A STUDY OF INTRINSIC DISORDER AND IT'S ROLE IN FUNCTIONAL PROTEOMICS

Amrita Mohan

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy in Informatics,
Indiana University

November 2009

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Thesis Committee

_____

Predrag Radivojac, PhD., Assistant Professor, Chair

_____

Mehmet Dalkilic, PhD., Associate Professor

_____

Sun Kim, PhD., Associate Professor

_____

Yuzhen Ye, PhD., Assistant Professor

October 23, 2009

To my father; for whom this is a cherished dream
To my mother; for her undying love and support

ACKNOWLEDGEMENTS

It is a pleasure to thank all those who made this dissertation possible.

It is difficult to overstate my gratitude to my Ph.D. supervisor, Professor Predrag Radivojac. As the biggest critic of my work, he has helped me set high standards for myself. It is hard to elucidate all the lessons that I have learnt from him in the past five years. This dissertation embodies only a fraction of all the invaluable knowledge that he has imparted to me.

I wish to thank my closest companion since freshman year and now my husband, for helping me get by all the difficult times, for the emotional support, camaraderie, entertainment and care he provided.

I am grateful to the Indiana University School of Informatics and Computing and, the Eli Lilly Foundation for providing me with the financial means to support my education. The school's entire administrative staff also deserves a note of thanks, for helping me cross each milestone. Linda Hostetter deserves a very special mention.

Last but by no means the least, I am indebted to all my student colleagues, both past and present for providing a fun environment in which to learn and grow. I am particularly thankful to Vikas and Anoop for sincerely proofreading this manuscript to help make sense of non-sense.

Amrita Mohan

A STUDY OF INTRINSIC DISORDER AND ITS ROLE IN FUNCTIONAL PROTEOMICS

The last decade has witnessed the emergence of an alternate view on how protein function arises. This view attributes the functionality of many proteins to the presence of an ensemble of flexible regions popularly as ‘intrinsically disordered’ or ‘unstructured’. Several proteomic studies have corroborated the existence of either wholly disordered proteins or proteins that contain regions of disorder in them. The purpose of this dissertation was to investigate the consistency of such regions across experiments, their mechanism of facilitating function via disorder-to-order transitions, their presence and significance in pathogenic versus non-pathogenic organisms and their promise of applicability towards the computational prediction of peptides involved in the most common class of post-translational modifications, phosphorylation. Besides these, a new algorithm exploiting the strong correlation between phosphorylation and intrinsic disorder has also been proposed to improve the detection of phosphorylated peptides via high-throughput methods such as tandem mass-spectrometry (LC-MS/MS). Results presented in this study, guide us in understanding the robustness of unstructured regions in proteins to sequence changes and environment, their role in facilitating molecular recognition as well as improving currently available methods for identification of post-translationally modified peptides. The findings and conclusions of this dissertation have the potential to impact ongoing structural genomics initiatives by suggesting alternative

methods for determining structure for targets containing regions of disorder. Additional ramifications of results from this work include directing attention towards the possible use of regions of intrinsic disorder by pathogenic organisms for host cell invasion. We believe that unlike the traditional reductionist approach in a scientific method, this study gathers strength and utility by investigating the role of intrinsic disorder on more than one front in order to provide a novel perspective to the understanding of complex interactions within biological systems. Concluding arguments presented in this study pique one's curiosity regarding the evolution of disordered regions and proteins in general. On a technological side, the findings from this study unequivocally support the viable use of informatics methods in gaining new insights about a relatively young class of proteins known as intrinsically disordered proteins and its applicability to improve our present knowledge of cellular physiology.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER ONE: INTRODUCTION

What is intrinsic disorder (ID)?

Traditional wisdom pertinent to the protein structure-function paradigm relies strongly on a single assumption that the three-dimensional structure of a protein is the key determinant of its function. However, results from many recent studies present evidence that an '*unstructured*' or '*intrinsically disordered*' conformation may in fact be responsible for the true functional state of some proteins.[1-22] Some other popular terms used to describe proteins or their regions without a specific 3-D structure include: '*flexible*', '*mobile*', '*partially folded*', '*natively unfolded*', '*pre-molten globule*'. This special class of flexible structure has been reported to be either partially or completely spanning the lengths of proteins. Despite the well adopted view that protein function is an immediate consequence of its three-dimensional structure, researchers discovered examples where fragments of proteins were found to be unstructured and contributing to the overall protein's functionality. The compelling evidence that, intrinsically disordered proteins exist *in vitro* as well as *in vivo* justifies treating them as a separate class within the protein universe.[2, 4, 10, 23, 24]

Function and significance of intrinsic disorder

Many studies have confirmed the presence of disordered proteins in a number of proteomes and that their increase in abundance is directly proportional to the increasing organism complexity.[25-27] [28-30] This increased prediction of disorder in eukaryotes in

comparison to prokaryotes or the *archaea* has been presupposed to be a result of the increased need for cell signaling and regulation.[27, 28, 30, 31] Two recent studies on the *Plasmodium falciparum* protozoan genome demonstrated the presence of asparagine-rich low complexity regions in as many as 50% of its translated genes.[32, 33] It is believed that these low complexity regions indicate the presence of disorder and contribute to the overall uniqueness of each individual species within this diverse group of early-branching eukaryotes. In the case of *Plasmodiu*m species, which cause malaria, such regions have also been reported to hinder the identification of homologues, thus making functional genomics exceptionally challenging.[34]

It is important to state here that even though the physical conformation of natively disordered proteins closely mimics the observed denatured states of structured (also known as ordered) proteins, the two are physiologically different. Unlike its denatured counterpart, proteins with disordered conformation do not lose their ability to function biologically.[4, 35] Since the sighting of such observations, the field of intrinsic disorder and protein functionality resulting from intrinsic disorder has steadily garnered attention from around the globe. Many literary articles[36-40] now contain confirmed reports of disordered regions that are crucial for protein function. The functional importance of protein disorder is further emphasized by its role in housekeeping cellular processes such as signal transduction, cell-cycle regulation, gene expression and molecular recognition as reported in.[9, 10, 13, 28, 41] The widespread prevalence and importance of these intrinsically disordered proteins has called for reassessing the understanding of the classical protein structure–function paradigm.[42] Among other functions, intrinsic disorder has been suggested to play an important role in molecular recognition as well as post-translational

modifications or PTMs. [5, 8, 41, 43-45] In addition to these, the functional importance of protein disorder in gene expression and cell cycle regulation has also been established.[1, 46] The following paragraphs provide a brief overview of molecular recognition and post-translational modifications.

Molecular recognition

Molecular recognition is defined as a process by which biological entities specifically interact with each other or with small molecules to form complexes. Complex formation is often a prerequisite for biological function, but also serves as a mechanism of functional modulation and signal transduction. Disorder-mediated molecular recognition although highly specific, occurs with low affinity and typically involves regions that are capable of binding to multiple partners by adopting a spectrum of conformations. Since these characteristics are also shared by signaling and regulation interactions, intrinsic disorder has often been implicated in cell signaling and regulation.

