of the constructed classification models using the following metrics: Accuracy (Acc) defined as the ratio of true prediction to the total number of prediction, Sensitivity (Sen) defined as the probability that a true example is classified as true, Specificity (Spe) defined as the probability that a negative example is classified as negative. The classification result of an instance in a binary classification can be fall into four categories: True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN). Using these counts the three metrics are approximated by: $Acc = (TP + TN)/(TP + TN + FP + FN)$, $Sen = TP/(TP + FN)$, and $Spe = TN/(TN + FP)$

Further, we use the area under the curve of the receiving operator characteristic curve (AUC ROC). AUC ROC is defined as the area under the (1-specificity, sensitivity)-curve, with each point corresponding to a specific threshold for class separation; a value of 1 performs perfect over the entire range of threshold values, with a random classifier having an AUC-value of 0.5.

To provide a benchmark for the expected performance we use n-fold Cross Validation (n-CV). In n-CV one randomly divides the original dataset into n equally sized bins, each classifier is then trained n times using $n - 1$ subsets. The omitted subset in each round is used for estimating the evaluation metrics of interest, the average of which is thus based on evaluation over all instances. For this work 20-CV was used.

5.2.3 Features

The association with membranes is known to be driven by a combination of general lipid binding mechanisms and the binding of key-residues with specific lipid head-groups. The general association mechanisms are modeled by quantifying the chemical and physical properties of the

domain structure as a collection of "patches" on the solvent exposed surface (SES). A patch is a well-defined area of a given property on the surface, i.e. an area of all positive electrostatic potential or a region of conserved hydrophobicity; here we use the area of five largest patches in each category. Further the surface propensities of the 20 amino acids are included as features for a total of 35 structural features. The steps of patch growing detailed below are outlined in (100). The basic idea is as follows: First, the surface is defined as a collection of neighboring triangles, second, a numerical representation of the quantity of interest is associated with each triangle, and finally the patches that are most highly correlated with the function of the structure are defined.

Assume that each triangle on the surface is associated with a numerical value corresponding to the quantity that forms the basis for patch growing. We will denote this value for a triangle t by $t.val$ and the distance between the centroids of triangles t_1 and t_2 by $dist(t_1, t_2)$. Furthermore, $t.neigh$ will denote the neighbor triangles of t, meaning those that share an edge with t, and $t.included$ will be a boolean flag indicating whether a given triangle has been included in a patch. The collection of patches is then found by repeating the following recursive procedure until all surface triangles have been included in a patch: Choose a random triangle that has not yet been included in a patch and extend the patch by adding neighbor triangles that satisfy $|t1.val - t2.val| < C$, where C is a constant.

In order to do the patch-growing we need to assign values from the quantity of interest to each triangle on the SES. In this work we use three quantities: (1) The electrostatic potential obtained from solving the Poisson-Boltzmann (PB) equation using APBS (6). The spatial

potential values are mapped onto the surface by taking a weighted average of the 8 discrete data points closest to the point 1 A from the triangle surface in the direction of its normal vector. (2) Hydrophobicity values are assigned to the surface based on the Kyte-Doolittle value of the amino acid that gave rise to the triangle of interest (99), (3) Hydrogen-bonding is mapped to the surface by determining if an atom is capable of forming a hydrogen bond, indicated by setting $t.val = 1$.

In addition to the patch-growing procedure, features solely based on statistics from the full collection of domain sequences are used. A score for each domain sequence is obtained by its similarity to other sequences in the dataset, this is done using a recursive functional classification (RFC) matrix inspired by Park *et. al* (146). A multiple sequence alignment of all domain sequences (both binding and non-binding) is created using a Hidden Markov Model-profile (for the C1, C2 and PH domain the PFAM models PF00130, PF00168, and PF00169 were used, respectively). Based on the alignment we calculate the probability of observing amino acid a at location i in the alignment. Denoting the probability for binding and non-binding cases by $P_{a,i,+}$ and $P_{a,i,-}$, respectively, each entry in the the RCF matrix is given by:

$$RCF_{a,i} = \log\left(\frac{P_{a,i,+}}{P_{a,i,-}}\right)$$

Thus a positive/negative entry in the matrix indicates that the presence of amino acid a at location i is evidence towards the domain being membrane binding/non-binding. We can summarize the evidence for a giving domain sequence S as being binding in the following score:

$$\text{RCF-score(S)} = \sum_{s_i \in S} RCF_{s_i,i}$$

For the sequence features we do, however, choose to decompose the RFC matrix into a series of residue subsequence features of lengths 3 to 6, to be able more specifically pin-point the exact local variation that was evidence for classification. Rather than using all possible rules, we include the 25 rules that provide the greatest degree of discriminatory power in the training set. The rule locations are selected through an initial bootstrap generation of several RFC matrices and a ranking of the rules that have the highest potential RCF score. The subsequence rules are thus intended to complement general membrane-binding mechanisms by identifying subsets of residues that correlate with specific binding modes. In general the quantifying local environment conservation has been shown to be of great utility in identifying remote similarity properties. Recently procedure focusing on local environment have been utilized with great success in the identification of DNA-binding protein domains (106) the and in more general purpose protocols for remote homology detection (19).

5.3 Results

The contribution of this work is two-fold. First, we show that machine learning models based on the sequence and structure features introduced below perform significantly better

than procedures based on sequence homology in separating MTDs from non-MTDs within families. Second, we demonstrate how ADtree models not only perform comparable to SVM based models, but also present us with specific evidence for the classification label allowing us to interpret the model within the context of current experimental observations.

TABLE VII: Performance comparison of models for C1, C2, and PH domain families.

Family	Algorithm	Acc.	Sen.	Spe.	AUC ROC
C1	ADTree	0.891	**0.939**	0.818	0.887
	SVM	**0.907**	0.909	**0.909**	**0.957**
	Seq Sim	0.57	0.43	0.78	-
C2	ADTree	**0.856**	**0.889**	**0.778**	**0.879**
	SVM	0.792	0.838	0.7	0.878
	Seq Sim	0.63	0.61	0.7	-
PH	ADTree	0.861	0.824	0.853	0.905
	SVM	**0.867**	**0.843**	**0.886**	**0.939**
	Seq Sim	0.64	0.64	0.62	-

To further illustrate the difficulty of the current classification task of intra-family separation, a simple unsupervised classification scheme aimed at predicting the membrane binding behavior of a domain based on the binding behavior of closely related homologs is fashioned. A sequence is predicted to be binding or non-binding from the majority vote decision of its three closest related sequence neighbors (as defined from a BLAST search(3)). Figure 22 depicts the prediction accuracy for the procedure as a function of varying levels of the maximum sequence similarity

allowed between domains in the dataset. It is evident that even at maximum sequence similarity levels as high as 85%, accuracies of no more than 75% can be achieved for any family, indicating the need for more sophisticated procedures to confidently identify MTDs within families.

5.3.1 Overall classifier performance

Table Table VII compares the performance of SVM, ADtree, and the sequence based nearest neighbor protocol for three domain families. In all families the two structure-based machine learning protocols perform significantly better domain separation than the sequence based nearest neighbor procedure at the 70% sequence identity level. For C1 domains both SVM and ADtree improve on the accuracy of the sequence based method by more than 30%-point, with both achieving accuracy rates in the 90%-range. The SVM protocol does, however, have a 6% better AUC ROC than the ADtree due to a better balance between sensitivity and specificity, making it the strongest classifier for this family.

Inspecting the models constructed for C2 and PH domains we again observe a far better performance of the machine learning methods over the sequence homology based classifier, with accuracy improvements of 22 percentage points for both families. Comparing the SVM and ADtree models for the C2 domain families, a similar performance over the entire specificity range is observed as the AUC ROC of the two models is almost identical, although the ADtree achieves a higher accuracy when comparing the points of best class separation for the two classifiers. For the PH domain family we again observe the two classifiers performing comparably, with a small advantage to the SVM algorithm of 3 percentage points as measured by AUC ROC. In sum, the results show that for the problem at hand the ADtree models perform comparably or

only slightly worse than SVM, indicating almost no loss in performance resulting from the use of a model that provides the benefit of human interpretability.

Though not presented here, we have experimented with machine learning protocols relying solely on sequence derived features. While these protocols did show better performance than the nearest neighbor homology based procedure used as the bench-mark above, they never achieved accuracies higher than 75%, which is significantly less than what was achieved by using both sequence and structure based features.

5.3.2 Knowledge-mining

Here we use the ADTree model to discover the rules that distinguish binding and non-binding domains. A graphical representation of the ADtree model constructed for the C1 domain family is depicted in Figure 23, yellow and blue splitter nodes signify structure and sequence based feature rules, respectively. Sequence based rules are divided into positive and negative patterns indicating subsequences in the domain family alignment for which there is a high/low RCF-score for binding/non-binding sequences. Feature names are followed by a number in parenthesis indicating the order in which the rules are added to the model, a measure that can be interpreted as an importance ranking of the rules. The fact that top-ranked rules are a mix of sequence and structure based features, indicate that both groups are adding orthogonal predictive power to the model, a feat also observed in the tree models for C2 and PH domains (not shown). In the following subsections we will interpret key rules in the three classification models in terms of their biological meaning in driving reversible membrane binding; the importance ranking is used when referring to specific rules in each model.

5.3.2.1 C1 model

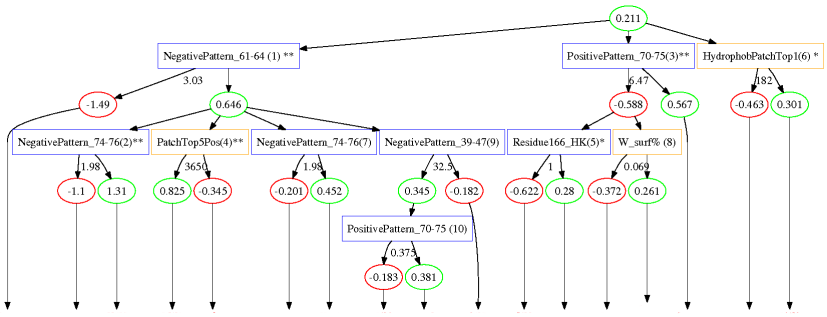


Figure 23: The ADtree model constructed for the C1 domain family. A single rule is represented by two elliptical *prediction nodes* and a rectangular *splitter nodes*. Each splitter node is associated with a real valued number indicating the rule condition, if the condition is true/false the path traversed by an instance will go through the left/right child node and accumulate the score in this node towards the overall classification of a domain. Splitter nodes colored in blue stem from sequence feature while yellow ones stem from structure features.

C1 domains are cystine-rich modules of approximately 50 amino acids in length, first discovered in Protein Kinase C (PKC) and subsequently found in signaling families such as PKDs, chimaerins, RasGRPs, and diacyglycerol kinases (DGK) (41).

The sequence of the known binding case PKCδ-C1a is utilized in Figure 24 for illustrating key rules learned for the entire C1-family. Membrane binding of C1 domains is known to be driven by specific binding of diacylglycerol (DAG) and phorbol esters in the membrane as well the association of key residue with the membrane surface and coordinated binding of Zn^{2+}, these

feature are highlighted in the PKC sequence. Rule 2 and 3 both overlap with the second group of membrane and DAG binding residues, indicating that two different kinds of conservation appear here one associated with membrane binding and one associated with other activity. NegRule_1 is observed to be high scoring if residues 11 and 13 are not aromatic, correlating well with the experimental evidence for binding, as interfacial penetration of the lipid bilayer has been found to be driven by aromatic residues (105; 184).

In addition to the conservation of specific sequence groups, two global structure mechanism are also discovered by the C1 tree. Rule 4 and 6 indicate that if the cumulative size of the five largest electrostatic patches and the size of the largest hydrophobic patch are greater than specific threshold values it is indicative of membrane binding. This observation correlates well with the fact that certain C1 domains are known to deeply penetrate the hydrophobic membrane core upon binding, an interaction that is only energetically favorable if non-polar surface-residue exists. Likewise, a somewhat positively charged surface is necessary for the initial recruitment of the domain to anionic lipid surfaces (82). The two mechanisms are illustrated in the lower part of Figure 24 by structure data from a binding and a non-binding C1 case, with electrostatic positive and negative isosurfaces superimposed and hydrophobic residues highlighted in yellow. For the binding case there is a large well-defined bulk of positive electrostatics and a cluster of hydrophobic residues, while the non-binding only displays sporadic regions of positive charge.

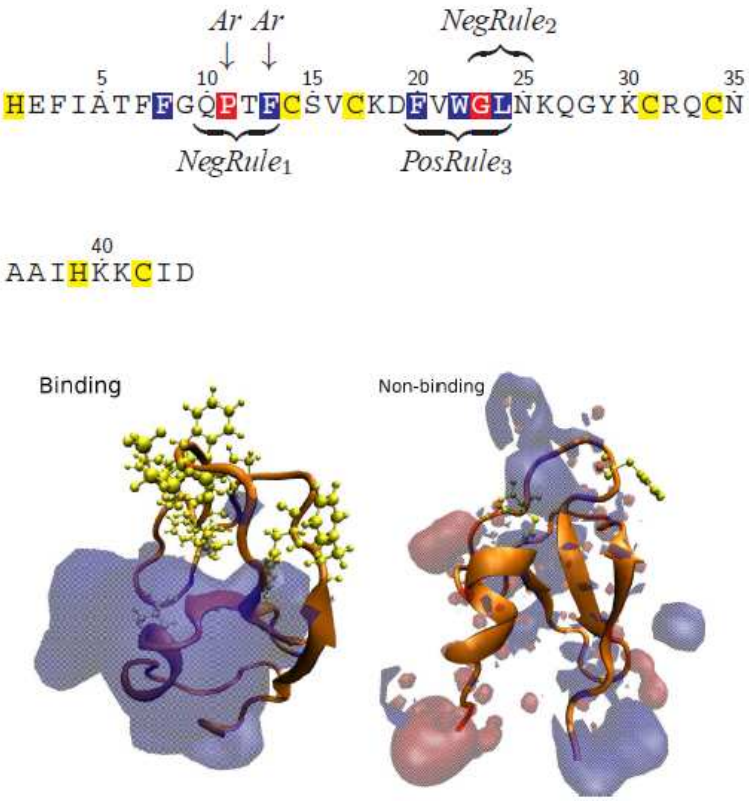


Figure 24: Rules learned for the C1-family. The sequence for PKCδ-C1a is used for illustrating key rules, the residue coloring used is membrane-binding (blue), DAG binding (red), and Zinc-binding (yellow). Structure models for a binding case and non-binding case are shown, with the positive and negative electrostatic isosurfaces color in blue and red, respectively. In addition, hydrophobic residues are highlighted in yellow.

5.3.2.2 C2 model

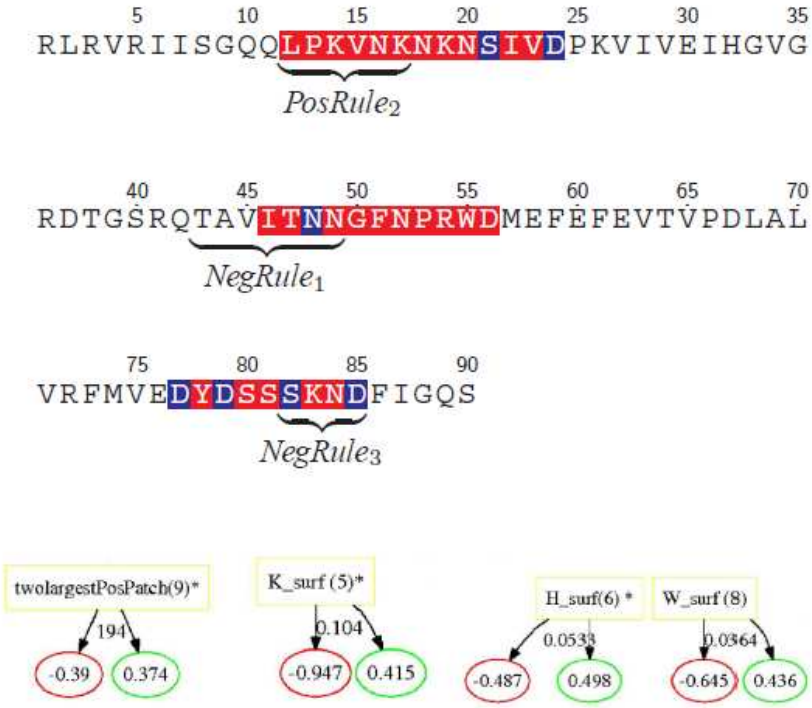


Figure 25: Rules learned for the C2-family. The sequence for PLCδ1 is used for illustrating key sequence rules, the residue coloring used is Ca^{2+} binding side-chain (blue) and Ca^{2+}-binding region (red). Further four structure rules from the C2 model are shown.

Most C2 domains require activation by divalent Ca^{2+} ions to bind to membranes with high affinity and show low affinity towards lipids otherwise (36; 82). In the majority of cases binding of Ca^{2+}-ions dramatically enhances the positive electrostatic potentials around the

Ca^{2+}-binding region that mediates the association with the anionic lipids (termed as an electrostatic switch) [10, 48] or induces a conformational change that accelerates binding (97; 175). As illustrated in Figure 25, the discovered sequence rules from the C2 domain model all overlap with Ca^{2+}-binding regions. Interestingly, the first negative rules also overlap with regions suggested to be involved in protein-protein interactions in PKCϵ (22) indicating the conservation of functional properties other than membrane binding in this region.

Further, a number of structure rules are utilized in the C2 model. Cationic residues on the surface (in corporation with Ca^{2+}-bridging) are important for anionic-lipid selectivity (i.e. Synaptotagmin) (36). We observe this in rules indicating a threshold on positive surface patches and surface propensity on K. Finally, selectivity for the lipid-head group PC in C2 domains is achieved through aromatic and aliphatic surface residues (i.e. observed in cPLA2), represented in the model as high surface propensities of amino-acids H and W being indicative of membrane binding (33).