Post-translational modifications

It is well known that the biological activity for many a proteins is regulated by different types of post-translational modifications (or PTMs). PTMs are chemically modified derivatives of translated proteins. Typically these modifications are the result of covalent additions of various chemical groups such as methyl, phosphoryl, glycosyl and acetyl to the side chains of particular amino acids such as serine, threonine, methionine etc. Occasionally, PTMs can also be the result of proteolytic cleavage of a peptide bond

by a special class of enzymes known as hydrolases.[47] There are more than 200 reported post-translation modification types known to occur in eukaryotes depending on the chemical group that attaches to a protein. These include (but are not limited to) phosphorylation, ubiquitination, acetylation, methylation, acylation, glycosylation, sulfation and deamidation. Sometimes, PTMs are also classified based on the small-molecule signaled for attaching at the site of modification (e.g., ubiquitination, SUMOylation) and based on the loss or gain of a chemical group on amino acids (deamidation, oxidation). Some proteins undergo a number of post-translational modifications to achieve their expected function. A good example for such types of proteins is provided by histones. Histones reportedly undergo methylation, phosphorylation, acetylation, ubiquitination, ADP-ribosylation, and SUMOylation at different time-intervals, to modulate histone–DNA interactions as well as histone–histone interations, that are closely involved in the control of nucleosome stability.[48] Chemically modified residues of a polypeptide chain after translation offer an extended range of functionalities to the protein.

Post-translationally modified instances of a protein play a crucial role in determining its active state, cellular localization, degradation, as well as interactions with other proteins. In signaling, for example, multiple kinase molecules are switched on and off by the reversible addition and removal of phosphate groups[49], and in the cell cycle, ubiquitination marks cyclins for destruction at designated time intervals.[50] In phosphorylation, molecules known as kinases attach phosphate groups to select amino acids such as serine (S), threonine (T) and tyrosine (Y). Ubiquitination on the other hand refers to the modification of a protein structure by the covalent attachment of one or more

ubiquitin monomers to a lysine side chain. Ubiquitination of proteins involves the combined action of three molecules viz., E1 or the ubiquitin-activating enzyme, E2 or the ubiquitin-conjugating enzymes and E3, the ubiquitin-protein ligase. It has been suggested that E3s work as 'docking proteins' that specifically bind to substrate proteins and specific E2s. At this stage, ubiquitin is transferred directly from E2s to substrates.[51] A relatively small molecule, ubiquitin comprises only of 76 residues and functions to regulate protein turnover in a cell by closely regulating the degradation of other target proteins. Protein degradation is a crucial step in most biological processes as it facilitates the elimination of non-functional or abnormal proteins. Past research suggests links between ubiquitination and phosphorylation with diseases involving a variety of cellular activities including neural and muscular degeneration, DNA transcription and repair, apoptosis or programmed cell death and cell division.[52] With a key role such as that of regulating a cell's housekeeping events, it is unsurprising to see unabated efforts by scientific groups to develop bigger and better information repositories of protein phosphorylation and ubiquitination sites.

In the past, Edman degradation was frequently employed in the identification of PTMs. Currently faster, more accurate methods such as the mass spectrometry (MS) based ones have gained popularity within the proteomics community. Notwithstanding the availability of advanced methods of detecting post-translationally modified peptides, their identification continues to entail major challenges due to the observation that at a given time only a fraction of the molecules of a given protein in the cell might actually be modified. Moreover, these modifications are often spread across multiple positions on a molecule resulting in a formidable heterogeneous population of the given protein at a

given point of time. As a consequence of this, a highly sensitive and selective analytical methodology is required for the thorough analysis of all the post-translationally modified copies of the proteins present in a cell.

Mode of facilitating protein function by intrinsically disordered proteins

Among other roles, disorder is believed to play a crucial role in the molecular recognition of protein molecules.[5, 8, 41, 43-45] Molecular recognition is defined as a process by which proteins and other biological entities specifically interact with each other or with small molecules to form complexes. Complex assembly is often a necessary step for biological activity, and serves as a mechanism of functional regulation and signal transduction. Common characteristics of intrinsic disorder-mediated molecular recognition include: (a) a combination of high specificity and low affinity; (b) binding diversity in which one region specifically recognizes different partners by structural rearrangement; (c) binding commonality in which multiple, distinct sequences recognize a common binding site, such that these sequences may or may not assume dissimilar folds. These same features are also believed to be important for interaction-mediated signaling and modulation, which indicates that unfolded proteins may play a central role in signal transduction.[1, 5, 8, 41, 43] Typically, functions of disordered proteins may arise directly due to its disordered state, switching between multiple disordered states, or from transitions between disordered and ordered states.

Characterization of Intrinsic Disorder

In the past decade, there has been significant progress in our understanding of the wide-spread prevalence and function of intrinsically disordered proteins.[1, 2, 6, 10, 23, 41, 53-55] What once seemed to be a set of exceptions to the traditional structure-to-function paradigm, where every protein was believed to have stable 3D structure to carry out function, turned into a field where computational and experimental approaches were developed and combined to accurately characterize disordered proteins [9, 56], understand their function [1, 6, 55, 57-59], mechanisms of binding [10, 45, 60, 61] and estimate their abundance in the protein universe.[10, 28, 31, 62, 63] Undoubtedly, bioinformatics analyses and methods, especially a set of predictors and statistical techniques, played a significant role in this process which estimated the broad functional repertoire of disordered proteins and provided early evidence of their prevalence in all kingdoms of life.[55, 64]

Experimental determination of intrinsic disorder

Several methods have been used to characterize disorder in proteins, each with its own strengths and limitations. Here we briefly discuss three of the leading methods for the experimental characterization of intrinsic disorder; X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and Circular Dichroism (CD).

X-ray crystallography

Regions with high flexibility found within proteins vary in location from one configuration to another and as a result fail to scatter X-rays as coherently. The absence

of X-ray scattering for such regions results in missing electron density. Therefore, presence of intrinsic disorder in a protein can be confirmed by the observation of missing electron density in protein structures determined using X-ray crystallography. This also explains why completely disordered proteins are difficult to crystallize. Crystallographers have often identified two configurations for disorder: static and dynamic.[65, 66] A dynamically disordered region has the potential to freeze into a single preferred conformation at lower temperatures, while static disorder can remain persistent regardless of temperature.[67] It is important to mention at this point that a missing region with one set of Ramachandran φ, ψ angles as determined by X-ray diffraction method, can also be a "wobbly domain" and not natively disordered owing to a flexible hinge in the protein's backbone allowing the region to adopt multiple positions in the crystal lattice as a rigid body. Missing electron density in protein structures can also arise from failure to solve the phase problem, from crystal defects or even from unintentional proteolytic removal during protein purification. Therefore, information from X-ray diffraction experiments may not always be a confident indicator of intrinsic disorder and missing electron density regions obtained using X-ray crystallization techniques should be treated with caution.

Nuclear magnetic resonance spectroscopy

3D structures can be determined for proteins in solution by the method of nuclear magnetic resonance (NMR) spectroscopy. In 1978, the same year that functional disorder was indicated by X-ray crystallography, Aviles et al used NMR and found the highly charged, functional tail of histone H5 to be disordered.[68] Since then, NMR 3D structural

determination has led to the characterization of several proteins containing functional, yet unstructured regions.[69, 70] Another surprising result of NMR protein structure determinations was the discovery of functional proteins that lacked any type of structural predisposition, i.e., functional proteins that are disordered from end-to-end.[71] Because NMR is more certain in its characterization of disorder than X-ray diffraction, the rediscovery of native disorder by NMR had significant impact.[42] Intrinsically disordered proteins have a multitude of dynamic conformations that interchange on a number of timescales. NMR spectroscopy can detect such a scale of molecular motion with reasonably high accuracy. Moreover, the absence of the requirement for crystallization in NMR rids it of any possible biased estimates arising due to the commonness of disorder ascertained using X-ray crystallography. A typical NMR experiment provides motional information for a protein on a residue-by-residue basis[72] by means of a variety of different isotopic labeling and pulse sequence experiments.[73] Of particular use is the ^{15}N-^{1}H hetereonuclear nuclear overhauser effect (NOE) measurement, which gives positive values for more slowly tumbling or ordered residues and negative values for more rapidly tumbling disordered residues.[73, 74]When these NOE data are looked up in reference to an amino-acid sequence, ordered regions can easily be identified by a series of consecutive positive values and disordered regions by contiguous stretches of negative values.[69]

Unlike ordered proteins, very few regions with persistent secondary structure and yet an apparent lack of any specific tertiary structure (also known as “molten globules”) have been successfully characterized by NMR, indicating significant experimental difficulties encountered. A big reason for this difficulty stems from the fact that molten globular regions frequently aggregate at the concentrations typically required for NMR

experiments. In addition to this, molten globules are inherently heterogeneous by nature and individual conformations typically interconvert on the millisecond timescale; this leads to extreme broadening of the side-chain NMR peaks. In short, native molten globular protein domains are significantly underrepresented in any collection of disordered proteins characterized using NMR.

Circular dichroism (CD) spectroscopy

Structural information for proteins in solution can also be determined using circular dichroism.[75] Far-UV CD spectra provide estimates of secondary structure and can help distinguish ordered and molten globular forms from random coil. On the other hand, near-UV CD show sharp peaks for aromatic groups when the protein is ordered, but these peaks disappear for molten globules and random coils due to motional averaging.[76-78] Therefore, the combined use of near and far-UV CD can help in determining whether a protein is ordered, molten globular, or random coil. However, this method provides only semi-quantitative data and lacks residue-specific information. In other words, circular dichroism spectroscopy does not provide clear information about proteins that contain both ordered and disordered regions.

Some other popular methods of determining disorder in protein molecules include protease digestion and Stoke's radius determination. Protease digestion relies on the assumption that a structured region needs to become unfolded over a length of 10 or more residues in order to be cut by standard proteases.[79] Studies by Fontana et al., [80, 81] demonstrate huge increases in digestion rates after the F-helix of myoglobin is converted

to a disordered state in apomyoglobin, with the cut loci for several different proteases occurring within the disordered region that arises from the F-helix. Although the exact disorder to order digestion rate ratio was not estimated, the authors indicated that typical rates are potentially in the 10^5 to 10^7 range. Thus, hypersensitivity to proteases is sure evidence of protein disorder. Protease digestion gives position-specific information. However, the requirement for protease-sensitive residues limits the demarcation of order/disorder boundaries by this method. Protease digestion is particularly helpful when used in combination with other methods such as X-ray diffraction to help sort out whether a region of missing electron density is due to a wobbly domain or due to intrinsic disorder.[82, 83] Protease digestion has also been used in conjunction with circular dichroism. Finally, the combination of proteolysis and mass spectrometry[10] for fragment identification shows special promise for indicating the presence of intrinsic disorder. In Stoke's radius determination method, the observation of significantly outsized radii for a given molecular weight serves as an indicator of disorder. This method has often been used in conjunction with CD spectroscopy,[84, 85] to test the presence of random coil structure in proteins. Besides these, the use of optical rotatory dispersion (ORD)[86], Fourier transform infrared[13], Raman spectroscopy/Raman optical activity[87] and fluorescence techniques[88] has also been explored to determine intrinsic disorder.

Regardless of the wide variety of methods available for the experimental determination of disorder, the outcome from each method greatly relies on experimental parameters (such as temperature, pH and salt concentration) supplied at the time of assay. Unfortunately, limited data is available to systematically evaluate the consequences of changes in experimental parameters on regions of disorder in polypeptides.

Computational prediction of intrinsic disorder

With the number of disordered proteins with verifiable functions continuing to increase dramatically, it is unsurprising to observe a proportional hike in the number of predictors of disorder including Predictors of Natural Disordered Region (PONDR)[89-92] (70 – 82%), GlobPlot[93] (62%), DISOPRED predictor[94] (74%), NORSp[95] (78%), DisEMBL[96] (62%), IUPred[97] (76%), RONN[98] (85%), FoldIndex[99] (77%), DISpro[100] (93%) , PreLINK (93%)[101] and Wiggle[102] (66%)). The inclusion of disorder prediction as an independent category in the past four CASP experiments (CASP5[103], CASP6[104],CASP7 and CASP8) has further generated increased interest in disorder prediction by structural biologists across the globe. CASP (Critical Assessment of techniques for protein Structure Prediction) a community-wide competition for protein structure prediction serves as a quality platform for comparing different intrinsic disorder and protein structure prediction software. Each CASP competition is strictly conducted every two years to measure the performance of existing software tools and servers capable of modeling protein structure including intrinsic disorder. Ensuring accurate and efficient prediction of disorder not only directly contributes towards the overall goal of structural genomics projects but also can also indirectly help identify potential drug targets with critical functions.

Why has little been known about intrinsic disorder until now?

Traditional structure determination methods are designed such that they favor the production and characterization of well-folded, functional proteins. A typical biochemical

procedure initiates with plant, animal or bacterial cells being isolated and homogenized for use in assaying the function of interest. The homogenized sample is subjected to fractionation, chromatography and/or gel filtration. Finally, all available fractions are tested for the function of interest and the associated active protein is purified for sequence and structure determination.

As can be realized by the order of aforementioned steps, such methodology is inherently driven to select only polypeptides with well-defined tertiary structures. Unfolded proteins under such conditions are much more prone to degradation by proteases under the conditions typically prevailing during such assays. In addition to this, typically disordered regions occupy a relatively smaller part of the larger proteins being studied and are therefore much more difficult for ascertainment.

With the recent increase in the wealth of data supporting the presence and significance of disordered proteins there has been growing parallel interest to simultaneously improve the *in vivo* characterization of disorder. This growing attention has led to the invention and development of alternative biochemical approaches conducive to the experimental determination of disorder. Such methods can begin with the formulation of a function of interest (e.g. gain of signaling event) followed by mapping the function to a key gene using gene-function mapping tools. The gene can next be transcribed into a protein and purified (with or without binding partner) for structure determination using NMR or circular dichroism spectroscopy.