5.3.2.3 PH model

PH domains are recruited to membranes by Phosphatidylinositol lipids such as Phosphatidylinositol (3,4,5)-trisphosphate (PIP_3) and phosphatidylinositol (4,5)-bisphosphate (PIP_2), and are, in example, found in $\beta\gamma$-subunits of heterotrimeric G proteins(199) and PKC (211). For the PH domain family, binding often occurs in two steps, an initial association is driven by non-specific electrostatic interactions followed by specific binding to anionic lipids (81). Key rules from the PH family model, illustrated in Figure 26, agree well with this overall process of membrane association. Two structure rules, both presenting minimum cut-offs on the size

of electrostatic patches, are present. As observed in other families, a large positive patch is indicative of binding. Interestingly, a smaller negative patch is also positively associated with binding in the case of PH domains. This apparent inconsistency (that both positive and negative charge promotes membrane binding) can be explained from the local arrangement of the electrostatic potential depicted in Figure 26. Here we see that the negative region is on the opposite side of the membrane associating surfaces, thus the repulsion of this side to a negatively charged membrane can help correctly position the domain relative to the membrane, a mechanism previously hypothesized (82).

In addition to the electrostatic mechanism for correct spatial orientation, sequence rules 1 and 2 overlap with residue experimentally determined to be important in Phosphatidylinositol and general membrane binding (82). The third sequence rule mapped does not immediately correlate with any known residues important for binding, thus making it a novel prediction, but does contain two positively charged residues that may be important in binding.

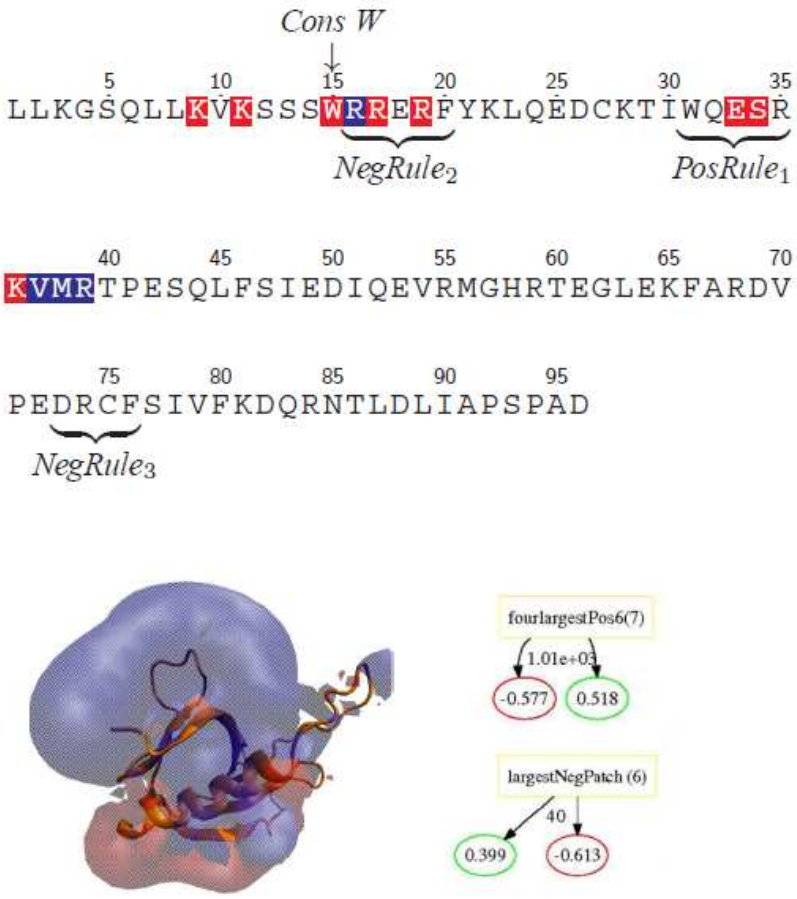


Figure 26: Rules learned for the PH-family. The sequence for PLCδ1 is used for illustrating key rules, the residue coloring used is Membrane binding (blue) and PIP_3/PIP_2 binding (red). A structure model for a binding case is shown with the positive and negative electrostatic isosurfaces color in blue and red, respectively. In addition, two key structure rules are depicted.

5.4 Discussion and Conclusion

The present work touches on two key challenges of computational biology: How do we efficiently organize and classify the vastly expanding body of data produced by experimentalists; and of even greater importance, how do we transform this data into biological knowledge in the form of testable hypotheses? It can be argued that simple rule mining would be an appropriate option to deduce classification rules. However, it is widely believed that discriminative approaches are far superior to generative ones given their simplicity. Moreover, discriminative classifiers have been shown to have a lower asymptotic error (20). Further, ADtrees have the ability to elicit more uncorrelated rules (by definition) that cover diverse features of the data.

The graphical models built for the three families highlight the general rules and features that set binding instances of peripheral proteins apart from non-binding ones. While some of these features are in agreement with previous studies, novel features are also proposed. In general we find that structural features, such as a specific cut-off for the size of positive electrostatic surface patches, are found in models for all domains and thusly constitute general mechanisms driving membrane binding. Sequence based features on the other hand, are more important in expressing unique binding properties for each family as they are rooted in local regions of the domains.

Characteristics elucidated from the rules learned in this work can be used to guide further experimental studies. For example, mutation of certain amino acids that are statistically over-represented in important rules could be suggested as pointers for experiments (such as Aromatic residues for C1/C2 domains). Similarly, if feasible, features like overall charge on the domain

could be tinkered with by multiple mutations of charged residues. Such guided studies are expected to reduce the effort and time required to reveal the mechanisms and features used by peripheral proteins and highlights the value of knowledge-mining over the “black-box“ type approaches that are often used in classification of biological data.

CHAPTER 6

MODELING THE UNFOLDING OF MECHANICAL PROTEINS WITH DISCRETE STATES

In previous chapters we covered methods for probabilistic identification of the protein entities making up proteomes from mass-spectrometry data as well as protocols for identification of functional classes in large protein domain datasets. In this chapter we turn our focus to the role of single domain dynamics in signaling. Specifically, we will use the atomic-scale time evolution models available from MD simulations to gauge how domain structures rearrange themselves in response to the onset of a signaling event. Understanding the elements of the protein structure responsible for structural stability as well as how these elements respond to changes in the physical or chemical environment of the structure is a key stepping-stone in the design of regimens for manipulation of protein behavior and by extension the higher level dynamics of protein interaction networks.

6.1 Introduction

The set of stable structural conformations a protein will assume in its native state is guided by the many local energy minima found in the complex energy landscape formed by the sum of additive contributions made by interstructure atomic interactions. Figure 27 illustrates how changing the chemical or physical environment of the protein structure alters the energy landscape and consequently biases the protein domain towards a new set of stable conformations,

possibly rendering it enzymatically active or enabling it to interact with other protein domains. Broadly defined, signaling mechanisms on the molecular level can thus be thought of as a controlled transfer of energy from one communicating party to the next. This transfer may take place through the build-up of a diffusion gradient as observed in the mithochondrial synthesis of ATP, the conversion of chemical to electrical energy as seen in the generation of neural action potential, or simply by cascades of phosphorylation events as observed in kinase networks (181).

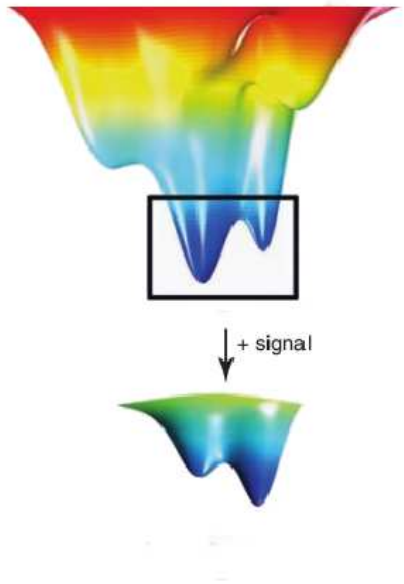


Figure 27: Conceptual depiction of a protein structure energy-landscape. Upon onset of a signaling event a dynamic shift in the relative likelihood of structural conformations occurs, resulting in a larger segment of the protein population being in the signal activated state.

Any of the signaling mechanisms mentioned will introduce a dynamic shift in the energy landscape of the involved protein entity and could thus potentially be studied by the methods presented in this chapter. We will, however, for the purpose of method development, focus on a completely different means of signaling, namely that of mechanical force. Mechanical force has been found to play a crucial role in many physiological processes by regulating the reversible folding and binding of single protein domains (101; 85). As a result, protein domains involved in these processes need to respond properly to mechanical strain in order to perform their function. Examples can be found in such diverse areas as stem cell differentiation (126), phosphorylation rate determination (87),and the differentiation of myotubes (54) (see recent reviews for a multitude of other examples (197)).

Mechanical proteins constitute a highly suitable model system for studying the dynamics 'rearrangements induced by the onset of signaling events (which in this specific case means partial structural unfolding). First, there is a clearly defined parameter indicating when a signaling event has taken place, namely the distance between the N- and C-terminal Cα atoms. This metric extending beyond what is observed in the native state of the protein indicates that partial unfolding of the protein domain has occurred. The completion of the signaling event is thus readily observable by measuring the end-to-end extension of the polypeptide chain. Second, there already exists a thoroughly tested computational protocol for imitating the effects of mechanical strain on single protein molecules, namely that of Steered Molecular Dynamics (SMD). Third, experimental single molecule measurements that can be used for verification of

the behavior predicted by of our model already exist in the form of Atomic Force Microscopy (AMF) (62).

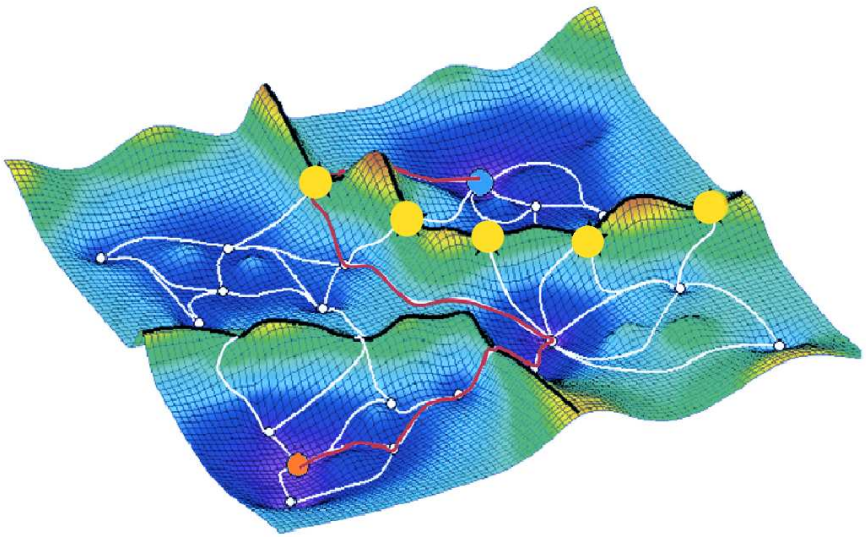


Figure 28: A transition network superimposed on the underlying energy landscape. The blue node indicates the starting state and the orange the end state. Each white node corresponds to a local energy minima. Stars indicate the potential crossing-over points for transitioning across the rate-limiting energy-barrier separating start- and end-state. The red path indicates one possible transition from start to end through the energy-landscape.

MD has already been utilized in numerous application for studying the dynamics of protein behavior. Specifically in the case of mechanical proteins the use of SMD has proven valuable in casting light on key events in mechanical unfolding of proteins. Many details of experimental observations do, however, remain elusive when simply observing the data generated by SMD in

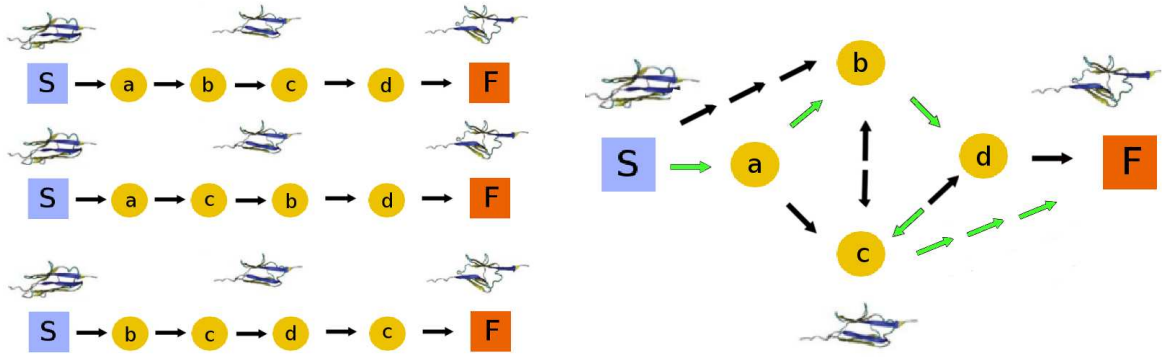
the form of an animation of a single trajectory, or when plotting key quantities from a single trajectory (such as the conservation of key hydrogen bonds or structural packing density). One reason why it is difficult to capture the finer details of mechanical unfolding (or any type of protein dynamics for that matter), is the stochastic nature of these processes. Figure 28 illustrates how a single trajectory will only sample one of many possible paths through a complex energy landscape. Thus, even if all macroscopic inputs of the system under consideration are the same (pressure, temperature, ion concentration etc.), one will not observe the same dynamics behavior every time a simulation is carried out due small random perturbations in the starting structure.

A single simulation thus only represents one of many possible routes between the starting and end point in the structural rearrangement occurring during a signaling event. Consequently, one will need to carry out numerous simulations to obtain a reasonable approximation of the major routes through the energy landscape responsible for observed higher level behavior in order to get a full picture of the underlying generative process. The development of computational procedures for modeling energy landscape dynamics have thus far mainly been focused on protein folding dynamics of small peptides. Most notably is the work by Noe *et al.* uncovering the folding states of Ala_{12} peptide from extended molecule to folded helix using multiple molecular dynamics trajectories. Likewise Chodera *et al.* developed a protocol for determining the metastable state in the folding of the engineered 12-residues $\beta - haipin$ trpzip2 protein and developed a kinetic model of the folding dynamics. These works do, however, not focus on the dynamics of a full-size protein once it is in the folded state, and the dynamics of near-native

conformations. Since functional properties of proteins occur in this part of the energy landscape it is necessary to develop methods for modeling the structural changes occurring in a folded protein.

We present a method for constructing integrative models in the form of a Markov Chain Model (MCM) representing the major dynamic events taking place during the partial unfolding of the muscle protein domain I27. We do so by grouping the observations from multiple SMD trajectories into one network model representing all major transition paths as illustrated in Figure 29. We use the developed framework to explore the effect of varying levels of mechanical force on the unfolding dynamics of I27 and investigate the changes in unfolding pathways observed in a number of mutant structures experimentally found to be of varying mechanical stability. We find that numerous unfolding pathways are present in I27, all contributing to the overall rate of unfolding at different levels depending on the mechanical pulling force applied.

The chapter is organized as follows: First, the theoretical framework needed for unifying numerous time-evolution samples of the same mechanical unfolding process in one MCM is developed. Second, we demonstrate the ability of SMD to correctly reproduce the ranking of mechanical strength of a number of I27 mutant structures. Finally, we use the developed framework to construct an unfolding network for I27 demonstrating the changes in the unfolding energy landscape induced by different pulling forces and different mutant structures.



(a) A collection of three single trajectory paths through the state-space. Each trajectory progress through a collection of the four states A, B, C, and D, representing the transition through several local minima before complete unfolding.

(b) Inferred state network from the collection of single trajectories. Arrows indicate the possible transitions between states, the path in green corresponds to a transition from start to finish that is not observed in any trajectory, but is revealed as possible by combining information from multiple trajectories.

Figure 29: The construction of a Markov Chain for modeling the dynamics of a protein energy landscape of mechanical protein unfolding from a collection of MD trajectories.

6.2 Methods

Figure 30 provides an overview of the procedure for constructing the MCM from a collection of SMD trajectories, in the following sections we will give details on the computations done in each step.

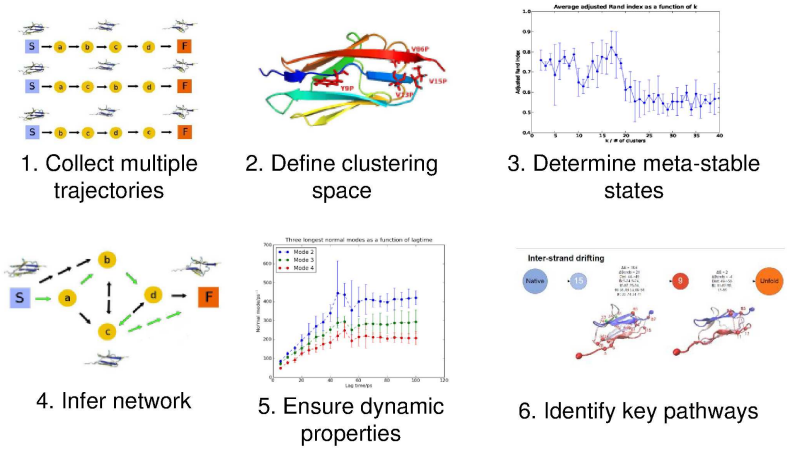


Figure 30: Overview of the procedure for constructing the MCM from a collection of SMD trajectories.

6.2.1 Markov Chains

The memoryless Master equation is often used to model the transition processes between conformational sub-states in a multi-state system. Assume $P(t)$ to be a m-dimensional vector representing the probability that a system is in one of m substates at time t, and K to be a rate

matrix, with K_{ij} designating the transition rate constant from state i to j, then the transition process is described by:

$$\frac{dP(t)}{dt} = \mathbf{K}P(t) \rightarrow P(t) = e^{\mathbf{K}t}P(0) \tag{6.1}$$

Alternatively, the transition dynamics can be modeled as a discrete-time Markov process with transition matrix $\mathbf{T}(\tau)$, where $\mathbf{T}_{ij}$ denotes the probability of the system to be in state i at time t and state j at time $t + \tau$. The analog to equation Equation 6.1 then becomes:

$$P(k\tau) = \mathbf{T}(k\tau)P(0) = \mathbf{T}^k(\tau)P(0) \tag{6.2}$$

The eigenstructure of the matrix $\mathbf{T}(\tau)$ is particularly valuable in describing the dynamic properties of the process it models (138; 139). We require $\mathbf{T}(\tau)$ to be ergodic, meaning that any state in the chain can be reached from any other state in a finite number of steps, resulting in $\mathbf{T}(\tau)$, having a unique eigenvector with eigenvalue 1. If we normalize this eigenvector we get the stable equilibrium distribution of the system denoted π. The sign-structure of all subsequent eigenvector, q_i, describes a "transition-mode" between states of the chain, while the corresponding eigenvalue λ_i denotes the percentage of molecules that have undergone the transition q_i after time τ (due to the way $\mathbf{T}(\tau)$ is constructed it will be positive-definite and thus have strictly positive eigenvalues).

For the above assertions to hold, it is crucial that the system modeled is "memoryless" or *Markovian*. This implies that the future state of the system only depends on its current

state and not past history. The most common cause of non-Markovian effects is the presence of internal energy barriers within the discrete substates. Two parameters can be adjusted to avoid such artifacts, the number of states in the model and the lag-time τ used when constructing the transition matrix. For a reasonable definition of the state-space, any model will be Markovian if sufficiently long lag-times are allowed. For the model to provide useful information on the dynamics studied one does, however, want to identify the shortest possible timescale at which the system becomes Markovian (37). To determine the minimal lag-time for a given topology models a memoryless process, we use the fact that a model which is Markovian if the lag-time τ was used for constructing the transitions matrix, will also be Markovian using lag-time τ', with $\tau' > \tau$. Thus any kinetic properties will converge in τ. One such property is the so-called *implied timescales* of the system, or the *relaxation time* of the different modes of the process, given by the transitions matrix eigenstructure as described above. Combining Equation 6.2 and Equation 6.1 we have $\mathbf{T}(\tau) = e^{\mathbf{K}\tau}$. From this relation we can determine the implied timescale of the i^{th} transition mode, τ_i^*, by:

$$\tau_i^* = \frac{\tau}{\ln(\lambda_i)} \tag{6.3}$$

In practice, using the above described framework to construct a Markov chain from a collection of SMD trajectories and ensuring Markovian properties of the derived model, requires the completion of three steps:

1. Determine a vector representation of conformations observed in trajectories and cluster these into a number of metastable sub-states.
2. Determine a statistically stable transition matrix and find the lag-time at which the process becomes Markovian.
3. If some states are not sufficiently sampled to provide a statistically stable transition matrix, re-sample these states by restarting simulations from a conformation within the state.

6.2.2 Defining the State Space

For the purpose of clustering the frameset from the trajectories into substates, a vectorial representation of each state is needed. Since we are mainly concerned with the order in which specific interactions are broken and formed during the course of the unfolding process, we use a modified contact matrix representation of the structure. Each entry in this vector corresponds to the euclidean distance between two residues (as measured from the center of mass of the residue). To reduce the size of the vector (and thus the clustering space) we only include contacts that are less than $8A$ in at least one trajectory frame. In example, only including short contacts that are likely to influence the free energy level of the entire structure will reduce the clustering space of a structure such as I27 from 3916 dimensions to around 500-600 depending on specific parameter choices.

We use a k-means algorithm for clustering the trajectory frames. The main challenge here is to determine the correct number of substates k from the data. Our strategy is to determine a value for k that provides the most consistent clustering of similar structural conformations.

More specifically, consider the two frame f_1 and f_2 which stem from the same state. Under the correct k we would expect these two frames to appear in the same cluster consistently even if small perturbations to the dataset are made. If this type of consistency is not observed it is an indication that a substates has been split into two clusters (in other words, too large a k has been chosen). To determine the cluster count that is most robust to perturbations in the dataset we deploy the procedure outlined in Figure 31.

```
Data = full set of trajectories
for k from 2 to 200:
   Create 20 bootstrap samples of Data called D = {d1...d20}
   C = []  //list of clustering on data in D
   for Sample in D:
      - do k-means clustering on Sample
      - store clustering of Sample in C
   calculate adjusted Rand index for C
```

Figure 31: Pseudocode for SMD trajectory clustering procedure.

Briefly described the procedure generates 20 clusterings on bootstrap samples of the frame data for each value of k in the range 2-200. For each k we calculate the Adjusted Rand index (ARI), a measure ranging from 0 to 1, indicating cluster consistency corrected for chance events (with few clusters it is more likely that two data points will cluster together by chance than

with many clusters) (196). We choose the k that provides the highest ARI and use this k to cluster the entire dataset.

6.2.3 Evaluation of clustering algorithms

To calculate the consistency between two clusterings of bootstrap samples from the same dataset we use ARI. We first define the simple Rand Index, R, a measure for comparing the consistency of two distinct partitions of a dataset into an arbitrary number of subsets. Consider a dataset, S, containing N elements, and two partitions of S to be compared $X = \{x_1 \dots x_r\}$ and $Y = \{y_1 \dots y_s\}$. We define a as the elements in S that are in the same sets in X and the same sets in Y, b as the elements in S that are in different sets in X and in different sets in Y, c as the elements in S that are in different sets in X and the same sets in Y, and c as the elements in S that are in the same sets in X and different sets in Y. The Rand Index is then defined as

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{N}{2}} \tag{6.4}$$

The Rand index does, however, suffer from one problem. For random data it will be higher for low cluster counts than for higher simply because two data-points are more likely to be clustered together by chance. The Adjusted Rand is a version of the Rand index corrected for chance. The following contingency table denotes the common objects of two clusterings, with n_{ij} denoting the number of common object in clusters x_i and y_j:

	y_1	y_2	$\dots$	y_n	Sum
x_1	n_{11}	n_{12}	$\dots$	n_{1c}	a_1
x_2	n_{21}	n_{22}	$\dots$	n_{2c}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_n	n_{r1}	n_{r2}	$\dots$	n_{rc}	a_r
Sum	b_1	b_2	$\dots$	b_n	

The ARI can be calculated as

$$ARI = \frac{Index - ExpectedIndex}{MaxIndex - ExprectedIndex} \tag{6.5}$$

or formulated in terms of the contingency table

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}} \tag{6.6}$$

6.2.4 Ensuring Markovian Properties of the Model

Once the substate-space is determined we can calculate the transition matrix entries for the MCM at a given lag-time τ. For each pair of micro-state (i, j) the entry is computed as follows, where $trans_{ij}(\tau)$ denotes the the number of transitions from state i to state j in lag-time τ and $start_i(\tau)$ the number of times remaining in state i:

$$T_{ij}(\tau) = \frac{trans_{ij}(\tau)}{start_i(\tau)} \tag{6.7}$$

We need to determine the lag-time τ at which the transition process becomes Markovian to construct a MCM that truthfully represents the dynamics of the process being modeled. This is achieved by calculating the transition matrix at different lag-times $(\tau, 2\tau, 3\tau, \ldots)$ and observing the development of implied timescales as defined in Equation 6.3 as a function of lag-time. At each lag-time we use a 10-fold bootstrap procedure to calculate a set of transition matrices and from these a distribution of implied time-scales are calculated. We define the criteria for convergence of the time-scale as the first lag-time at which the five following lag-times are not statistically significantly different at level 5%. Further, to ensure that all transition probabilities are sufficiently sampled we carry out a multiple hypothesis test to determine which probabilities can be said statistically significantly different from zero.

6.3 Results

In this section we will first demonstrate the ability of SMD to correctly reproduce experimental observations regarding the relative strength of a number of I27 mutants pulled at constant force. Thereafter we construct a MCM representing the diverse unfolding pathways of wild-type I27 pulled at different forces and show how diverse sets of state transition sequences dominate the unfolding process at low and high pulling forces. Finally, we utilize an MCM of four I27 mutants to explain the change in mechanical unfolding pathways induced by the changes in the amino-acid sequence of the domain.

6.3.1 Diverse mechanical properties of I27 mutants predicted by SMD

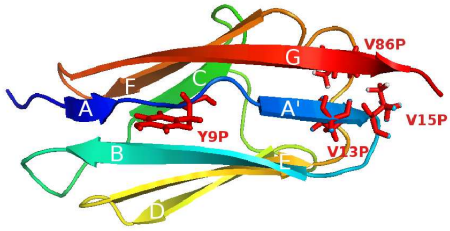
It is well established that the mechanical stability of a protein domain cannot simply be determined from its structural topology. While protein structures with anti-parallel β-strand of

the N- and C-terminal residues, such as I27, protein-G, and Ubiquitin, all display a markedly higher mechanical stability than almost all other known folds, there is still a high degree of variance in unfolding force within these topologically similar domains that remains unexplained. There is, in other words, no well established model for determining how the unique sequence features and local structural elements play together in determining the exact mechanical properties of otherwise highly similar protein domains.

This notion was established in an AFM study of four I27 mutants by Li *et al* (112). In this work it was demonstrated how four different mutant structures of I27 displayed significantly different unfolding properties. Figure 32 summarizes the maximum unfolding force observed from constant velocity pulling at 0.6 nm/ms^{-1} (result for mutant V86A extrapolated from (202)). The mutants V13P, V15P, and V86A display lower mechanical stability than wild-type I27, while the mutant Y9P has a markedly higher peak force at 269 pN.

Mutant	Unfolding force/pN	Samples
Wild-type	204	266
V13P	132	384
V15P	159	259
Y9P	268	340
V86A	148	210

(a) Peak unfolding forces observed for WT I27 and four mutants from constant velocity pulling experiments at pulling speed 0.6 nm/ms^{-1}.



(b) Spatial location of I27 mutations labeled on the WT structure of I27 (PDB id 1TIT)).

Figure 32: Overview of kinetic properties for five I27 mutant species.

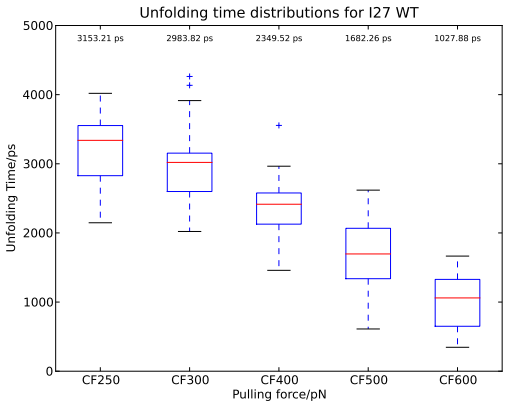
These experiments clearly demonstrate that single-point mutations in the A' and G β-strands can serve as a means by which the mechanical properties of the structure can be tuned to selectively unfold at specific forces. It is, however, not clear why some mutants make the structure stronger while others weaken its ability to withstand mechanical strain. To explore the mechanisms of unfolding we conduct a series of constant force SMD simulations at five discrete forces. The number of simulations carried out for each protein mutant at each force are summarized in Table VIII, with the cumulative simulation time indicated in parenthesis. Each simulation is started at a random time point in a 30 ns equilibration simulation and stopped once the N- C-terminal distance of the structure extends beyond 64A (corresponding to the point at which all hydrogen-bonds between the G- and A/A'-strand are broken).

Protein Species	Force/pN				
	250	300	400	500	600
WT	38 (120ns)	66 (197ns)	50 (117ns)	230 (386ns)	38 (70ns)
Y9P	1 (4ns)	18 (68ns)	28 (88ns)	34 (73ns)	20 (27ns)
V13P	20 (70ns)	30 (75ns)	32 (48ns)	32 (21ns)	32 (11ns)
V15P	16 (50ns)	28 (74ns)	34 (55ns)	34 (29ns)	32 (15ns)
V86A	22 (61ns)	22 (61ns)	32 (53ns)	34 (32ns)	32 (19ns)

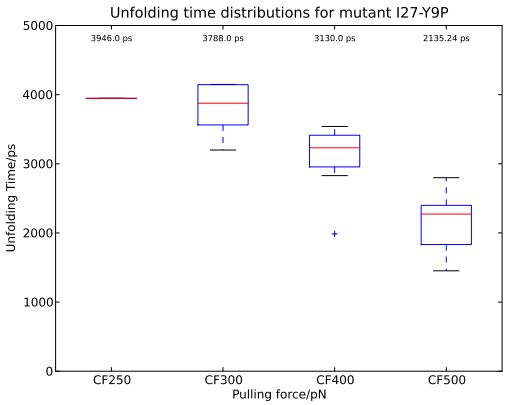
TABLE VIII: Simulation statistics for WT I27 and four mutants. For each protein species the number of simulations carried out at a given force along with the cumulative time simulated in nanoseconds is indicated.

Each simulation is thus characterized by a specific break-time, t, at which the structure unfolds. This time is an indication of the specific mutant species' ability to withstand the pulling force applied, with higher average values of t indicating a stronger protein structure. The box-plots in Figure 33 show the distribution of break-times for the five protein species. As expected there is a clear correlation between break-time and the applied force, with break-times growing shorter as the force is increased. In general wider distributions are observed at lower forces, which may indicate that a larger range of different unfolding pathways is possible when the force is lowered, whereas higher forces appear to only utilize a few pathways.

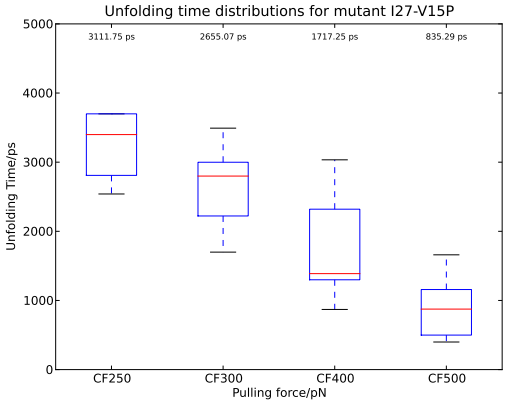
A full comparison of the unfolding time distributions for all mutants as a function of the pulling force applied in the constant force protocol is depicted in Figure 34. When comparing the unfolding time ranking with experimental results it is evident that our SMD simulations reproduce the relative mechanical stability ranking of the five protein species. Any direct comparison between SMD results and the experimental measurements is not possible as they were obtained using different protocols (constant velocity and constant force pulling protocols, respectively). The fact that the Y9P mutant displays significantly longer unfolding times at all forces than wild-type I27, and the mutants V13P, V15P and V86A, display significantly shorter unfolding times at all forces is indicative of SMD simulations correctly representing the dynamics of the system. Thus it is likely that any mechanism we may find explaining differential mechanical strength of the four mutants will be a reflection of the true dynamics observed in wet-lab experiments.



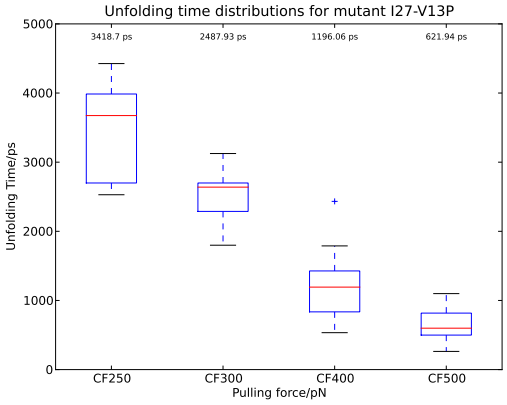
(a) Wildtype.



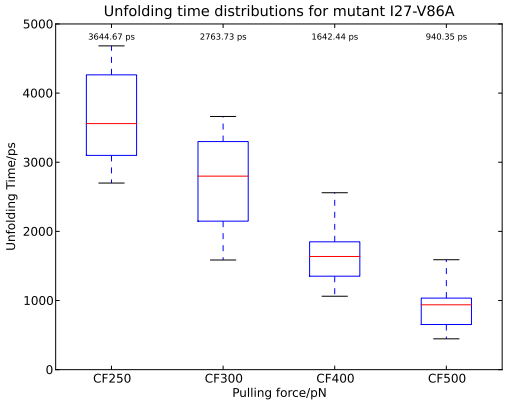
(b) Y9P.



(c) V15P.



(d) V13P.



(e) V86A.

Figure 33: Overview of break-time distributions from constant force-pulling of I27 mutants at five forces.

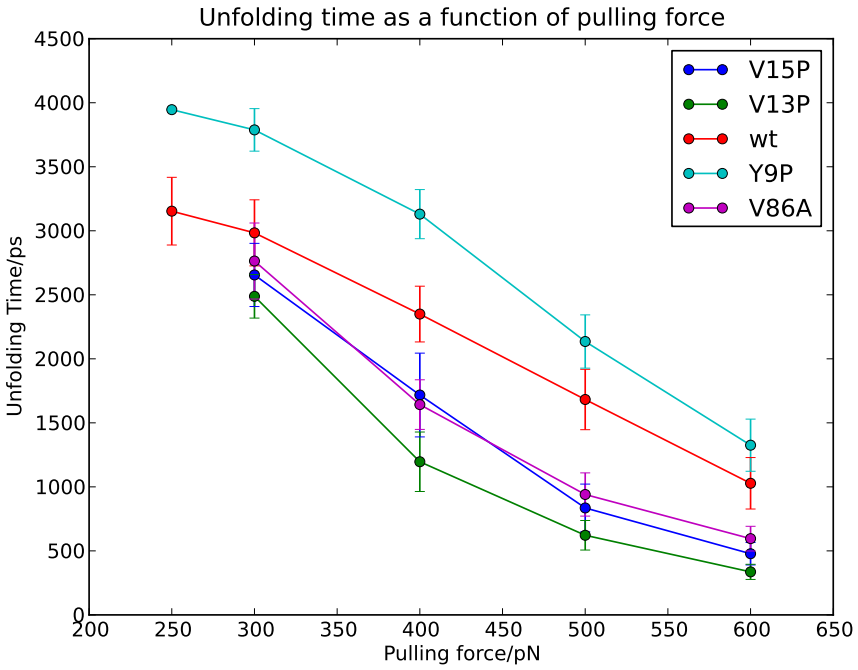


Figure 34: The average unfolding time for I27-WT and 4 mutants. The average unfolding time is indicated as a function of pulling force with the error bars indicating the standard deviation of each data-point. Note that only one data-point was obtained for Y9P at 250 pN.