CHAPTER TWO: STATEMENT OF PROBLEMS

The goal of this dissertation is to seek answers to some key questions regarding specific aspects of a relatively new class of protein conformation, intrinsic disorder and study its role in molecular recognition and post-translational modifications such as phosphorylation and explore its prevalence in a large group of pathogenic organisms in comparison to non-pathogenic organisms. In doing so, we pose the following questions:

Are disordered regions perfectly repeatable experimentally?

Despite the fact that since 1997, the number of predictors of disordered protein regions has continued to rise, with balanced accuracies within the 70 – 93% range, the individual disorder prediction accuracies of most sequence-based predictors of disorder at CASP7 were much lower.[105] In addition to this, statistical evaluation of disorder predictions by various groups on CASP targets in CASP7[106] revealed that despite similar distribution of the disordered segments between CASP7 and CASP6[104] targets, no significant improvement in disorder predictions was made in CASP7 in comparison to CASP6. Assessors' evaluations for the disorder prediction category in CASP8 (Brik et al., in press) concluded that the latest CASP experiment witnessed a marginal boost in the performance of disorder predictions in comparison to CASP7 however none of the submitted predictors achieved >85% accuracy. (Figure 1)

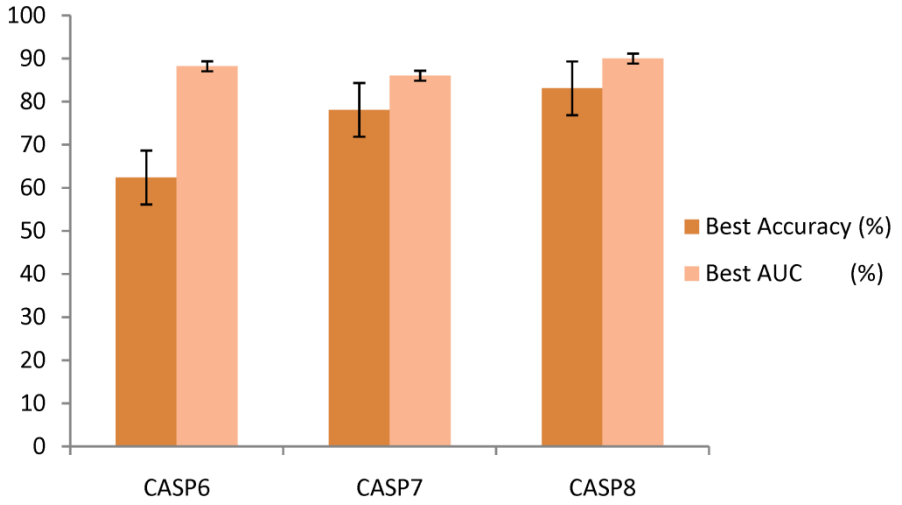

Figure 1: Best blind-test accuracies in CASP 6, 7 and 8 in disorder prediction category.

Given this finding and the continued interest in predicting disorder from structural disorder researchers, at least two key questions arise about the assumptions behind their construction and whether further improvement in classification models is still possible. Are experimentally determined disordered regions preserved if the same or similar protein is characterized under the same or slightly different experimental conditions? Is there an upper limit to the accuracy of disorder prediction stemming from this experimental repeatability?

We suspect that this apparent lack of improvement in prediction accuracies stems from the fact that all sequence based predictors of disorder stand to be limited by the type of data used to train them. Alternatively, learning to predict disorder accurately may imply perfect in vivo repeatability of disorder. Although it is known that the increasing variety in the lengths and compositions of disordered regions makes it harder for sequence-specific predictors to anticipate such regions with very high accuracy[91], we believe that the reason behind this stagnation in prediction accuracies is as simple as the

existence of a practical limit to the experimental repeatability of disordered regions. Any upper limit to the repeatability of experimental disorder will also serve to restrict the highest achievable accuracy for intrinsic disorder prediction. In other words, the achievement of higher prediction accuracies for intrinsic disorder may, in fact, be limited due to the absence of perfect identity between disordered regions from highly similar sequences characterized in different experiments using differing experimental conditions.

What distinguishes disorder-mediated interactions from other interactions?

Previously, intrinsically disordered proteins have been shown to be prevalent across a multitude of eukaryotic proteomes[25, 27] and are believed to play a central role in the process of molecular recognition by small molecules[35], and especially so in the case of interaction-mediated signaling events. There are abundant advantages of disordered proteins or regions catering to this role. These include the decoupling of specificity and affinity[44], the ability to recognize multiple partners through the accommodation of different conformations[107] and faster coupling, perhaps, through the fly-casting[108] mechanism. A majority of disordered proteins perform their functions by undergoing disorder-to-order transitions upon binding to their target proteins.[44, 45] A recent article cites the use of this disorder-to-order phenomenon to devise a methodology for obtaining the structure of a disordered protein that had previously failed in the high-throughput structure determination pipeline of structural genomics.[109] The approach of crystallizing unfolded proteins in the presence of their molecular binding partners promises to greatly increase the number of proteins amenable to the traditional structure determination

methods. A previous work[44] focused on a small set of fragments that adopted an α-helical structure after binding to their respective partners. However, the class of all such fragments is thought to be much broader and believed to include fragments of β-strand or irregular structure type. To improve our current understanding of how disorder-mediated interactions are facilitated a much larger data set of disorder fragments that gain structure when bound to their partners is needed. Clearly, the development of a complete data set of disordered proteins known to adopt well-formed three-dimensional conformations upon binding to their targets is the first step in this direction. The availability of such a data set, besides facilitating the development of a model for intrinsic disorder-mediated interactions will help us in answering other important questions. A few such queries include, what features distinguish disorder-based molecular recognition from other types of protein interactions? Are interaction surfaces on such fragments different from other interactions? Is there preference for a specific local secondary structure type in such fragments? Do inter-chain or intra-chain interactions have effect on the conformation of such fragments?

Can intrinsic disorder be applied to improve the detection of post-translational modifications?

An interesting study by Xie et al.,[58] demonstrated that post-translational modifications can be segregated into two classes based on the conformational state of the modified site: the first class includes all PTMs where the modified sites are present within regions of structure and the second class includes all PTMs where the site of

modification is associated with intrinsic disorder. It is suggested that the inherent flexibility of disordered regions in proteins facilitates the easy binding of some of the important small-molecules such as the likes of kinases, phosphatases and ubiquitin ligases to their respective targeted residues. The former class encompasses PTMs that are vital for the execution of catalytic reactions, enzymatic activities and for stabilizing protein structure. Examples of such PTMs include oxidation, formylation and covalent attachment of polypeptides with organic heteroatoms. The latter group of modifications is closely involved with interactions relying on low affinity and high specificity binding (e.g. signaling events) attaching groups and their substrates. Examples of PTMs from this group include phosphorylation, ubiquitination, methylation, acetylation, prenylation, acylation, adenylylation and SUMOylation among others. Of these, phosphorylation and ubiquitination (also known as ubiquitylation) collectively account for more than half of the modifications crucial to signaling and transduction processes in biological systems. The dynamic nature of both these modifications makes them extremely important in cellular regulation mechanisms warranting the need for reliable, fast and sensitive methods for their identification. A recent study by Radivojac et al., [110] presented multiple lines of evidence indicating that a significant fraction of ubiquitination sites may be located in intrinsically disordered regions. The authors describe a novel random forest based predictor of ubiquitination sites, UbPred, designed to identify candidate ubiquitination sites using local sequence information, including the propensity of a residue to be disordered as predicted by VSL2B. UbPred achieved balanced accuracy and area under receiver operator curve of 72% and 80%, respectively. Besides this, many other studies have also presented arguments supporting the pivotal role intrinsic disorder

plays in ubiquitination.[111-113] Previous research has also suggested that disordered regions are preferentially enriched in phosphorylation sites.[114] This theory has repeatedly been validated by a number of recent studies including one that demonstrated that a majority of phosphorylation sites found in the mouse forebrain proteins are present within unstructured regions[115] and disordered proteins in the *Sacchromyces cervesiae* proteome are putative substrates of a large number of kinases.[112] It is therefore unsurprising to witness consistently rising interest in developing methodologies that strictly focus on the detection of post-translationally modified peptides such as phospeptides and those containing sites of ubiquitination as well as assembling comprehensive, independent data sets of phosphorylation and ubiquitination sites.