6.3.2 A model for the unfolding of WT I27

As demonstrated, the statistics derived from a large number of SMD simulations of I27 clearly show a correlation between unfolding time and pulling force. We will now construct an MCM for the unfolding of wild-type I27 from simulations data obtained at forces in the range 250-600pN to uncover the unfolding pathway space dominant at different forces.

6.3.2.1 Determining the clustering space

The first step in constructing a network model from a collection of trajectories is to define the set metastable states that exist in the unfolding process. To determine the state space of the process each trajectory is split into an ordered sequence of snapshots of static structures by storing the structural conformation assumed at discrete time points. It is assumed that each metastable state is characterized by a collection of highly structurally similar conformations. Thus to determine the correct number of metastable states we have to determine the number of clusters that best partition the complete collection of structural snapshots from all trajectories.

First, a suitable/feasible vector-space for comparing the set of structure snapshots and a metric for measuring the similarity between two structures need established. The simplest choice would be to use the root mean square distance (RMSD) between two structure snapshots, thereby measuring the overall average distance change between all residues. Using this metric would, however, introduce certain complications that may lead to an incorrect picture of the energy landscape. For instance, the change in position of a residue centrally located in the structure will likely impact a high number of energy terms and thus cause a significant shift in the energy landscape, while the change in position of a peripherally placed residue may have

little influence; yet both changes would be treated similarly. To overcome this potential artifact we instead use the "contact-space" of the all trajectories under consideration. The contact-space is determined as the complete set of residue-pair for which the distance between the center of mass is within a pre-specified cut-off distance for any of the structures in the clustering analysis. In other words, the contact-space is the full set of all residue-residue contacts observed in any frame in any trajectory. Since the global energy landscape of a protein structure is mainly defined by the set of non-bonded contacts, the suggested definition of contact space is believed to better represent the changes in the energy landscape occurring, when the protein is plied with mechanical force. Another key advantage of only using the set of contacts that are actually observed to be within a given cut-off range (rather than the full contact matrix), is a significant reduction in the clustering space.

Table IX summarizes the size of the contact-space for a number of simulation subsets at contact cut-off definitions of 6, 7, and 8 A. As one would expect, for all subsets of trajectories the contact count grows with the cut-off values. Comparing the number of contacts as a function of the pulling force, we interestingly observed that a larger number of contacts occur at lower forces indicating that at high forces the unfolding pathways do in general explore a smaller region of the contact-space prior to unfolding. Another interesting observation is the number of contacts found when determining the combined space of wild-type I27 and one or more mutants. In these instances we also find an increase in the size of the contact space indicating that the mutant explores different structural arrangements on the path to unfolding than wild-type I27.

		Cut-off/A		
Protein Species	Forces/pN	6	7	8
WT	250,300,400,500,600	642	858	1083
WT	300,400,500,600	522	725	930
WT,Y9P	300,400,500,600	617	834	1060
WT,V13P	300,400,500,600	565	771	979
WT,V15P	300,400,500,600	584	798	1022
WT,V86A	300,400,500,600	581	786	997
WT,All Mutants	300,400,500,600	674	904	1135
WT	250	556	764	980
WT	300	464	667	878
WT	400	344	558	679
WT	500	357	537	714
WT	600	321	512	667

TABLE IX: The interaction count for a number of protein species groups for different cut-off values. For each species the set of pulling forces used are indicated. For comparison, the full contact matrix of I27 will have 3916 non-redundant entries.

For the purpose of clustering we apply the k-means bootstrapping procedure outlined in the Methods section to all wild-type trajectories using the contact-space defined by a cut-off value of 7A. Figure 35 depicts the distribution of the ARI as a function of the the number of clusters. It is evident that the highest average level of consistency among bootstrapped clusterings is achieved for $k = 18$. For higher numbers of clusters a significantly lower consistency level is observed; we will thus use the clustering with 18 states for further analysis.

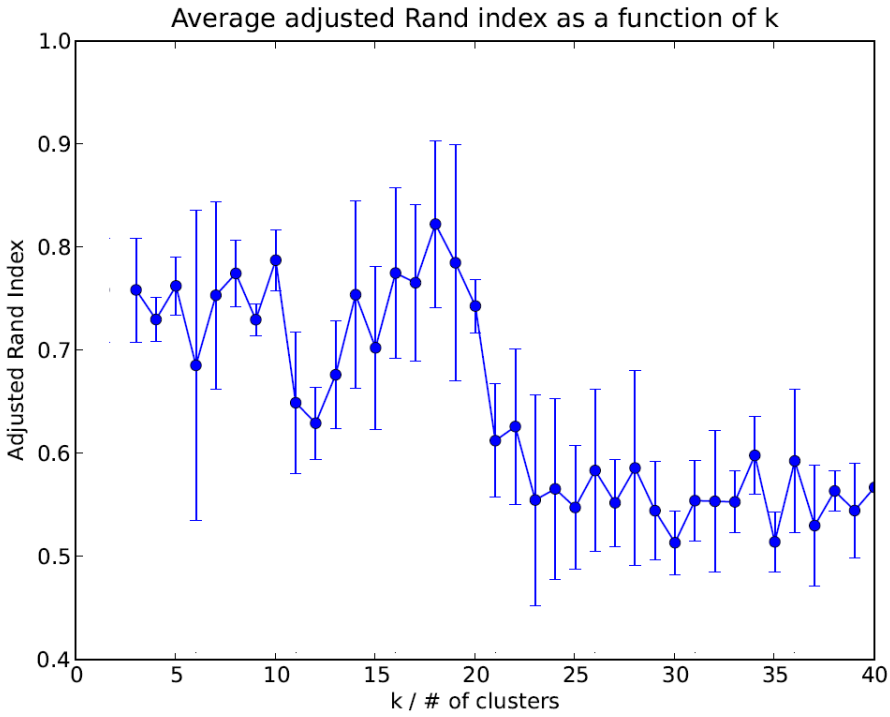
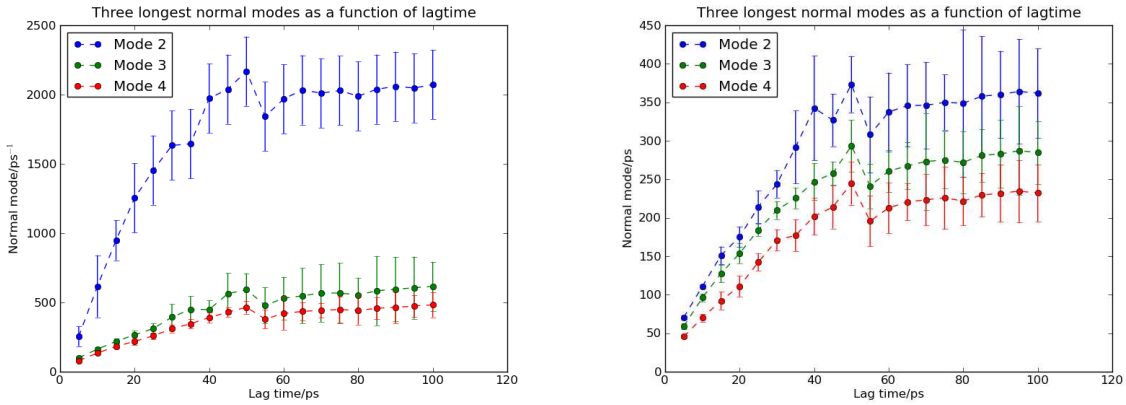


Figure 35: The adjusted Rand index depicted as a function of the k used in k-means clustering. For each k the average and standard deviation of the index is shown for a 20-fold bootstrap procedure.



(a) CF 250 pN.

(b) CF 500 pN.

Figure 36: The three longest normal modes for the 250 and 500 pN transition matrices as a function of lag-time used in matrix construction.

6.3.2.2 Ensuring Markovian properties

To determine the lag-time at which the 18-state Markov chain becomes memoryless, meaning the lag-time for which the implied timescale of the system converges, we calculate the implied time-scale of the model for a number of transition matrices. Figure 36 shows development in the three longest normal modes as a function of lag-time used for constructing the transition matrix for forces 250 and 500 pN, respectively. For each lag-time the standard deviation of the implied time-scale distribution from bootstrap generations of multiple transition matrices is shown as error bars. For both the 250 and 500 pN cases we observe a convergence in the time-scale at lag-time 50 ps, in further analysis we will use this lag-time when analyzing the dynamics of the system.

6.3.2.3 Unfolding networks at different forces

Although we have constructed MCMs for five different forces, our discussion here will mainly focus on a comparison of the networks for 250 and 500 pN as comparing these are representative of the key modifications that occur in the unfolding process when the pulling force is changed.

Figure 37 shows a graphical representation of the MCM constructed for the unfolding of I27 at 250 pN. Each node in the network represents a metastable state in the network, with the size of the node indicating the relative time spent in the state during the unfolding process. The label in each node serves as a unique ID for the state, which we will use as reference in our discussion of the network dynamics. Two special nodes are placed at the top and bottom of the network, labeled "Native" and "Unfolded," respectively, and serve to indicate the common starting point for all trajectories and the point at which the protein structure can be said to be completely unfolded. An edge in the network indicates that a transition between two states is possible, or, more specifically, that a none-zero entry between the states exists in the transition matrix representing the unfolding process.

Three different colorings of the network representing the mean distribution of the quantities N-,C-terminal distance, mode partitioning, and hydrogen bonding energy for each state have been used. In Figure 37(a) each state is shaded according to the average distance between the two Cα atoms of the terminal residues. In agreement with the general understanding of the unfolding process of I27 we observe that the nodes roughly fall into three distinct groups, an initial stage observed in states 14 and 15 with an extension around 46-47 A (close to that found in the native state), a large middle group with a distance in the interval 51 to 53, and a final stage

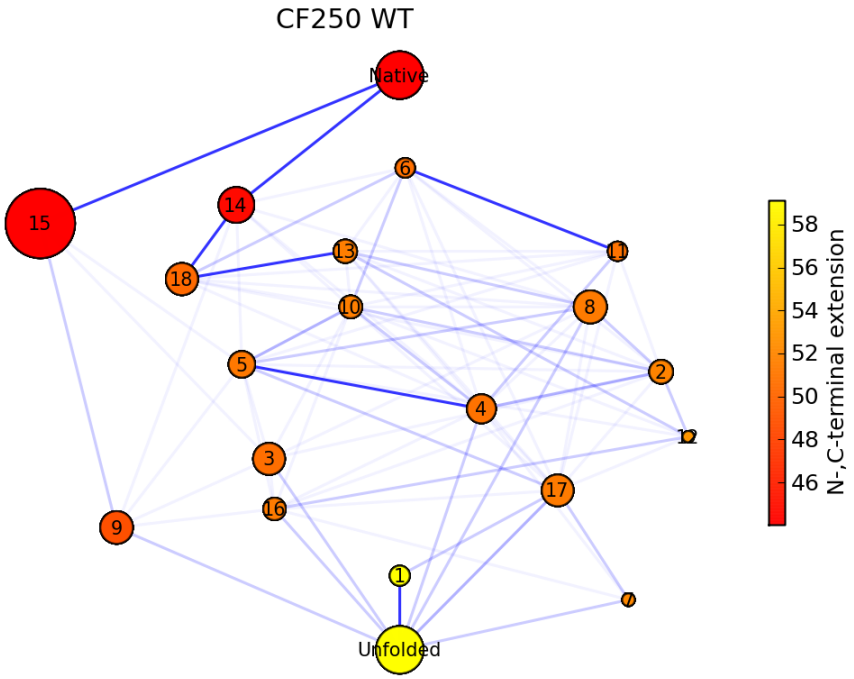
where the distance extends beyond 54 A (states 1,7,9,17). Likewise, Figure 37(c) displaying the average hydrogen bonding energy, agrees well with what we would expect from our current knowledge of I27 unfolding. We observe that the near native states have the lowest energy, with the energy increasing as we move closer towards the unfolded state. There are, however, two important observations to be made when inspecting the distribution of hydrogen bonding energy. First, we observe that the energy is more gradually diminished as we move closer to the unfolded state, it thus evident that even though no or little observable change is seen in the key reaction coordinate (the unfolding distance) there is a continuous structural rearrangement going on as we transition through the network with a resulting increase in hydrogen bonding energy. Second, the currently established model of mechanical protein stability focuses on a patch of five key hydrogen bonds between the A' and G strand being the main mechanical barrier preventing unfolding, we do, however, observe that many more structural elements play a role as the difference in hydrogen bond count between the native and unfolded is 38.

Finally, Figure 37(b) depicts the partitioning the of state-space according to modes of process obtained from the eigenstructure of the transition matrix. The first mode, corresponding to the slowest motion or largest energy barrier in the transition process, is colored in pink, the second slowest mode is colored in purple. The pattern of a major and a minor energy barrier is observed at all unfolding forces, albeit both are lowered when the pulling is increased.

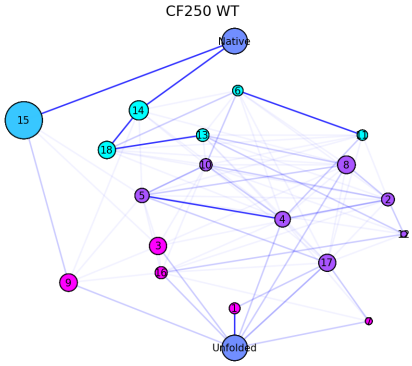
A comprehensive view of the common elements found in mode decomposition is given in Figure 38. Each node in the graph represents a state in the transition pathway with the numbering indicating the node id. The coloring of the nodes are according to which sides of

the major energy barriers they reside on. The slowest mode in the network is the transition between pink nodes on one side and green+purple nodes on the other, while the second slowest mode is defined by green and purple nodes, respectively. The blue and red nodes (nodes 15 and 9) are specific pathways only observed at low forces. Each edge indicates a transition across an energy barrier. For each transition the forward rates are shown for the forces 250, 300, 400, and 500 pN in blue, red, yellow, and purple, respectively. Unconnected nodes are states that only serve interior meta-states in the basin on one side of an energy barrier.

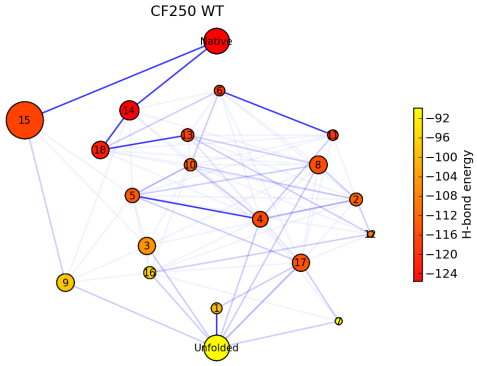
The transition rates given from the native state to nodes 15, 14, and 6 indicate the probability of the unfolding process starting in each of these states when force is first applied. It is evident that as the force is increased there is a redistribution from state 15 to 14. For instance, 37% of the trajectories generated at 250 pN start in state 15 a number that is reduced to 18% at 300 pN. As a result the propensity of trajectory starting points in state 14 go from 63% to 98% when the pulling force is increased from 250 to 500 pN. When inspecting the general trends of transition rates we observe that as the force is increased the transition rates across the energy barriers increase as well for all transition, albeit not by the same factor. State transitions are characterized by the same structural rearrangements regardless of pulling force, and thus the path through the energy landscape. The higher transitions rates make it clear that force lowers the height of the individual energy barriers between states, thereby making a transitions more likely at higher force.



(a) Coloring according to unfolding distance.



(b) Coloring according to mode partitioning.



(c) Coloring according to cummulative hydrogen bond energy.

Figure 37: Unfolding network for wild-type I27 pulled at 250 pN. Each node in the graphs indicate a metastable state, with the size of the node representing the relative time spend in the state during the unfolding process. An edge between two states indicates that a transition is possible, the darker the edge the more likely the transition. Two special nodes have been included at the top and bottom of the graph, labeled *Native* and *Unfolded*, respectively. *Native* indicated the starting point of all trajectories, whereas *Unfolded* is an artificial state all trajectories go to when their N-,C-terminal extension exceeds 64 A. The node coloring in the three networks represents the unfolding distance, the mode partitioning, and the average hydrogen bond energy for each state, respectively.

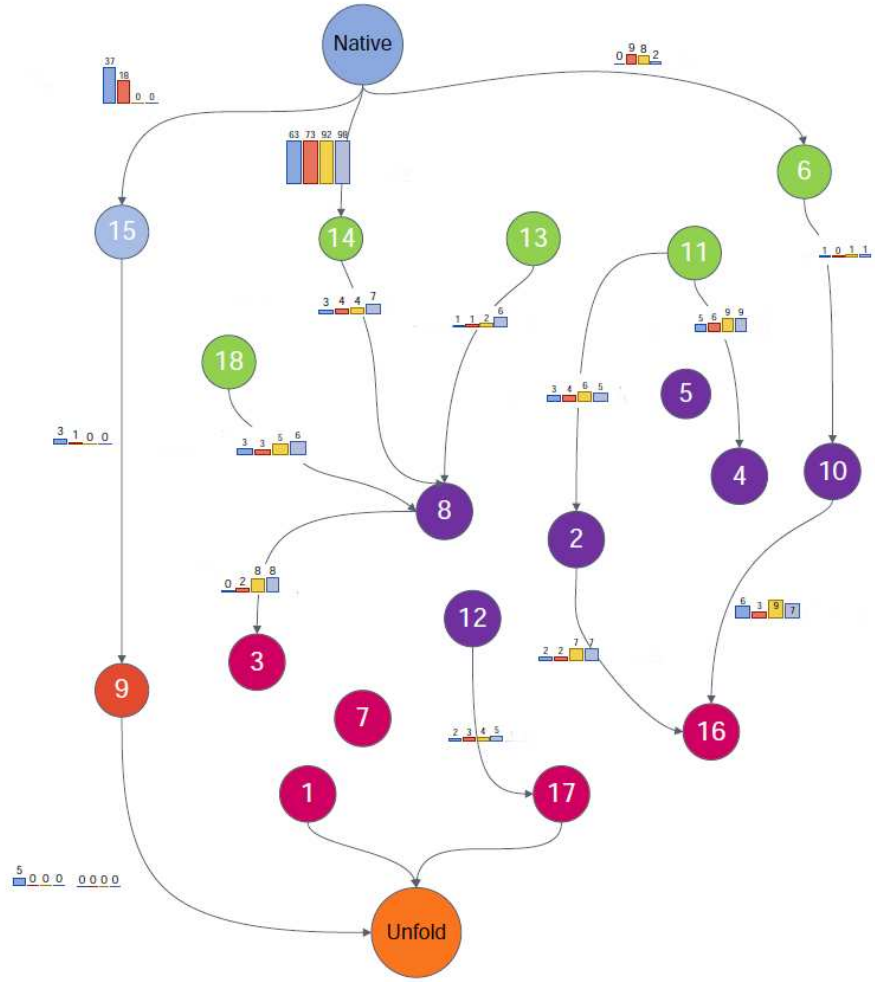


Figure 38: A unified representation of I27 unfolding at a number of pulling forces. Each node in the graph represents a state in the transition pathway with the numbering indicating the node id. The coloring of the nodes is according to which sides of the major energy barriers they reside on. The slowest mode in the network is the transition between pink nodes on one side and green+purple nodes on the other, while the second slowest mode is defines by green and purple nodes, respectively. The blue and red node (nodes 15 and 9) are specific pathway only observed a low forces. Each edge indicates a transition across an energy barrier. For each transition the forward rates are shown for the forces 250, 300, 400, and 500 pN in blue, red, yellow, and purple, respectively. Unconnected nodes are states that only serve interior meta-states in the basin on one side of an energy barrier.

6.3.2.4 Key transition pathways

Inspecting the flow across the key barriers reveals four key mechanisms of unfolding in wild-type I27, each dominating different ranges of the force spectrum.

A mechanism observed only at lower forces is the *Inter-strand drifting* unfolding pathway illustrated in Figure 39. This mode of unfolding proceeds through two states: First, state 15 represents only a minor rearrangement relative to the native state with all key hydrogen-bonds intact. The transition to state 9 is characterized by the drifting of 21 inter-strand bonds, weakening the interaction between most key bonds connection the loop regions of the parallel β-strands. This drifting procedure results in a moderate extension of the structure of 3 A putting further strain on the bonds connecting the A' and G strand eventually leading to complete unfolding from state 9.

While the inter-strand drifting is most prominent at lower forces, as force increases the *peeling of A* mechanism becomes gradually more important. As illustrated in Figure 40, this mode of unfolding is characterized by three transitions. First, the transition from the state 13, 14, and 18 (all on one side of the energy barrier characterizing the second slowest mode of unfolding) to 8 is characterized by an extension of the distance between the terminal structural residues to 52 A. This extension is brought about by the breakage of hydrogen-bonds between the A and B β-strand leading to a peeling of the A strand from the protein structure. The transition from state 8 to 3 does not lead to any further extension of the structure, but is characterized by the rearrangement of hydrogen-bonds between loop regions of the C, E and F strand. Finally, unfolding is observed as the A'-G- strand bonds are broken. Interestingly,

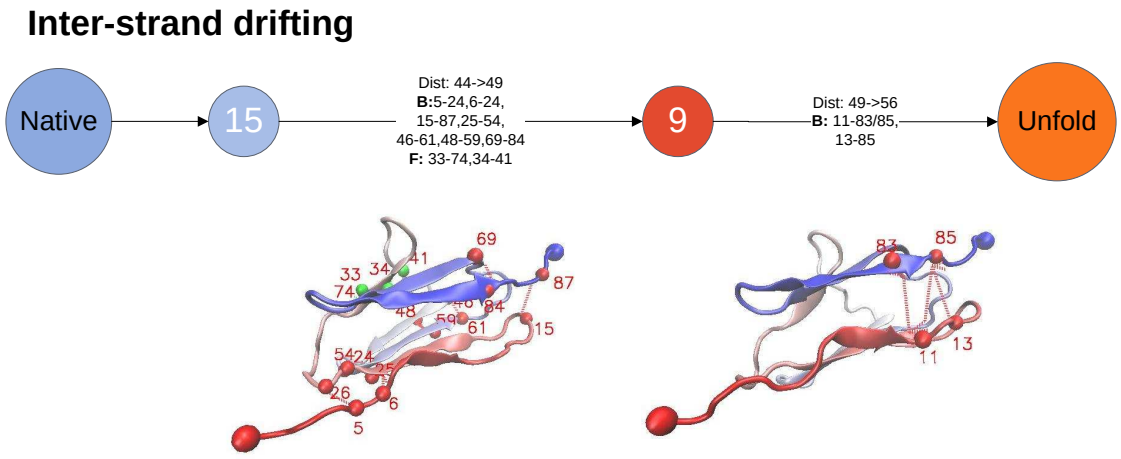


Figure 39: The inter-strand drifting unfolding pathway. Each state transition is characterized by three key metrics changing between the two states: Dist indicating the N-,C-terminal distance, B listing the residue-pairs between which hydrogen bonds are broken, and F listing the residue-pairs between which hydrogen bonds are formed. In the structural depiction of each transition, residues for which hydrogen bonds are broken and formed are highlighted in red and green, respectively.

the hydrogen-bonds between the A and B strand originally weakened are reformed in this step, likely due to mechanical strain being transferred to other regions of the structure upon the breakage of the central hydrogen-bond patch.

The *peeling of the A strand* pathway accounts for the majority of unfolding events, there is, however, a minor fraction that occur through the *Unravel from both ends* pathway illustrated in Figure 41, mainly at pulling forces of 300 and 400 pN. Unlike previous pathways described this mode of unfolding works by compromising the structural integrity through destabilization of hydrogen-bonds at both end of the structure simultaneously. First, the transition from state 6 to 10 results in a 1.5 A extension with weakening of hydrogen-bonds at both the N- and C-terminal loop regions. A further extension from 51.5 to 52 A is seen in the transition from state with further weakening of other loop-region hydrogen bonds, as well as the bond between residues 11 and 83 in the main hydrogen-bond patch of the structure. As observed in other pathways the final unfolding of the structure occurs through the breakage of the hydrogen-bond pairs 11-83, 11-85, 13-85, and 15-87. Similar to the *peeling of A strand* pathway, the breakage of the patch hydrogen-bond occur in conjunction with the re-connection of the A and B strand sheet.

The final transition pathway to unfolding depicted in Figure 42 differs from the three pathways described above by not being directly reachable from the native state. Thus transition into this pathway occurs through one of the other states on the same side of the transition barrier as state 11, highlighting the fact that 'cross-talk' is possible between different pathways. This pathway is characterized by a rearrangement of the support strand network prior to the

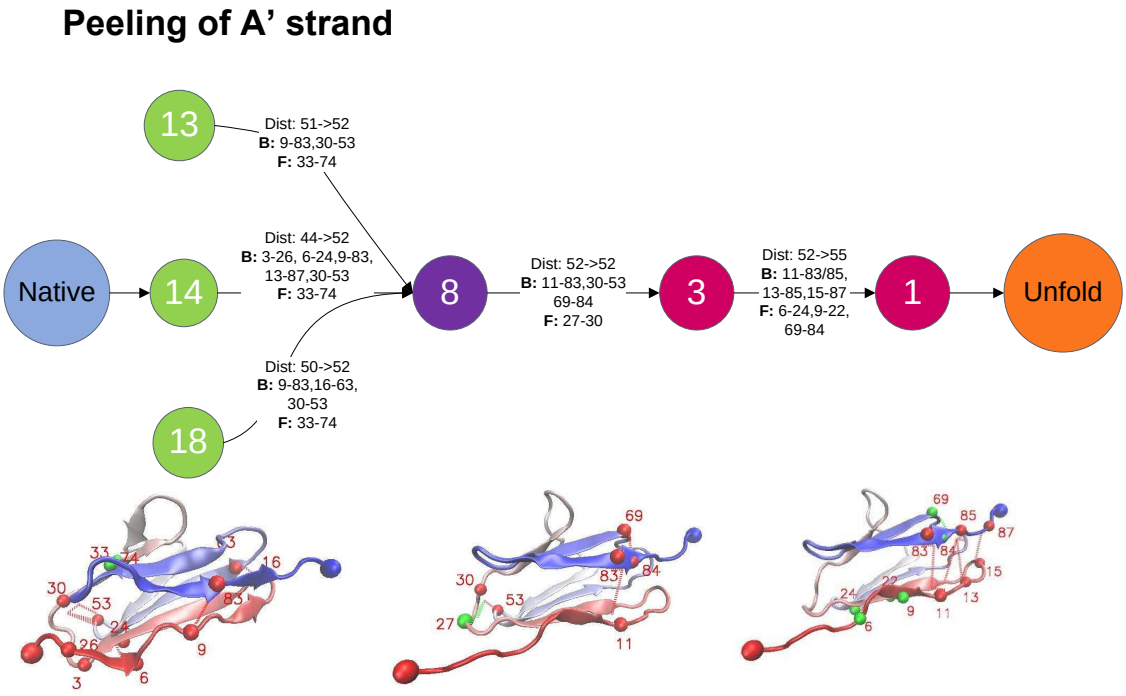


Figure 40: The peeling of A' strand unfolding pathway. Each state transition is characterized by three key metrics changing between the two states: Dist indicating the N-,C-terminal distance, B listing the residue-pairs between which hydrogen bonds are broken, and F listing the residue-pairs between which hydrogen-bonds are formed. In the structural depiction of each transition, residues for which hydrogen bonds are broken and formed are highlighted in red and green, respectively.

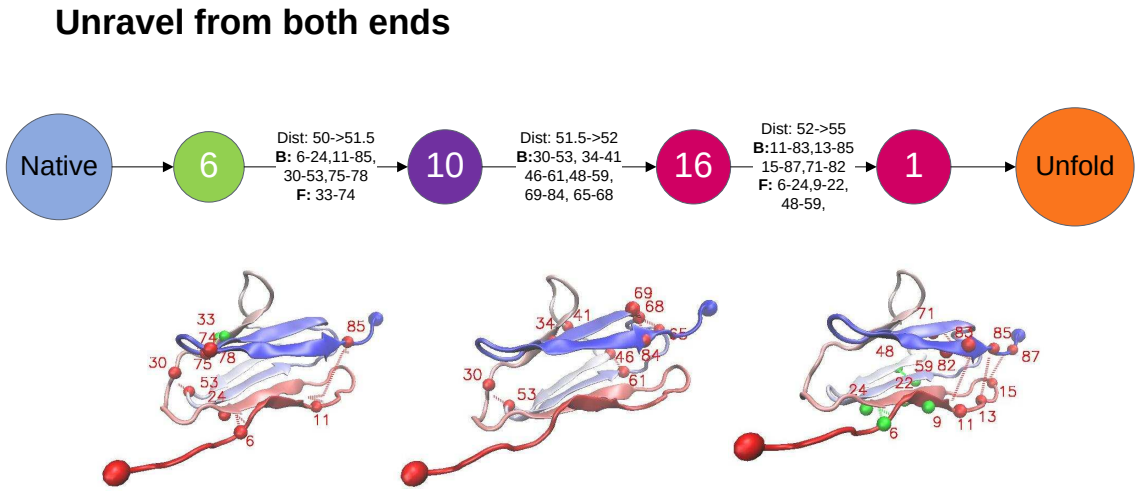


Figure 41: The unraveling from both ends unfolding pathway. Each state transition is characterized by three key metrics changing between the two states: Dist indicating the N-,C-terminal distance, B listing the residue-pairs between which hydrogen bonds are broken, and F listing the residue-pairs between which hydrogen-bonds are formed. In the structural depiction of each transition, residues for which hydrogen bonds are broken and formed are highlighted in red and green, respectively.

breakage of the key hydrogen-bond patch. Specifically, the transition from state 11 to states 2 and 4 is characterized in formation of a hydrogen-bond between 33 and 74 as well as weakening of bond between residue pairs 6-24, 17-63, and 30-53, all of which are outside the core bonds between the A/A' and G strands. The transition from state 4 to 16 sees a further weakening of non-core hydrogen bonds as well as the patch bond 15-87. The structural weakening induced by the rearrangement of the support strands shift the mechanical load onto the patch hydrogen-bonds leading to unfolding from state 16.

6.3.3 Explaining the effect of the Y9P mutant on the mechanical stability of I27

As demonstrated in the MCM for the unfolding process of wild-type I27 at different forces, the energy landscape of the protein structure is altered significantly at different modes of pulling, leading to the preference of different unfolding pathways. We saw previously that SMD simulations correctly predicted the relative strength of several I27 mutants. When interpreting the effect of the two Pro mutants V13P and V15P in context of the unfolding pathways presented above, it is not surprising that these mutant do not withstand force as well as wild-type I27. Both of these mutations are of residues that are part of the central hydrogen-bond patch of the structure, the breakage of which constitutes the key mechanical barrier to unfolding. The replacement of Valine with Proline removes a backbone hydrogen-bond in the patch thus effectively lowering the energy barrier preventing unfolding. The mutant V86A is also observed to have lower mechanical resistance than wild-type I27. While the mutated residue is not directly involved in any of the hydrogen-bonds constituting the mechanical barriers its surrounding

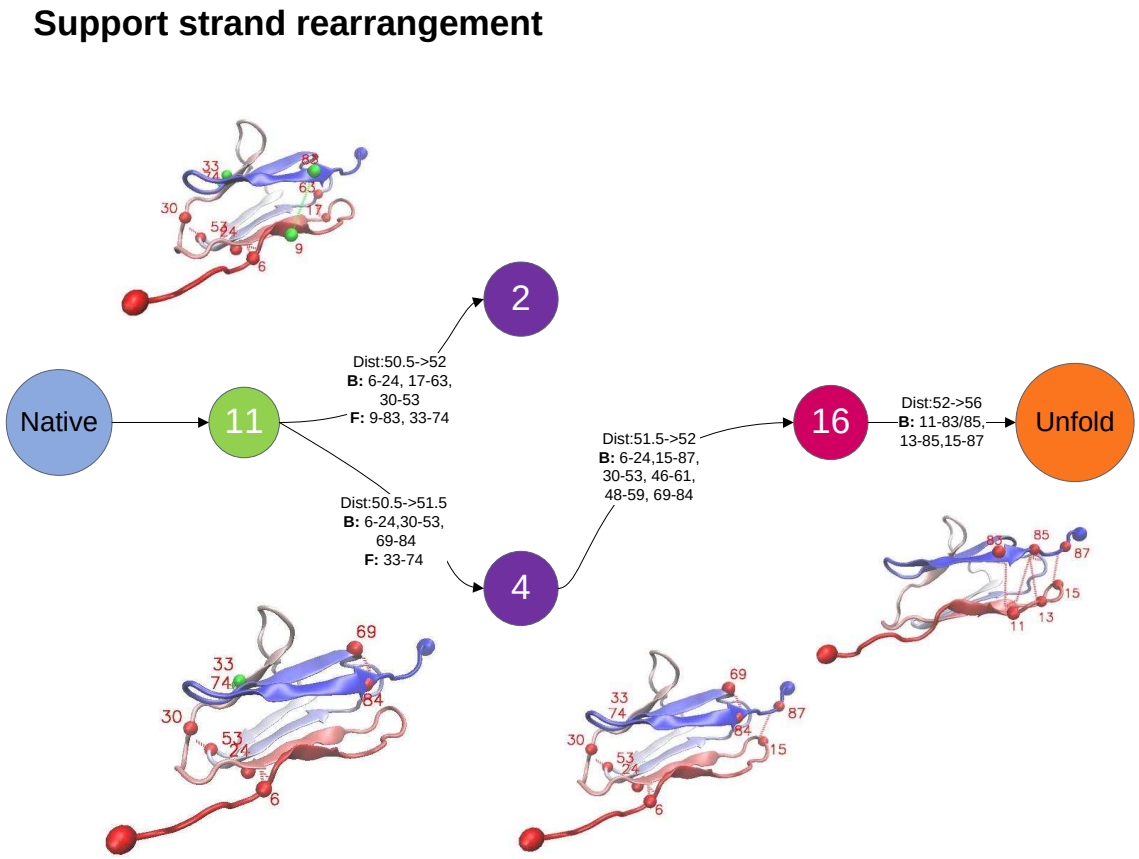


Figure 42: The support strand rearrangement unfolding pathway. Each state transition is characterized by three key metrics changing between the two states: Dist indicating the N-,C-terminal distance, B listing the residue-pairs between which hydrogen bonds are broken, and F listing the residue-pairs between which hydrogen-bonds are formed. In the structural depiction of each transition, residues for which hydrogen bonds are broken and formed are highlighted in red and green, respectively.

residues 83,85,87,68, and 69 are all key in the mechanical resistance of the structure, thus it is not surprising that the introduction of a novel amino-acid in this position could interfere with the hydrogen-bonding network formed by these residues.