Does structural disorder content in early-branching pathogenic organisms differ from non-pathogenic organisms?

The study of early-branching eukaryotic cells carries great potential to provide valuable insights into the evolutionary landscape of cell developmental biology. Some of the oldest eukaryotic species are single-celled protozoa, a diverse array of organisms that live freely or have evolved into parasitic entities. Investigation of the parasitic varieties not only offers the benefit of studying organisms with limited knowledge about their phylogenetic neighbors, but also may have rewards of therapeutic relevance. Among the myriad of parasitic protozoa, are notorious pathogens that cause significant morbidity and mortality in humans and livestock. Consequently, parasitic protozoal infections also have profound economic and socioeconomic ramifications. Additionally, many of the genes in

Plasmodium species have been found to encode relatively large proteins that contain a large number of low complexity regions.[32, 33, 116] Another study has reported that nearly 90% of all proteins from chromosomes 2 and 3 in *Plasmodium falciparum* contain low complexity regions.[34] The uniqueness of such genomes and the presence of a large number of low complexity regions cause difficulties in identifying homologues of *Plasmodium* proteins.[34] Many *Plasmodium* proteins are also shown to be amenable to expression in heterologous systems.[117] One of the leading explanations for low expression yields is the presence of intrinsically disordered regions.[34]

Summary of problem statements

Within the last two decades, the field of intrinsically disordered proteins has gradually metamorphosed from a novel hypothesis to a steadily maturing paradigm in cell biology. The arguments presented in the leading sections of this topic, allow an easy route to identify the fascinating diversity of biological processes and functions that make use of intrinsic disorder. From well-documented cases of signaling and transcription related proteins (MoRFs) to post-translational modifications (phosphorylation) to low-complexity regions in a model pathogen, understanding the role of intrinsically disordered proteins in realizing functionality within each of these complex phenomena cannot easily be overlooked. With this idea in the background, this dissertation focuses on improving our understanding of how *intrinsic* is intrinsic disorder and investigate the role it can potentially play in helping expand our knowledge about other biological functions such phosphorylation, protein interactions and pathogenesis.

CHAPTER THREE: CONSISTENCY OF DISORDERED REGIONS

Background

Successful protein crystallization depends on the complexity of its structure but also on a number of experimental or environmental factors including purity of the protein sample, temperature, ionic strength, pH, and precipitants such as ammonium sulfate or polyethylene glycol.[118] Traditionally, crystallization begins with procuring a sizable quantity of a protein (15mg to 1g) before growing a crystal.[119] The experimental conditions are then systematically varied to determine optimal conditions for crystal formation. Once a reasonable estimate of conditions is made, the protein solution is supersaturated to facilitate crystal formation typically using vapor diffusion. Despite a number of steps that differ from the physiological conditions, there is evidence that protein structure, though a static representation, often corresponds to its native state.[118] To the best of our knowledge, no study to-date has singularly focused on the influence of experimental conditions at the time of crystallization on the global structure, especially to their influence on disordered protein regions.

A survey of existing literature showed results from two recent studies that document the effects of varying pH conditions on regions of intrinsic disorder in the same protein.[120, 121] More specifically, Palaninathan et al., [121] report conformational changes observed in the tertiary and quaternary structures in the crystals of the native human transthyretin (TTR) at pH = 4.0 and pH = 3.5. The crystal structure of TTR at pH value of 4.0 reveals that the native fold of the tetramer, including the crucial functional EF helix-loop region between residues 75 and 90, remains mostly undisturbed. In contrast, in the

crystal structure at pH = 3.5, the EF helix-loop region was completely disordered. Both of these structures were studied at 1.7Å resolution. Zurdo et al., [120] studied two yeast ribosomal stalk proteins, P1α and P2β that despite high sequence similarity have different functional roles. Concluding arguments presented in by Zurdo et al., [120] suggest that differences in function could be associated to structural differences between both proteins. Even though neither protein is compact and regular in solution, under physiological pH and temperature, P1α is mostly disordered with very little helical content in comparison to P2β. This residual structure is reported to disappear at temperatures below 30°C, but is reportedly regained under low pH conditions or with addition of trifluoroethanol. In addition to experimental studies, computational analyses of redundant sets of experimental protein structures for identical proteins provide evidence of the existence of numerous protein fragments observed in ordered and disordered states.[122] Results presented by Zhang et al., suggest that disorder may not always be an intrinsic (or physiological) feature for some fragments of proteins. In other words, some disordered protein fragments may not necessarily lack structure in all proteins containing them. The authors also hint towards the existence of a new class of fragments that lie precariously on the boundary between order and disorder, and, therefore, are more likely to be found in one state or another depending on environmental conditions or post-translational modifications. Owing to their presence in two different states, such fragments have also been referred to as the 'dual-personality' fragments.[122]

A quick search of Protein Data Bank (PDB) by us resulted in the discovery of additional examples where slight changes in experimental conditions strongly correlated with the presence or absence of regions of intrinsic disorder. The following paragraphs

discuss two such proteins in further detail.

(i) NusG. N-utilization substance G

(NusG), an important elongation/termination modulator has been well studied in the past for its involvement in translational regulation. Two of its structure representatives in PDB include 1M1H-A and 1NPR-A. Structural alignment between these two molecules returned a z-score of 26 and rmsd of 2Å. Both these molecules are monomers with 100% sequence identity and have been crystallized at identical temperature of 291K and pH values of 5.8 and 7.5 respectively, resulting in different space groups (I222 for 1M1H and $C222_1$ for 1NPR). Despite similar experimental methods used to crystallize both these proteins (vapor diffusion, sitting drop) 1M1H has a 62 residue stretch (G187-I248) at its C-terminal that is completely disordered and missing in the crystal structure (Figure 2). This region is observed in its identical sequence, 1NPR-A. A closer look at the crystal contact information for 1M1H revealed that as many as 31 of the 62 disordered residues are found participating in contact formation.

Figure 2: Molecule 1M1H-A (blue) and 1NPR-A (pink) were crystallized under different pH values (5.8 vs. 7.5) and solved in different space groups. Regions that are observed as disordered in 1M1H-A are colored in red.

(ii) Cyclophilin 40.

Cyclophilin 40 (Cyp40) is one of the principal members of a family of large immunophilins found in mammals. Although the exact biological function of large immunophilins is not well understood, they are believed to be strongly associated with Hsp90 and play a crucial regulatory role in the upkeep of steroid receptor activity. In PDB, this protein has been stored as 1IIP-A and 1IHG-A. 1IIP-A is the tetraclonal conformation of cyclophilin 40, whereas 1IHG-A is its monoclinic form. Both structures have been obtained using the vapor diffusion, hanging drop method with recorded

temperature as 277K, but 1IIP-A was crystallized at a pH of 8.0, while 1IHG-A was crystallized at 6.1. Despite 100% sequence identity between the two proteins, an rmsd of 14.2Å, z-score of 38.6 1IIP-A has a region between residues A299 and Y365 that is absent from the structure of 1IHG-A (Figure 3). A quick analysis of crystal contact formation in 1IIP-A using the CryCo software[123] revealed that as many as 55 of the 67 residues found disordered in it and ordered in 1IHG-A participate in crystal contacts.

Figure 3: Molecule 1IIP-A (blue) and 1IHG-A (pink) were crystallized under different pH values (8.0 vs. 6.1) and solved in different space groups. Regions that are observed as disordered in 1IHG-A are colored in red.

With little doubt it can be stated that the examples presented above shed light on the influence that experimental parameters can exert on disordered residues found in crystallized proteins (a complete list is provided in the supplementary section of Mohan et al[124]). The discovery of such protein pairs also suggests that minor variations in experimental conditions could potentially trigger local structural changes that, in turn, have consequences on the reproducibility of intrinsically disordered regions even for the same protein sequence. In such cases, a protein may crystallize in a different space group that is either caused by a changed structure or can lead to a different set of crystal contacts that stabilize an otherwise disordered region. Alternatively, the presence of a small molecule may influence structural changes.

In the following paragraphs we propose experiments to attempt answers to these questions by investigating reproducibility of disordered regions with variations of sequence and environment and by estimating the upper limit of predictability of disordered proteins. We start with a hypothesis that the experimental reproducibility of disordered regions, e.g. those in crystallized structures, provides the upper limit on the predictability of disordered regions. In addition, we hypothesized that differences in experimental conditions during crystallization can play an important role in limiting the prediction accuracy of computational models.

Materials and Methods

Sequence data sets

Our initial data set *S* comprised of 18,884 protein chains from PDB (March 2008).

Each of these chains was characterized using X-ray crystallography with a high-resolution 2Å (Supplementary, Mohan et al.,[124]). This data set consisted of two subsets: *D*–a set of 14,646 chains containing at least one disordered region of length ≥ 3, identified by missing C-α atoms in the ATOM fields; and *OD*–a set of 4,238 completely ordered chains such that each sequence was ≥ 90% identical to one or more sequences in *D*. For each sequence in *S* we extracted experimental conditions: temperature, pH value, and concentration of salt (e.g. ammonium sulfate, potassium sodium tartrate, sodium cacodylate, and a number of others), whenever available (1 sequence in *D* and 1502 sequences in *OD*, did not have any experimental conditions extracted due to differences in file format). While temperature and pH value can be obtained from designated fields in PDB, salt concentration was mined from REMARK200 and REMARK280 fields and manually confirmed in a number of cases. For simplicity of our analysis, each experimental condition was clustered into two groups, high and low (Figure 4). Temperature was clustered into group high (T_h), containing temperatures greater than or equal to 200 K and group low (T_l), containing temperatures below 200 K at the time of experiment. pH value was clustered into P_h and P_l based on threshold 6.5, while the salt concentration was clustered into S_h and S_l based on the threshold of 100 mM.

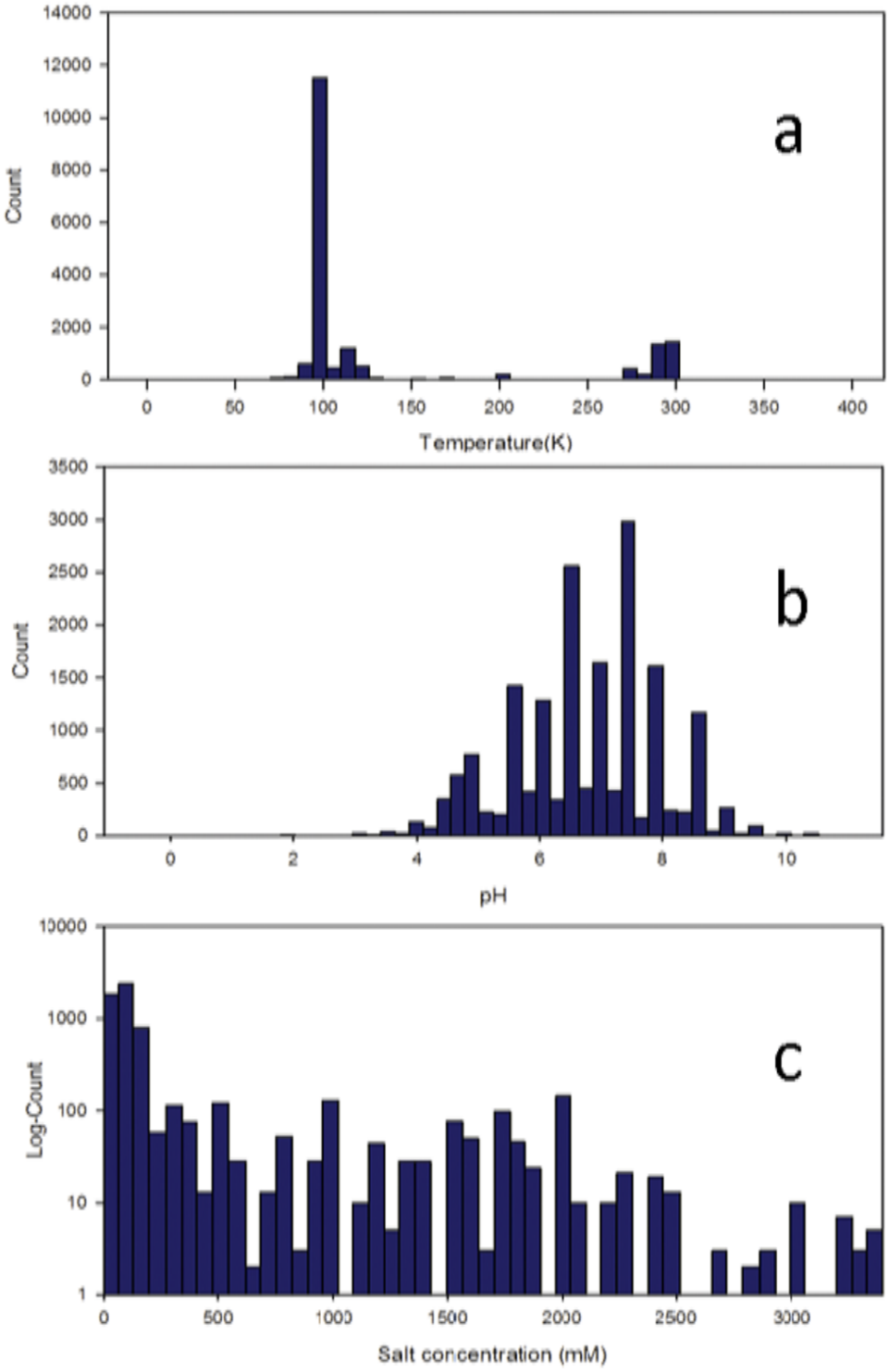

Figure 4: Histogram of observed (a) temperature (b) pH and (c) salt concentration data used.