The most interesting of the four mutants is Y9P as it is observed to make the structure stronger. One would expect that the introduction of a Pro residue at residue 9 would lower the energy barrier preventing unfolding as there would one fewer hydrogen-bond to break (the bond between residues 9 and 83 would be eliminated).

To uncover the change in unfolding pathway leading to a stronger structure we constructed a second MCM model based on simulations from all wild-type and mutant trajectories. This model has 23 metastable states indicating that the mutant structures do indeed explore additional structural conformation in their unfolding pathways compared with the wild-type structure. Figure 43 illustrates the most common unfolding pathway observed at a pulling force of 400 pN (the mutated residue is marked in yellow). The pathway differs from wild-type modes of unfolding by leaving hydrogen-bonds formed by the A strand between the B and G strand, respectively, intact in the initial stage. Instead we observe a marked structural rearrangement of the E-F and C-F strand loop regions, eventually undermining the integrity of the top part of the structure, resulting in the G strand being pulled free of the A strand and thus breaking the main hydrogen-bond patch. We cannot say for sure why the bonds formed by the A strand appear to be better preserved, thereby leading to an alternative (and more mechanically stable) mode of unfolding. It could, however, be hypothesized that the introduction of the more rigid Pro residue in place of the substantially larger Tyr residue leads to a more stable strand

arrangement, due to less thermal fluctuation being picked up by this smaller residue. Thus the hydrogen-bond arrangement sees less random fluctuation and can consequently withstand mechanical strain better.

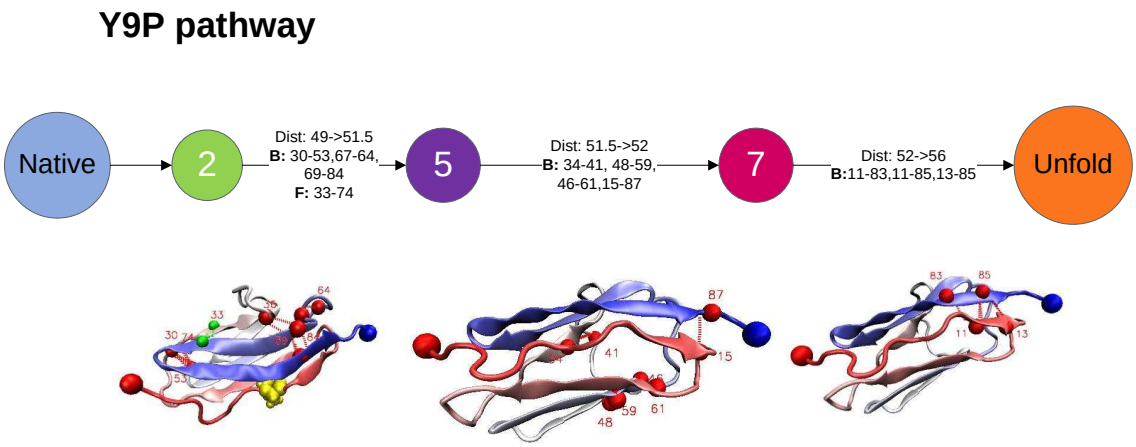


Figure 43: The Y9P unfolding pathway. Each state transition is characterized by three key metrics changing between the two states: Dist indicating the N-,C-terminal distance, B listing the residue-pairs between which hydrogen bonds are broken, and F listing the residue-pairs between which hydrogen-bonds are formed. In the structural depiction of each transition, residues for which hydrogen bonds are broken and formed are highlighted in red and green, respectively. The mutated residue is marked in yellow in the first structure cartoon diagram.

6.4 Discussion and Conclusion

We have developed a method for modeling the changes in single molecule dynamics introduces by a signaling event as a discrete state Markov Chain model. Specifically, we use

the partial unfolding of so-called mechanical proteins by ways of steered molecular dynamics to demonstrate how the protein energy landscape is altered when different external mechanical forces are applied. By probing the protein structure with a range of forces, we show that the transitions pathways taking the protein structure from folded to partially unfolded vary depending on the external input. The constructed model is instrumental in uncovering the specific structural changes associated with unfolding, allowing us to pin-point the specific residue-interactions responsible for global structural properties.

We demonstrated how SMD simulations correctly predict the mechanical strength of experimentally characterized I27 mutants, and utilized a MCM constructed from several mutant trajectories to characterize the main unfolding pathway of the Y9P-I27 mutant. In general, the representation of the unfolding process as a network of discrete state with transitions characterized by a the change in a few structural elements, allows us to suggest novel mutants that will affect the global unfolding rate of the structure, thus providing a tool for guiding the design of protein structures with specifically tuned mechanical properties. In example, the interaction between residues 30 and 53 is observed to be broken in several of the unfolding pathways from wild-type I27, while not being in the main hydrogen-bond patch constituting the major energy barrier of the unfolding process. As preliminarily test we mutate residues 30 to Pro (mutant V30P), thus removing any potential backbone hydrogen-bonds between the residue 30 and 53, and pull the structure at 400 pN (7 samples). We do indeed observe slightly faster unfolding rates than those in wild-type I27, though the decrease in unfolding time (mean unfolding time

for the mutant 2278 ps) is not as dramatic as that observed in the other mutant structures investigated.

In sum, the presented framework may be instrumental in characterizing the specific properties of structurally similar mechanical proteins responsible their mechanical strength, revealing how structural elements play together in forming global mechanical properties. In principle, a similar framework could be used for investigating the dynamics of structural change occurring during any type of signaling event, assuming the system of interest can be sufficiently sampled given the currently accessible simulation time-scales in MD studies. Alternatively, techniques such as targeted molecular dynamics where a biased force-fields is used to drive the protein of interest from it native conformation to a predefined end conformation may be used to reduce the simulation time required for large system.

As the growth in computational power makes longer MD simulation time-scales feasible, representing the energy-landscape dynamics of biological macromolecules as a network of discrete states may become a standard concept in the analysis of protein dynamics. There are, however, two key challenges that need addressed further to make this type of method generally applicable. First, while observing the development of implied time-scales may give us an indication of when a system becomes Markovian, there is no theoretical guarantee that the system is indeed memoryless when convergence occurs. It would be desirable to have a method providing a more rigorous test of this property. Second, in this work we assume that the contact space of the protein structure can be used to approximate its energy-state and thus identify the meta-stable state of the unfolding process. For the purpose of our application this approximation seems to

be justified. This may, however, not always be the case, thus it would be desirable to have a general purpose representation of the protein structure on which an algorithm for discovering meta-stable states could be based.

CHAPTER 7

CONCLUSION

The study of the molecular interactions that make up signal transduction pathways are key in understanding the regulation of cellular function. The research described in this thesis was carried out to address a number of major challenges in the study of signal transduction mechanisms using proteomics data. We have presented computational methods addressing three key challenges in the quest to construct a more complete picture of protein signaling pathways, namely, confident identification of proteins in a sample, functional classification of large-scale proteomics data, and characterization of the dynamic conformational changes in protein structures.

First, we developed a probabilistic protocol for identification of short peptide fragments characterized by tandem mass-spectromety (MS/MS). A machine learning procedure for correctly matching peptides with mass spectra was constructed. Further, we demonstrated how the developed model can be represented as an interpretable tree of rules, thereby effectively removing the 'black-box' notion often associated with machine learning classifiers, making the underlying model clearer to end-users. Finally, using a probabilistic framework, a method for protein identification based on the peptide predictions was proposed and tested.

Second, a genome-wide functional classification protocol for identifying dual specificity membrane- and protein-binding domains was developed. Experimental characterization of 90 PDZ domains showing that 40% had submicromolar membrane affinity was used for building

a model utilized to predict the membrane binding properties of 2000 PDZ domains from 20 species. We demonstrated that reversible membrane binding is a key component in the spatially regulation protein interaction networks and further proposed a mechanistic classification of dual-specificity binding. As an extension to the PDZ domain models, we build a knowledge-mining procedure for learning the general mechanisms of membrane-binding, using C1, C2, and PH domains as test-beds. We demonstrated how this method was able to uncover properties of each family known to be important for binding.

Last, we presented a method for modeling the changes in single molecule dynamics induced by a signaling event as a discrete state Markov Chain model. Specifically, we used the partial unfolding of so-called mechanical proteins by way of steered molecular dynamics to demonstrate how the protein energy landscape is altered when different external mechanical forces are applied. By probing the protein structure with a range of forces, we show that the transitions pathways taking the protein structure from folded to partially unfolded vary significantly depending on the external input. The constructed model is instrumental in explaining experimental single molecule studies of the unfolding of the protein domain I27, as well as the changes in mechanical properties of a number of I27 mutant structures.

Many pathological conditions are the result of changes to signal transduction systems. In some instances we are already able to provide treatment by administering drugs targeting signaling proteins such as protein kinases (171). Our ability to further the development of such treatment regimens depends on the availability of accurate functional prediction methods regarding the role of proteins and the effects of modifying the protein networking properties in

specific systems. It is the hope that the computational models presented in this thesis will be a contribution towards this final goal.

VITA

Morten Källberg

Education

University of Illinois at Chicago , Chicago, IL
Ph.D., Bioinformatics, (Aug 07 - August 12)
Dissertation: Computational Investigation of Signaling Regimens using Proteomics Data.

University of Southern Denmark , Odense, Denmark
B.Sc., Bioinformatics, (Aug 03 -June 06)

Awards

1. **Fulbright Fellowship** Fulbright Foundation/University of Illinois (2006-7)
2. **FMC Fellowship** FMC Technologies Fund (2009-10, 2010-11, and 2011-12)
3. **ISCB Travel Fellowship** National Science Foundation (2010)
4. **Erasmus Travel Grant** (2006)

Peer-reviewed Publications

1. **Källberg** and Lu. *Reconstructing the Unfolding Pathways of Titin I27 from Steered Molecular Dynamics Simulations.* (In preparation)
2. **Källberg** and Lu. *A structure based protocol for learning the family specific mechanisms of membrane binding domains.* Bioinformatics. 2012. (Accepted).
3. **Källberg**, Peng, Zhiang, Lu, Xu. *RaptorX: A protein function and structure prediction protocol.* Nature Protocols. Nature Protocols 7, 1511-1522. 2012.
4. Chen/Sheng/**Källberg***, Silkov, Tun, Bhardwaj, Kurilova, Hall, Honig, Lu, Cho. *Genome-Wide Identification and Functional Annotation of Dual Specificity Protein- and Lipid-Binding Modules That Modulate Protein Interactions at the Membrane.* Molecular Cell. April 27., 2012.
   *Authors contributed equally to this work
5. **Källberg** and Lu. *An improved machine learning protocol for the identification of correct Sequest search results.* BMC Bioinformatics. 11:591. 2010.
6. Yoon, Tong, Lee, Albanese, Bhardwaj, **Källberg**, Digman, Lu, Gratton, Shin, Cho. *Molecular basis of the potent membrane remodeling activity of the epsin-1 ENTH domain.* Journal of Biological Chemistry, 285, 531-540. 2009.
7. Genchev, **Källberg**, Gursoy, Mittal, Dubey, Perisic, Fang and Lu. *Mechanical Signaling on the Single Protein Level Studied Using Steered Molecular Dynamics.* Cell Biochemistry and Biophysics. 1085-91,95. 2009.

8. **Källberg** and Lu. *Structural Feature Extraction Protocol for Classifying Reversible Membrane Binding Protein Domains.* Conf Proc IEEE Eng Med Biol Soc. 6735-8. 2009.

Book Chapters

1. Bhardwaj, **Källberg**, Cho, Lu. *MeTaDoR: Online Resource and Prediction Server for Membrane Targeting Peripheral Proteins.* In: Algorithmic and AI Methods for Protein Bioinformatics. Yi Pan and Albert Y. Zomaya ed. New Jersey: Wileys. 2012.

Talks and Presentations

1. **Källberg** and Lu. *Learning from structure and sequence.* IEEE Eng Med., Minneapolis, 2010. Platform.
2. **Källberg** and Lu. *A Machine Learning Protocol for Distinguishing Intrafamily Peripheral Membrane-Targeting Domain Properties using Sequence and Structure.* Biophysical society meeting, Boston, 2009. Platform.

CITED LITERATURE

1. R. Aebersold and D. R. Goodlett. Mass spectrometry in proteomics. *Chem Rev*, 101(2):269–295, Feb 2001.

2. Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

3. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

4. D. C. Anderson, Weiqun Li, Donald G Payan, and William Stafford Noble. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide ms/ms spectra and sequest scores. *J Proteome Res*, 2(2):137–146, 2003.

5. A. Bairoch. The enzyme database in 2000. *Nucleic Acids Res*, 28(1):304–305, Jan 2000.

6. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, Oct 2001.

7. M. M. Balamurali, Deepak Sharma, Anderson Chang, Dingyue Khor, Ricky Chu, and Hongbin Li. Recombination of protein fragments: a promising approach toward engineering proteins with novel nanomechanical properties. *Protein Sci*, 17(10):1815–1826, Oct 2008.

8. Nitin S Baliga, Min Pan, Young Ah Goo, Eugene C Yi, David R Goodlett, Krassen Dimitrov, Paul Shannon, Ruedi Aebersold, Wailap Victor Ng, and Leroy Hood. Coordinate regulation of energy transduction modules in halobacterium sp. analyzed by a global systems approach. *Proc Natl Acad Sci U S A*, 99(23):14913–14918, Nov 2002.

9. A. Ball, R. Nielsen, M. H. Gelb, and B. H. Robinson. Interfacial membrane docking of cytosolic phospholipase a2 c2 domain using electrostatic potential-modulated spin relaxation magnetic resonance. *Proc Natl Acad Sci U S A*, 96(12):6637–6642, Jun 1999.

10. Helen M Berman, Tammy Battistuz, T. N. Bhat, Wolfgang F Bluhm, Philip E Bourne, Kyle Burkhardt, Zukang Feng, Gary L Gilliland, Lisa Iype, Shri Jain, Phoebe Fagan, Jessica Marvin, David Padilla, Veerasamy Ravichandran, Bohdan Schneider, Narmada Thanki, Helge Weissig, John D Westbrook, and Christine Zardecki. The protein data bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):899–907, Jun 2002.

11. Marshall Bern, David Goldberg, W. Hayes McDonald, and John R Yates. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*, 20 Suppl 1:i49–i54, Aug 2004.

12. Robert B Best, Susan B Fowler, Jos L Toca Herrera, Annette Steward, Emanuele Paci, and Jane Clarke. Mechanical unfolding of a titin ig domain: structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J Mol Biol*, 330(4):867–877, Jul 2003.

13. Nitin Bhardwaj, Robert Langlois, Guijun Zhao, and Hui Lu. Structure based prediction of binding residues on dna-binding proteins. *Conf Proc IEEE Eng Med Biol Soc*, 3:2611–2614, 2005.

14. Nitin Bhardwaj, Robert E Langlois, Guijun Zhao, and Hui Lu. Kernel-based machine learning protocol for predicting dna-binding proteins. *Nucleic Acids Res*, 33(20):6486–6493, 2005.

15. Nitin Bhardwaj and Hui Lu. Residue-level prediction of dna-binding sites and its application on dna-binding protein predictions. *FEBS Lett*, 581(5):1058–1066, Mar 2007.

16. Nitin Bhardwaj, Robert V Stahelin, Robert E Langlois, Wonhwa Cho, and Hui Lu. Structural bioinformatics prediction of membrane-binding proteins. *J Mol Biol*, 359(2):486–495, Jun 2006.

17. Nitin Bhardwaj, Robert V Stahelin, Guijun Zhao, Wonhwa Cho, and Hui Lu. Metador: a comprehensive resource for membrane targeting domains and their host proteins. *Bioinformatics*, 23(22):3110–3112, Nov 2007.

18. Roby P Bhattacharyya, Attila Remnyi, Brian J Yeh, and Wendell A Lim. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem*, 75:655–680, 2006.

19. A. Biegert and J. Sding. Sequence context-specific profiles for homology searching. *Proc Natl Acad Sci U S A*, 106(10):3770–3775, Mar 2009.

20. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin., 2006.

21. Nichole R Blatner, Robert V Stahelin, Karthikeyan Diraviyam, Phillip T Hawkins, Wanjin Hong, Diana Murray, and Wonhwa Cho. The molecular basis of the differential subcellular localization of fyve domains. *J Biol Chem*, 279(51):53818–53827, Dec 2004.

22. Relly Brandman, Marie-Hlne Disatnik, Eric Churchill, and Daria Mochly-Rosen. Peptides derived from the c2 domain of protein kinase c epsilon (epsilon pkc) modulate epsilon pkc activity and identify potential protein-protein interaction surfaces. *J Biol Chem*, 282(6):4113–4123, Feb 2007.

23. D. Bray. Signaling complexes: biophysical constraints on intracellular communication. *Annu Rev Biophys Biomol Struct*, 27:59–75, 1998.

24. L Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

25. L. Breiman. Random forrest. *Machine Learning*, 45:5–32, 2001.

26. Anthony Bretscher, Kevin Edwards, and Richard G Fehon. Erm proteins and merlin: integrators at the cell cortex. *Nat Rev Mol Cell Biol*, 3(8):586–599, Aug 2002.

27. Michal Brylinski and Jeffrey Skolnick. Findsite: a threading-based approach to ligand homology modeling. *PLoS Comput Biol*, 5(6):e1000405, Jun 2009.

28. Pietro De Camilli, Hong Chen, Joel Hyman, Ezequiel Panepucci, Alex Bateman, and Axel T Brunger. The enth domain. *FEBS Lett*, 513(1):11–18, Feb 2002.