Non-redundant data sets were constructed in two ways to best support the experiments in this study. In the first approach, the initial data set was split into overlapping subsets, where each subset set D_i contained proteins crystallized at experimental conditions E_i, where $E_i \in \{T_h, T_l, T_hP_h, T_hP_l, \ldots, T_lP_lS_l\}$. For example, data set containing proteins crystallized at conditions T_hP_h, had proteins solved at high temperature and high pH value, but the salt concentration could be from the entire range or unknown. Each data set D_i was also filtered into a non-redundant set $D_{i\text{-}nr}$ such that no two chains had sequence identity greater than or equal to 25% on a global level (BLOSUM62 matrix, gap opening penalty = −11, and gap extension penalty = −1). This approach of defining non-redundant sets was used for estimating experimental reproducibility of disordered regions from class E_i to E_j.

In the second approach, data set D was first filtered into a non-redundant set D_{nr}, and then split into non-overlapping subsets $D_{nr\text{-}i}$, based on experimental conditions E_i (clearly, $|D_{nr\text{-}i}| \leq |D_{i\text{-}nr}|$). This approach was used for evaluating predictors of disordered regions trained for the specific experimental conditions E_i, because it was necessary to ensure that no two proteins within $D_{nr\text{-}i}$ and across different subsets $D_{nr\text{-}i}$ and $D_{nr\text{-}j}$ are similar at 25% or more. The final outcome is then an average over a non-redundant data set. The size of each data set is shown in Table 1.

		Temperature		Salt		pH	
		T_{high}	T_{low}	S_{high}	S_{low}	P_{high}	P_{low}
D	# proteins	3,675	14,822	4,413	1,986	11,715	6,136
	# disordered residues	41,868	220,068	55,870	24,191	158,063	96,378
	# ordered residues	788,496	3,150,810	831,521	393,542	2,534,009	1,306,568
D_{nr}	# proteins	556	1,600	700	392	1,393	846
	# disordered residues	10,196	33,815	13,699	7,724	27,717	18,695
	# ordered residues	161,864	455,274	188,698	106,142	401,679	232,957

Table 1: Number of proteins with available temperature, salt, and pH value data (pre- and post removal of redundant proteins) along with respective number of disordered and ordered residues in each class.

Calculation of experimental reproducibility of disordered regions

Experimental reproducibility of disordered regions was estimated by calculating the mean overlap of ordered and disordered regions in similar or identical protein chains, crystallized at the same or different experimental conditions. Two protein chains were considered to be similar if their global sequence identity was ≥ 90%. This threshold was selected to ensure not only highly similar 3-D structure between two proteins [117, 125], but also similar function.[126]

The mean overlap between two globally aligned proteins $p \in D_{i-nr}$ and $q \in S_j$, where the sequence identity (*si*) between p and q was ≥90%, was calculated as follows. Let O_p and D_p be the sets of positions of ordered and disordered residues in protein p, and O_q and D_q sets of positions of ordered and disordered residues in protein q, respectively, as shown in Figure 5. The residues corresponding to insertions and deletions were ignored.

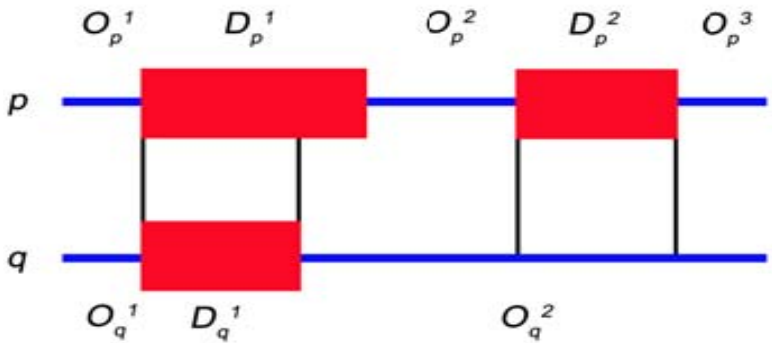

Figure 5: Calculation of the mean overlap between ordered and disordered residues between two homologous proteins p and q.

We calculate the overlap between ordered (o_o) and disordered regions (o_d) as,

$$o_o(p,q) = \frac{1}{2} \cdot \left(\frac{|O_p \cap O_q|}{|O_p|} + \frac{|O_p \cap O_q|}{|O_q|} \right)$$

Equation 1: Calculation of order overlap

$$o_d(p,q) = \begin{cases} \frac{1}{2} \cdot \left(\frac{|D_p \cap D_q|}{|D_p|} + \frac{|D_p \cap D_q|}{|D_q|} \right) & \text{if } |D_q| > 0 \\ \frac{|D_p \cap D_q|}{|D_p|} & \text{if } |D_q| = 0 \end{cases}$$

Equation 2: Calculation of disorder overlap.

The average reproducibility of ordered and disordered regions for a pair (p, q) is calculated as,

$$acc(p,q) = \frac{1}{2} \cdot \left(o_o(p,q) + o_d(p,q) \right)$$

Equation 3: Calculation of mean order and disorder overlap.

We use the term accuracy for the mean overlap due to its similarity to a prediction

process in which ordered and disordered regions in one protein serve as predictions for the other protein and vice versa.

These overlaps can now be generalized to the level of the data sets. An average accuracy for chain p is first calculated over all sequences q that are ≥90% identical to p, denoted by $si(p, q) \geq 0.9$. Then, the average accuracy between data sets D_{i-nr} and S_j, corresponding to experimental conditions E_i and E_j, is calculated as the mean over all proteins p. We formalize the entire calculation as,

$$acc(E_i, E_j) = \frac{1}{N_i} \sum_{p \in D_{i-nr}} \frac{1}{N_j^p} \cdot \sum_{\substack{q \in S_j \\ si(p,q) \geq t}} acc(p, q)$$

Equation 4: Calculation of mean order and disorder overlap over all non-redundant data sets.

Here, $N_i = |D_{i\_nr}|$ and N_j^p is the number of sequences $q \in S_j$ that when aligned to p have sequence identity at least 90%. Note that q can be a completely ordered sequence.

Assuming that the maximum prediction accuracy of intrinsically disordered regions is limited by experimental reproducibility of similar proteins, this approach serves to provide an estimator of the upper limit of the balanced sample accuracy over the given two sets of experimental conditions.

Calculation of amino acid compositions

Disordered residues from sequences belonging to six experimental groups T_h, T_l, S_h, S_l, P_h, and P_l were used to study trends of amino acid compositions for the given experimental groups. Fractional difference of disordered residues from proteins crystallized at two sets of experimental conditions as,

$$\frac{C_{high} - C_{low}}{C_{low}}$$

Equation 5: Fractional amino acid composition at high and low temperature, pH and salt conditions.