29. Yi Cao, Teri Yoo, and Hongbin Li. Single molecule force spectroscopy reveals engineered metal chelation is a general approach to enhance mechanical stability of proteins. *Proc Natl Acad Sci U S A*, 105(32):11152–11157, Aug 2008.

30. Kilpatrick Carroll, Carlos Gomez, and Lawrence Shapiro. Tubby proteins: the plot thickens. *Nat Rev Mol Cell Biol*, 5(1):55–63, Jan 2004.

31. David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The

amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–1688, Dec 2005.

32. Yong Chen, Ren Sheng, Morten Kllberg, Antonina Silkov, Moe P Tun, Nitin Bhardwaj, Svetlana Kurilova, Randy A Hall, Barry Honig, Hui Lu, and Wonhwa Cho. Genome-wide functional annotation of dual-specificity protein- and lipid-binding modules that regulate protein interactions. *Mol Cell*, 46(2):226–237, Apr 2012.

33. W. Cho. Membrane targeting by c1 and c2 domains. *J Biol Chem*, 276(35):32407–32410, Aug 2001.

34. W. Cho, L. Bittova, and R. V. Stahelin. Membrane binding assays for peripheral proteins. *Anal Biochem*, 296(2):153–161, Sep 2001.

35. Wonhwa Cho. Building signaling complexes at the membrane. *Sci STKE*, 2006(321):pe7, Feb 2006.

36. Wonhwa Cho and Robert V Stahelin. Membrane-protein interactions in cell signaling and membrane trafficking. *Annu Rev Biophys Biomol Struct*, 34:119–151, 2005.

37. John D Chodera, Nina Singhal, Vijay S Pande, Ken A Dill, and William C Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J Chem Phys*, 126(15):155101, Apr 2007.

38. Chia-Lin Chyan, Fan-Chi Lin, Haibo Peng, Jian-Min Yuan, Chung-Hung Chang, Sheng-Hsien Lin, and Guoliang Yang. Reversible mechanical unfolding of single ubiquitin molecules. *Biophys J*, 87(6):3995–4006, Dec 2004.

39. Marek Cieplak, Trinh Xuan Hoang, and Mark O Robbins. Folding and stretching in a go-like model of titin. *Proteins*, 49(1):114–124, Oct 2002.

40. Marek Cieplak and Piotr E Marszalek. Mechanical unfolding of ubiquitin molecules. *J Chem Phys*, 123(19):194903, Nov 2005.

41. Francheska Coln-Gonzlez and Marcelo G Kazanietz. C1 domains exposed: from diacylglycerol binding to protein-protein interactions. *Biochim Biophys Acta*, 1761(8):827–837, Aug 2006.

42. Senena Corbaln-Garcia, Susana Snchez-Carrillo, Josefa Garca-Garca, and Juan C Gmez-Fernndez. Characterization of the membrane binding mode of the c2 domain of pkc epsilon. *Biochemistry*, 42(40):11661–11668, Oct 2003.

43. Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

44. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

45. Robertson Craig and Ronald C Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, Jun 2004.

46. Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *the 23rd International Conference on Machine learning, Pittsburgh, Pennsylvania*, 2006.

47. Carmen L de Hoog and Matthias Mann. Proteomics. *Annu Rev Genomics Hum Genet*, 5:267–293, 2004.

48. Hendrik Dietz and Matthias Rief. Exploring the energy landscape of gfp by single-molecule mechanical experiments. *Proc Natl Acad Sci U S A*, 101(46):16192–16197, Nov 2004.

49. Jonathan P DiNitto, Thomas C Cronin, and David G Lambright. Membrane recognition and targeting by lipid-binding domains. *Sci STKE*, 2003(213):re16, Dec 2003.

50. Jonathan P DiNitto and David G Lambright. Membrane and juxtamembrane targeting by ph and ptb domains. *Biochim Biophys Acta*, 1761(8):850–867, Aug 2006.

51. O. du Roure, A. Buguin, H. Feracci, and P. Silberzan. Homophilic interactions between cadherin fragments at the single molecule level: an afm study. *Langmuir*, 22(10):4680–4684, May 2006.

52. Joshua E Elias, Francis D Gibbons, Oliver D King, Frederick P Roth, and Steven P Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol*, 22(2):214–219, Feb 2004.

53. Jimmy K. Eng, Ashley L. McCormack, and John R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–989, 1994.

54. Adam J Engler, Maureen A Griffin, Shamik Sen, Carsten G Bnnemann, H. Lee Sweeney, and Dennis E Discher. Myotubes differentiate optimally on substrates with tissue-like stiffness: pathological implications for soft or stiff microenvironments. *J Cell Biol*, 166(6):877–887, Sep 2004.

55. Jan Eriksson and David Feny. Probity: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J Proteome Res*, 3(1):32–36, 2004.

56. Eran Eyal and Ivet Bahar. Toward a molecular understanding of the anisotropic response of proteins to external forces: insights from elastic network models. *Biophys J*, 94(9):3424–3435, May 2008.

57. J. A. Falkner, J. W. Falkner, and P. C. Andrews. Proteomecommons.org jaf: reference information and tools for proteomics. *Bioinformatics*, 22(5):632–633, Mar 2006.

58. Jayson A Falkner, Maureen Kachman, Donna M Veine, Angela Walker, John R Strahler, and Philip C Andrews. Validated maldi-tof/tof mass spectra for protein standards. *J Am Soc Mass Spectrom*, 18(5):850–855, May 2007.

59. Jianwen Fang, Yinghua Dong, Todd D Williams, and Gerald H Lushington. Feature selection in validating mass spectrometry database search results. *J Bioinform Comput Biol*, 6(1):223–240, Feb 2008.

60. Wei Feng and Mingjie Zhang. Organization and dynamics of pdz-domain-related supramodules in the postsynaptic density. *Nat Rev Neurosci*, 10(2):87–99, Feb 2009.

61. K. M. Ferguson, J. M. Kavran, V. G. Sankaran, E. Fournier, S. J. Isakoff, E. Y. Skolnik, and M. A. Lemmon. Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains. *Mol Cell*, 6(2):373–384, Aug 2000.

62. Julio M Fernandez and Hongbin Li. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, Mar 2004.

63. A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–1773, Sep 2000.

64. Peter L Freddolino, Sanghyun Park, Benot Roux, and Klaus Schulten. Force field bias in protein folding simulations. *Biophys J*, 96(9):3772–3780, May 2009.

65. Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *the 16th International Conference on Machine Learning, Bled, Slovenia*, 1999.

66. Yoav Freund and Rovert E. Schapire. Experiments with a new boosting algorithm. In *13th International Conference on Machine Learning, Bari, Italy*, 1996.

67. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting. Technical report, Department of Statistics Stanford University, 1998.

68. Rodrigo Gallardo, Ylva Ivarsson, Joost Schymkowitz, Frdric Rousseau, and Pascale Zimmermann. Structural diversity of pdz-lipid interactions. *Chembiochem*, 11(4):456–467, Mar 2010.

69. Mu Gao, Hui Lu, and Klaus Schulten. Unfolding of titin domains studied by molecular dynamics simulations. *J Muscle Res Cell Motil*, 23(5-6):513–521, 2002.

70. Hlne Gary-Gouy, Julie Harriague, Ali Dalloul, Emmanuel Donnadieu, and Georges Bismuth. Cd5-negative regulation of b cell receptor signaling pathways originates from tyrosine residue y429 outside an immunoreceptor tyrosine-based inhibitory motif. *J Immunol*, 168(1):232–239, Jan 2002.

71. Anne-Claude Gavin, Markus Bsche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jrg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Hfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, Jan 2002.

72. Paul Grayson, Emad Tajkhorshid, and Klaus Schulten. Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. *Biophys J*, 85(1):36–48, Jul 2003.

73. Tina Guina, Samuel O Purvine, Eugene C Yi, Jimmy Eng, David R Goodlett, Ruedi Aebersold, and Samuel I Miller. Quantitative proteomic analysis indicates increased synthesis of a quinolone by pseudomonas aeruginosa isolates from cystic fibrosis airways. *Proc Natl Acad Sci U S A*, 100(5):2771–2776, Mar 2003.

74. Bianca Habermann. The bar-domain family of proteins: a case of bending and binding? *EMBO Rep*, 5(3):250–255, Mar 2004.

75. Gregory Hannum, Rohith Srivas, Aude Gunol, Haico van Attikum, Nevan J Krogan, Richard M Karp, and Trey Ideker. Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet*, 5(12):e1000782, Dec 2009.

76. Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, Dec 2007.

77. Geoffrey Hinton and Terrence J. Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. MIT Press, 1999.

78. Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Sren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R Willems, Holly Sassi, Peter A Nielsen, Karina J Rasmussen, Jens R Andersen, Lene E Johansen, Lykke H Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D Srensen, Jesper Matthiesen, Ronald C Hendrickson, Frank Gleeson, Tony Pawson, Michael F Moran, Daniel Durocher, Matthias Mann, Christopher W V Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, Jan 2002.

79. Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, Mar 2010.

80. W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, Feb 1996.

81. J. H. Hurley and S. Misra. Signaling and subcellular targeting by membrane-binding domains. *Annu Rev Biophys Biomol Struct*, 29:49–79, 2000.

82. James H Hurley. Membrane binding domains. *Biochim Biophys Acta*, 1761(8):805–811, Aug 2006.

83. Wei Jiang, David J Hardy, James C Phillips, Alexander D Mackerell, Klaus Schulten, and Benot Roux. High-performance scalable molecular dynamics simulations of a polarizable force field based on classical drude oscillators in namd. *J Phys Chem Lett*, 2(2):87–92, 2011.

84. Brett R Johnson, Ryan T Nitta, Richard L Frock, Leslie Mounkes, David A Barbie, Colin L Stewart, Ed Harlow, and Brian K Kennedy. A-type lamins regulate retinoblastoma protein function by promoting subnuclear localization and preventing proteasomal degradation. *Proc Natl Acad Sci U S A*, 101(26):9677–9682, Jun 2004.

85. Colin P Johnson, Hsin-Yao Tang, Christine Carag, David W Speicher, and Dennis E Discher. Forced unfolding of proteins within cells. *Science*, 317(5838):663–666, Aug 2007.

86. W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, Dec 1983.

87. B. Z. Katz, E. Zamir, A. Bershadsky, Z. Kam, K. M. Yamada, and B. Geiger. Physical state of the extracellular matrix regulates the structure and molecular composition of cell-matrix adhesions. *Mol Biol Cell*, 11(3):1047–1060, Mar 2000.

88. Marcelo G Kazanietz and Patricia S Lorenzo. Phorbol esters as probes for the study of protein kinase c function. *Methods Mol Biol*, 233:423–439, 2003.

89. Andrew Keller, Jimmy Eng, Ning Zhang, Xiao jun Li, and Ruedi Aebersold. A uniform proteomics ms/ms analysis platform utilizing open xml file formats. *Mol Syst Biol*, 1:2005.0017, 2005.

90. Andrew Keller, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search. *Anal Chem*, 74(20):5383–5392, Oct 2002.

91. Nicole King, Christopher T Hittinger, and Sean B Carroll. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science*, 301(5631):361–363, Jul 2003.

92. M. Kinter and N. Sherman. *Protein Sequencing and Identification using Tandem Mass Spectroemtry*. Jon Wiley and Son, New York, 2000.

93. Kenneth Kinzler and Bert Vogelstein. *The genetic basis of human cancer*. New York: McGraw-Hill, 2002.

94. D. K. Klimov and D. Thirumalai. Stretching single-domain proteins: phase diagram and kinetics of force-induced unfolding. *Proc Natl Acad Sci U S A*, 96(11):6166–6170, May 1999.

95. D. K. Klimov and D. Thirumalai. Native topology determines force-induced unfolding pathways in globular proteins. *Proc Natl Acad Sci U S A*, 97(13):7254–7259, Jun 2000.

96. A. Krammer, H. Lu, B. Isralewitz, K. Schulten, and V. Vogel. Forced unfolding of the fibronectin type iii module reveals a tensile molecular recognition switch. *Proc Natl Acad Sci U S A*, 96(4):1351–1356, Feb 1999.

97. Shilpa Kulkarni, Sudipto Das, Colin D Funk, Diana Murray, and Wonhwa Cho. Molecular basis of the specific subcellular localization of the c2-like domain of 5-lipoxygenase. *J Biol Chem*, 277(15):13167–13174, Apr 2002.

98. Tzu-Ling Kuo, Sergi Garcia-Manyes, Jingyuan Li, Itay Barel, Hui Lu, Bruce J Berne, Michael Urbakh, Joseph Klafter, and Julio M Fernndez. Probing static disorder in arrhenius kinetics by single-molecule force spectroscopy. *Proc Natl Acad Sci U S A*, 107(25):11336–11340, Jun 2010.

99. J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982.

100. Morten Kllberg and Hui Lu. Structural feature extraction protocol for classifying reversible membrane binding protein domains. *Conf Proc IEEE Eng Med Biol Soc*, 2009:6735–6738, 2009.

101. Joseph M Laakso, John H Lewis, Henry Shuman, and E. Michael Ostap. Myosin i can act as a molecular force sensor. *Science*, 321(5885):133–136, Jul 2008.

102. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk,

S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

103. Robert Langlois and Hui Lu. Intelligible machine learning with malibu. In *Proceedings of the 30th Annual International Conference of the IEEE, EMBC*, August 20-24 2008.

104. Robert Langlois and Hui Lu. Machine learning for protein structure and function prediction. *Annual Reports in Computational Chemistry*, 4:41–66, 2008.

105. Robert E Langlois, Matthew B Carson, Nitin Bhardwaj, and Hui Lu. Learning to translate sequence and structure to function: identifying dna binding and membrane binding proteins. *Ann Biomed Eng*, 35(6):1043–1052, Jun 2007.

106. Robert E Langlois and Hui Lu. Boosting the prediction and understanding of dna-binding domains from sequence. *Nucleic Acids Res*, 38:3149–3158, Feb 2010.

107. D. Leckband. The surface apparatus–a tool for probing molecular protein interactions. *Nature*, 376(6541):617–618, Aug 1995.

108. Chang S Lee, Il S Kim, Jong B Park, Mi N Lee, Hye Y Lee, Pann-Ghill Suh, and Sung H Ryu. The phox homology domain of phospholipase d activates dynamin gtpase activity and accelerates egfr endocytosis. *Nat Cell Biol*, 8(5):477–484, May 2006.

109. M. A. Lemmon and K. M. Ferguson. Signal-dependent membrane targeting by pleckstrin homology (ph) domains. *Biochem J*, 350 Pt 1:1–18, Aug 2000.

110. Mark A Lemmon. Membrane recognition by phospholipid-binding domains. *Nat Rev Mol Cell Biol*, 9(2):99–111, Feb 2008.

111. P. F. Lenne, A. J. Raae, S. M. Altmann, M. Saraste, and J. K. Hrber. States and transitions during forced unfolding of a single spectrin repeat. *FEBS Lett*, 476(3):124–128, Jul 2000.

112. H. Li, M. Carrion-Vazquez, A. F. Oberhauser, P. E. Marszalek, and J. M. Fernandez. Point mutations alter the mechanical stability of immunoglobulin modules. *Nat Struct Biol*, 7(12):1117–1120, Dec 2000.

113. Hongbin Li, Wolfgang A Linke, Andres F Oberhauser, Mariano Carrion-Vazquez, Jason G Kerkvliet, Hui Lu, Piotr E Marszalek, and Julio M Fernandez. Reverse engineering of the giant muscle protein titin. *Nature*, 418(6901):998–1002, Aug 2002.

114. Pai-Chi Li, Lei Huang, and Dmitrii E Makarov. Mechanical unfolding of segment-swapped protein g dimer: results from replica exchange molecular dynamics simulations. *J Phys Chem B*, 110(29):14469–14474, Jul 2006.

115. Mary S Lipton, Ljiljana Pasa-Tolic', Gordon A Anderson, David J Anderson, Deanna L Auberry, John R Battista, Michael J Daly, Jim Fredrickson, Kim K Hixson, Heather Kostandarithes, Christophe Masselon, Lye Meng Markillie, Ronald J Moore, Margaret F Romine, Yufeng Shen, Eric Stritmatter, Nikola Tolic', Harold R Udseth, Amudhan Venkateswaran, Kwong-Kwok Wong, Rui Zhao, and Richard D Smith. Global analysis of the deinococcus radiodurans proteome by using accurate mass tags. *Proc Natl Acad Sci U S A*, 99(17):11049–11054, Aug 2002.

116. Wei Liu, Markus Eilers, Ashish B Patel, and Steven O Smith. Helix packing moments reveal diversity and conservation in membrane protein structure. *J Mol Biol*, 337(3):713–729, Mar 2004.

117. A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proc Natl Acad Sci U S A*, 96(10):5482–5485, May 1999.

118. H. Lu, B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J*, 75(2):662–671, Aug 1998.

119. H. Lu, A. Krammer, B. Isralewitz, V. Vogel, and K. Schulten. Computer modeling of force-induced titin domain unfolding. *Adv Exp Med Biol*, 481:143–60; discussion 161–2, 2000.

120. H. Lu and K. Schulten. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins*, 35(4):453–463, Jun 1999.

121. H. Lu and K. Schulten. The key event in force-induced unfolding of titin's immunoglobulin domains. *Biophys J*, 79(1):51–65, Jul 2000.

122. David MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

123. Margineantu and Dietterich. Pruning adaptive boosting. In *International Conference on Machine Learning*, volume 14, pages 211–218. Cavtat-Dubrovnik, Croatia,, Morgan Kaufmann, 1997.

124. P. E. Marszalek, H. Lu, H. Li, M. Carrion-Vazquez, A. F. Oberhauser, K. Schulten, and J. M. Fernandez. Mechanical unfolding intermediates in titin modules. *Nature*, 402(6757):100–103, Nov 1999.

125. M. A. Mart-Renom, A. C. Stuart, A. Fiser, R. Snchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325, 2000.

126. Rowena McBeath, Dana M Pirone, Celeste M Nelson, Kiran Bhadriraju, and Christopher S Chen. Cell shape, cytoskeletal tension, and rhoa regulate stem cell lineage commitment. *Dev Cell*, 6(4):483–495, Apr 2004.

127. M. A. McCloskey and M. M. Poo. Contact-induced redistribution of specific membrane components: local accumulation and development of adhesion. *J Cell Biol*, 102(6):2185–2196, Jun 1986.

128. M. A. McCloskey and M. M. Poo. Rates of membrane-associated reactions: reduction of dimensionality revisited. *J Cell Biol*, 102(1):88–96, Jan 1986.

129. Stuart McLaughlin and Diana Murray. Plasma membrane phosphoinositide organization by protein electrostatics. *Nature*, 438(7068):605–611, Dec 2005.

130. Kris Meerschaert, Moe Phyu Tun, Eline Remue, Ariane De Ganck, Ciska Boucherie, Berlinda Vanloo, Gisle Degeest, Jol Vandekerckhove, Pascale Zimmermann, Nitin Bhardwaj, Hui Lu, Wonhwa Cho, and Jan Gettemans. The pdz2 domain of zonula occludens-1 and -2 is a phosphoinositide binding domain. *Cell Mol Life Sci*, 66(24):3951–3966, Dec 2009.

131. Roger E Moore, Mary K Young, and Terry D Lee. Qscore: an algorithm for evaluating sequest database search results. *J Am Soc Mass Spectrom*, 13(4):378–386, Apr 2002.

132. Anna Mulgrew-Nesbitt, Karthikeyan Diraviyam, Jiyao Wang, Shaneen Singh, Paul Murray, Zhaohui Li, Laura Rogers, Nebojsa Mirkovic, and Diana Murray. The role of electrostatics in protein-membrane interactions. *Biochim Biophys Acta*, 1761(8):812–826, Aug 2006.

133. V. Muoz, E. R. Henry, J. Hofrichter, and W. A. Eaton. A statistical mechanical model for beta-hairpin kinetics. *Proc Natl Acad Sci U S A*, 95(11):5872–5879, May 1998.

134. E. A. Nalefski and J. J. Falke. The c2 domain calcium-binding motif: structural and functional diversity. *Protein Sci*, 5(12):2375–2390, Dec 1996.

135. Alexey I Nesvizhskii and Ruedi Aebersold. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem ms. *Drug Discov Today*, 9(4):173–181, Feb 2004.

136. Alexey I Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, 75(17):4646–4658, Sep 2003.

137. Sean P Ng, Ross W S Rounsevell, Annette Steward, Christian D Geierhaas, Philip M Williams, Emanuele Paci, and Jane Clarke. Mechanical unfolding of tnfn3: the unfolding pathway of a fniii domain probed by protein engineering, afm and md simulation. *J Mol Biol*, 350(4):776–789, Jul 2005.

138. Frank No and Stefan Fischer. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol*, 18(2):154–162, Apr 2008.

139. Frank No, Christof Schtte, Eric Vanden-Eijnden, Lothar Reich, and Thomas R Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci U S A*, 106(45):19011–19016, Nov 2009.

140. A. F. Oberhauser, P. E. Marszalek, H. P. Erickson, and J. M. Fernandez. The molecular elasticity of the extracellular matrix protein tenascin. *Nature*, 393(6681):181–185, May 1998.

141. Andres F Oberhauser, Carmelu Badilla-Fernandez, Mariano Carrion-Vazquez, and Julio M Fernandez. The mechanical hierarchies of fibronectin observed with single-molecule afm. *J Mol Biol*, 319(2):433–447, May 2002.

142. E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J Mol Biol*, 288(3):441–459, May 1999.

143. E. Paci and M. Karplus. Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc Natl Acad Sci U S A*, 97(12):6521–6526, Jun 2000.

144. Lifeng Pan, Hao Wu, Chong Shen, Yawei Shi, Wenying Jin, Jun Xia, and Mingjie Zhang. Clustering and synaptic targeting of pick1 requires direct interaction between the pdz domain and lipid membranes. *EMBO J*, 26(21):4576–4587, Oct 2007.

145. Gilbert Di Paolo and Pietro De Camilli. Phosphoinositides in cell regulation and membrane dynamics. *Nature*, 443(7112):651–657, Oct 2006.

146. Wei Sun Park, Won Do Heo, James H Whalen, Nancy A O'Rourke, Heather M Bryan, Tobias Meyer, and Mary N Teruel. Comprehensive identification of pip3-regulated ph domains from c. elegans to h. sapiens by model prediction and live imaging. *Mol Cell*, 30(3):381–392, May 2008.

147. T. Pawson and J. D. Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, Dec 1997.

148. Tony Pawson. Specificity in signal transduction: from phosphotyrosine-sh2 domain interactions to complex cellular systems. *Cell*, 116(2):191–203, Jan 2004.

149. Tony Pawson and Piers Nash. Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452, Apr 2003.

150. Jian Peng and Jinbo Xu. Boosting protein threading accuracy. *Res Comput Mol Biol*, 5541:31–45, 2009.

151. Jian Peng and Jinbo Xu. Low-homology protein threading. *Bioinformatics*, 26(12):i294–i300, Jun 2010.

152. Jian Peng and Jinbo Xu. A multiple-template approach to protein threading. *Proteins*, 79(6):1930–1939, Jun 2011.

153. Jian Peng and Jinbo Xu. Raptorx: exploiting structure information for protein alignment by statistical inference. *Proteins*, 79 Suppl 10:161–171, 2011.

154. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999.

155. M. F. Perutz and F. S. Mathews. An x-ray study of azide methaemoglobin. *J Mol Biol*, 21(1):199–202, Oct 1966.

156. James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kal, and Klaus Schulten. Scalable molecular dynamics with namd. *J Comput Chem*, 26(16):1781–1802, Dec 2005.

157. Ursula Pieper, Benjamin M Webb, David T Barkan, Dina Schneidman-Duhovny, Avner Schlessinger, Hannes Braberg, Zheng Yang, Elaine C Meng, Eric F Pettersen, Conrad C Huang, Ruchira S Datta, Parthasarathy Sampathkumar, Mallur S Madhusudhan, Kimmen Sjlander, Thomas E Ferrin, Stephen K Burley, and Andrej Sali. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*, 39(Database issue):D465–D474, Jan 2011.

158. A. F. Quest, E. S. Bardes, and R. M. Bell. A phorbol ester binding domain of protein kinase c gamma. high affinity binding to a glutathione-s-transferase/cys2 fusion protein. *J Biol Chem*, 269(4):2953–2960, Jan 1994.

159. J. R. Quinlan. Induction of decission trees. *Machine Learning*, 1:81–106, 1986.

160. K. S. Ravichandran, M. M. Zhou, J. C. Pratt, J. E. Harlan, S. F. Walk, S. W. Fesik, and S. J. Burakoff. Evidence for a requirement for both phospholipid and phosphotyrosine binding via the shc phosphotyrosine-binding domain in vivo. *Mol Cell Biol*, 17(9):5540–5549, Sep 1997.

161. Jane Razumovskaya, Victor Olman, Dong Xu, Edward C Uberbacher, Nathan C VerBerkmoes, Robert L Hettich, and Ying Xu. A computational method for assessing peptide- identification reliability in tandem mass spectrometry analysis with sequest. *Proteomics*, 4(4):961–969, Apr 2004.

162. M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by afm. *Science*, 276(5315):1109–1112, May 1997.

163. Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, Jun 2010.

164. J. Rizo and T. C. Sdhof. C2-domains, structure and function of a universal ca2+-binding domain. *J Biol Chem*, 273(26):15879–15882, Jun 1998.

165. D. R. Robinson, Y. M. Wu, and S. F. Lin. The protein tyrosine kinase family of the human genome. *Oncogene*, 19(49):5548–5557, Nov 2000.

166. M. Rodbell. The role of hormone receptors and gtp-regulatory proteins in membrane transduction. *Nature*, 284(5751):17–22, Mar 1980.

167. Ambrish Roy, Narayanaswamy Srinivasan, and Venkatraman S Gowri. Molecular and structural basis of drift in the functions of closely-related homologous enzyme domains: implications for function annotation based on homology searches and structural genomics. *In Silico Biol*, 9(1-2):S41–S55, 2009.

168. Rovshan G Sadygov, Daniel Cociorva, and John R Yates. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods*, 1(3):195–202, Dec 2004.

169. Rovshan G Sadygov and John R Yates. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem*, 75(15):3792–3798, Aug 2003.

170. M. F. Sanner, A. J. Olson, and J. C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, Mar 1996.

171. Charles Sawyers. Targeted cancer therapy. *Nature*, 432(7015):294–297, Nov 2004.

172. M. Saxena, S. Williams, K. Taskn, and T. Mustelin. Crosstalk between camp-dependent kinase and map kinase through a protein tyrosine phosphatase. *Nat Cell Biol*, 1(5):305–311, Sep 1999.

173. Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

174. J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. Smart, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A*, 95(11):5857–5864, May 1998.

175. X. Shao, I. Fernandez, T. C. Sdhof, and J. Rizo. Solution structures of the ca2+-free and ca2+-bound c2a domain of synaptotagmin i: does ca2+ induce a conformational change? *Biochemistry*, 37(46):16106–16115, Nov 1998.

176. Deepak Sharma, Gang Feng, Dingyue Khor, Georgi Z Genchev, Hui Lu, and Hongbin Li. Stabilization provided by neighboring strands is critical for the mechanical stability of proteins. *Biophys J*, 95(8):3935–3942, Oct 2008.

177. M. Sheng and C. Sala. Pdz domains and the organization of supramolecular complexes. *Annu Rev Neurosci*, 24:1–29, 2001.

178. Paul Sherwood, Bernard R Brooks, and Mark S P Sansom. Multiscale methods for macromolecular simulations. *Curr Opin Struct Biol*, 18(5):630–640, Oct 2008.

179. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1):209–225, Apr 1997.

180. Shaneen M Singh and Diana Murray. Molecular modeling of the membrane targeting of phospholipase c pleckstrin homology domains. *Protein Sci*, 12(9):1934–1953, Sep 2003.

181. Robert G Smock and Lila M Gierasch. Sending signals dynamically. *Science*, 324(5924):198–203, Apr 2009.

182. Zhao Song, Luonan Chen, and Dong Xu. Confidence assessment for protein identification by using peptide-mass fingerprinting data. *Proteomics*, 9(11):3090–3099, Jun 2009.

183. David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. Gromacs: fast, flexible, and free. *J Comput Chem*, 26(16):1701–1718, Dec 2005.

184. Robert V Stahelin, Fei Long, Brian J Peter, Diana Murray, Pietro De Camilli, Harvey T McMahon, and Wonhwa Cho. Contrasting membrane interaction mechanisms of ap180 n-terminal homology (anth) and epsin n-terminal homology (enth) domains. *J Biol Chem*, 278(31):28993–28999, Aug 2003.

185. Hanno Steen and Matthias Mann. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5(9):699–711, Sep 2004.

186. Harald Stenmark, Rein Aasland, and Paul C Driscoll. The phosphatidylinositol 3-phosphate-binding fyve finger. *FEBS Lett*, 513(1):77–84, Feb 2002.

187. Michael A Stiffler, Jiunn R Chen, Viara P Grantcharova, Ying Lei, Daniel Fuchs, John E Allen, Lioudmila A Zaslavskaia, and Gavin MacBeath. Pdz domain binding selectivity is optimized across the mouse proteome. *Science*, 317(5836):364–369, Jul 2007.

188. Shiro Suetsugu, Kiminori Toyooka, and Yosuke Senju. Subcellular membrane curvature mediated by the bar domain superfamily proteins. *Semin Cell Dev Biol*, 21:340–349, Dec 2009.

189. Takuma Sugi, Takuji Oyama, Kosuke Morikawa, and Hisato Jingami. Structural insights into the pip2 recognition by syntenin-1 pdz domain. *Biochem Biophys Res Commun*, 366(2):373–378, Feb 2008.

190. Sonia Snchez-Bautista, Consuelo Marn-Vicente, Juan C Gmez-Fernndez, and Senena Corbaln-Garca. The c2 domain of pkcalpha is a ca2+ -dependent ptdins(4,5)p2 sensing domain: a new insight into an old pathway. *J Mol Biol*, 362(5):901–914, Oct 2006.

191. Johannes Sding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960, Apr 2005.

192. D.L. Tabb, Jimmy K. Eng, and John R. Yates III. *Proteome Research: Mass Spectrometry.*, chapter Protein Identification by SEQUEST. Springer, Berlin., 2001.

193. L. Tskhovrebova, J. Trinick, J. A. Sleep, and R. M. Simmons. Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature*, 387(6630):308–312, May 1997.

194. Peter J Ulintz, Ji Zhu, Zhaohui S Qin, and Philip C Andrews. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol Cell Proteomics*, 5(3):497–509, Mar 2006.

195. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen,

M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

196. Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clustering comparison: Is a correction for chance necessary? In *CML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM.*, page 10731080, 2009.

197. Viola Vogel and Michael Sheetz. Local force and geometry sensing regulate cell functions. *Nat Rev Mol Cell Biol*, 7(4):265–275, Apr 2006.

198. Jinrong Wan, Michael Torres, Ashwin Ganapathy, Jay Thelen, Beverly B DaGue, Brian Mooney, Dong Xu, and Gary Stacey. Proteomic analysis of soybean root hairs after infection by bradyrhizobium japonicum. *Mol Plant Microbe Interact*, 18(5):458–467, May 2005.

199. D. S. Wang, R. Shaw, J. C. Winkelmann, and G. Shaw. Binding of ph domains of beta-adrenergic receptor kinase and beta-spectrin to wd40/beta-transducin repeat containing regions of the beta-subunit of trimeric g-proteins. *Biochem Biophys Res Commun*, 203(1):29–35, Aug 1994.

200. M. P. Washburn, D. Wolters, and J. R. Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*, 19(3):242–247, Mar 2001.

201. S. H. White, W. C. Wimley, A. S. Ladokhin, and K. Hristova. Protein folding in membranes: determining energetics of peptide-bilayer interactions. *Methods Enzymol*, 295:62–87, 1998.

202. Philip M Williams, Susan B Fowler, Robert B Best, Jos Luis Toca-Herrera, Kathryn A Scott, Annette Steward, and Jane Clarke. Hidden complexity in the mechanical properties of titin. *Nature*, 422(6930):446–449, Mar 2003.

203. Matthew J Winters, Rachel E Lamson, Hideki Nakanishi, Aaron M Neiman, and Peter M Pryciak. A membrane binding domain in the ste5 scaffold synergizes with gbetagamma binding to control localization and signaling in pheromone response. *Mol Cell*, 20(1):21–32, Oct 2005.

204. Hao Wu, Wei Feng, Jia Chen, Ling-Nga Chan, Siyi Huang, and Mingjie Zhang. Pdz domains of par-3 as potential phosphoinositide signaling integrators. *Mol Cell*, 28(5):886–898, Dec 2007.

205. Sitao Wu, Jeffrey Skolnick, and Yang Zhang. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol*, 5:17, 2007.

206. Sitao Wu and Yang Zhang. Muster: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, 72(2):547–556, Aug 2008.

207. Jinbo Xu and Ming Li. Assessment of raptor's linear programming approach in cafasp3. *Proteins*, 53 Suppl 6:579–584, 2003.

208. Jinbo Xu, Ming Li, Dongsup Kim, and Ying Xu. Raptor: optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1(1):95–117, Apr 2003.

209. Y. Xu, L. F. Seet, B. Hanson, and W. Hong. The phox homology (px) domain, a new player in phosphoinositide signalling. *Biochem J*, 360(Pt 3):513–530, Dec 2001.

210. ChengFeng Yang and Marcelo G Kazanietz. Divergence and complexities in dag signaling: looking beyond pkc. *Trends Pharmacol Sci*, 24(11):602–608, Nov 2003.

211. L. Yao, Y. Kawakami, and T. Kawakami. The pleckstrin homology domain of bruton tyrosine kinase interacts with protein kinase c. *Proc Natl Acad Sci U S A*, 91(19):9175–9179, Sep 1994.

212. Youngdae Yoon, Park J Lee, Svetlana Kurilova, and Wonhwa Cho. In situ quantitative imaging of cellular lipids using molecular sensors. *Nat Chem*, 3(11):868–874, Nov 2011.

213. Haiyuan Yu, Pascal Braun, Muhammed A Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-Franois Rual, Amlie Dricot, Alexei Vazquez, Ryan R Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E Hudson, Juyong Park, Xiaofeng Xin, Michael E Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P Roth, Albert-Lszl Barabsi, Jan Tavernier, David E Hill, and Marc Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct 2008.

214. Yang Zhang. I-tasser: fully automated protein structure prediction in casp8. *Proteins*, 77 Suppl 9:100–113, 2009.

215. Yang Zhang. Protein structure prediction: when is it useful? *Curr Opin Struct Biol*, 19(2):145–155, Apr 2009.

216. M. M. Zhou, K. S. Ravichandran, E. F. Olejniczak, A. M. Petros, R. P. Meadows, M. Sattler, J. E. Harlan, W. S. Wade, S. J. Burakoff, and S. W. Fesik. Structure and ligand recognition of the phosphotyrosine binding domain of shc. *Nature*, 378(6557):584–592, Dec 1995.

217. Pascale Zimmermann. The prevalence and significance of pdz domain-phosphoinositide interactions. *Biochim Biophys Acta*, 1761(8):947–956, Aug 2006.

218. Pascale Zimmermann, Kris Meerschaert, Gunter Reekmans, Iris Leenaerts, J. Victor Small, Jol Vandekerckhove, Guido David, and Jan Gettemans. Pip(2)-pdz domain binding controls the association of syntenin with the plasma membrane. *Mol Cell*, 9(6):1215–1225, Jun 2002.