Where, C_{high} represents the average amino acid composition of disordered regions for T_h, S_h or P_h groups, and C_{low} represents the average amino acid composition of disordered for T_l, S_l or P_l groups, respectively. Average compositions and confidence intervals were obtained using bootstrapping on the protein level, iterated over 200 independent trials.

Predictor development and evaluation

During predictor development each residue was represented as a vector of 21 features. Twenty relative amino acid frequencies as well as Shannon's entropy were computed over a sliding window $w \in \{21, 31 \text{ and } 41\}$. Before training and testing our predictors, we used the t-test to select the most significant features for our data sets using the p-value threshold of 0.1. All selected features were normalized using the z-score approach before performing a principal component analysis (with 95% of retained variance) in order to further reduce the dimensionality and internal correlation within the data set.

SVM^{light} software[127] was used to predict disordered regions. We evaluated both linear and non-linear kernels, where non-linear kernels were polynomial (quadratic) and Gaussian ($\sigma = 10^{-4}$). The default value was used for capacity C in all experiments. As the main goal of this study was not to refine and train the best predictor for each group of experimental conditions, all parameters were selected based on prior experience with

similar problems, and only the data set representation window w and the kernel type were varied.

In the case when predictors were evaluated within the group $D_{nr\text{-}i}$, we used per-protein 10-fold cross validation to estimate prediction accuracy, while in the case of inter-group validation the predictor was constructed on data set $D_{nr\text{-}i}$ and then evaluated on $D_{nr\text{-}j}$, where $i \neq j$. We estimated sensitivity (sn), specificity (sp), balanced-sample accuracy $acc = ½\cdot(sn + sp)$, and area under the ROC curve (AUC) to evaluate predictor's performance. Sensitivity is defined as the prediction accuracy on the disordered regions and specificity corresponds to the prediction accuracy on the ordered regions. ROC curve shows sn as a function of $1 - sp$ over an entire range of decision thresholds.

Results

Experimental consistency of intrinsically disordered regions & its implications on predictors of intrinsic disorder

To estimate the experimental reproducibility of disordered regions and the limits of its predictability, we studied the overlap of disordered regions in pairs of highly similar proteins crystallized under the same and different experimental conditions. At least one protein sequence in a pair was required to contain one or more disordered regions of length ≥ 3 and two proteins were considered similar if their global sequence identity was ≥ 90%. We investigated the influence of temperature, pH value, and salt concentration used at the time of experiment on the overall reproducibility of disordered

regions. To facilitate this analysis, each experimental condition was clustered into two groups, low and high. Thus, we refer to the experiments carried out under conditions clustered in the same or different groups as *same* (*similar*) and *different* (*dissimilar*) experimental conditions, respectively.

Figure 6 shows the mean agreement of disordered residues obtained in pairs of identical proteins and proteins with sequence identity in the range [90, 100) %. When all experimental conditions were similar, the agreement of disordered residues for identical sequences was 92% (95% for monomers alone). For the same set of experimental conditions, however, and sequence identity in the range [90, 100) %, the agreement of disordered regions decreased to 52% for the set of all protein chains (p-value = $1.4 \cdot 10^{-48}$; Wilcoxon test) and 50% for monomers (p-value = $5.5 \cdot 10^{-10}$; Wilcoxon test). We also investigated the situation when at least one experimental condition was different (e.g. temperature, salt concentration, and/or pH value). For both identical proteins and those in the [90, 100)% range, the reduction of the mean agreement of residues designated as disordered was about 11 percentage points[124]. In an attempt to estimate which of the experimental conditions had the largest influence on the variability of observed disordered regions, a count for each condition was incremented for each protein pair with inexact matches of disordered regions whenever this condition differed. We found that salt concentration had slightly larger impact (39%) than temperature (31%) and pH value (30%), as shown in Figure 6 (inset). Furthermore, we found that, in general, an increase in temperature (6%) and pH value (7%) lead to an increase in the number of disordered residues in identical or similar protein chains. In contrast, an increase in salt concentration (11%) leads to a decrease in the number of observed disordered residues.

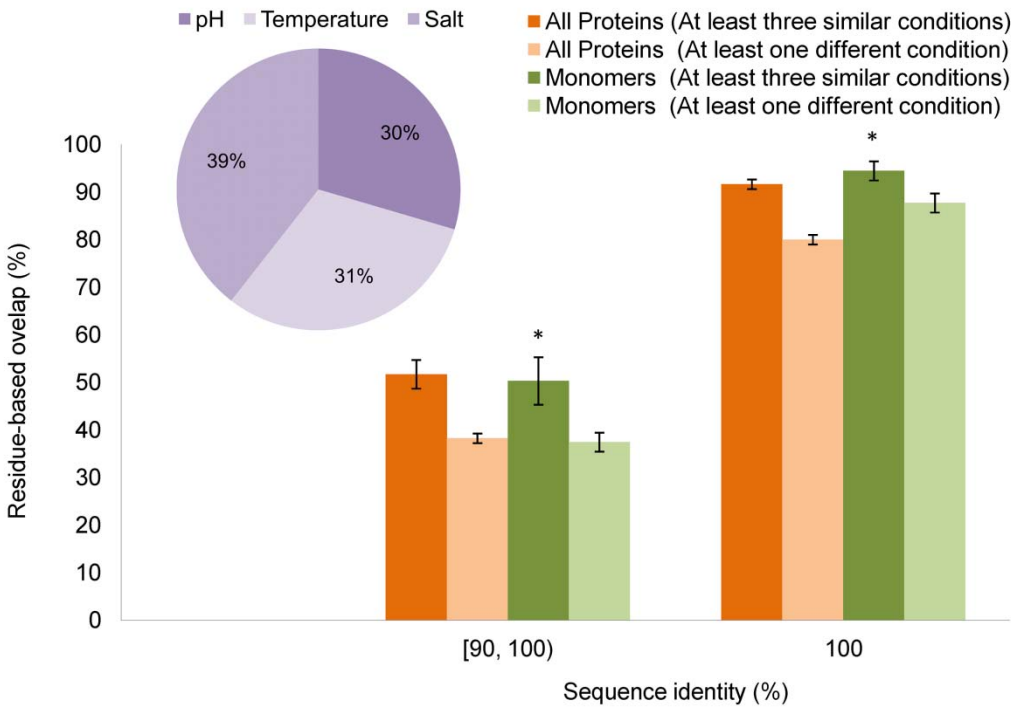

Figure 6: Percentage of overlap of disordered residues between protein pairs with sequence identity [90, 100)% and identical proteins.

We also grouped all pairs of sequences with identity ≥90% into those solved using at least one, two, or three similar experimental conditions and at least one, two, or three different experimental conditions. We estimate that, assuming unchanged experimental platforms for structure determination, the mean agreement of intrinsically disordered residues is 73% (79%, 83%) if one (two, three) or more experimental conditions are similar (Figure 7, left). When different experimental conditions were considered, the agreement of disordered residues was consistently around 50%.

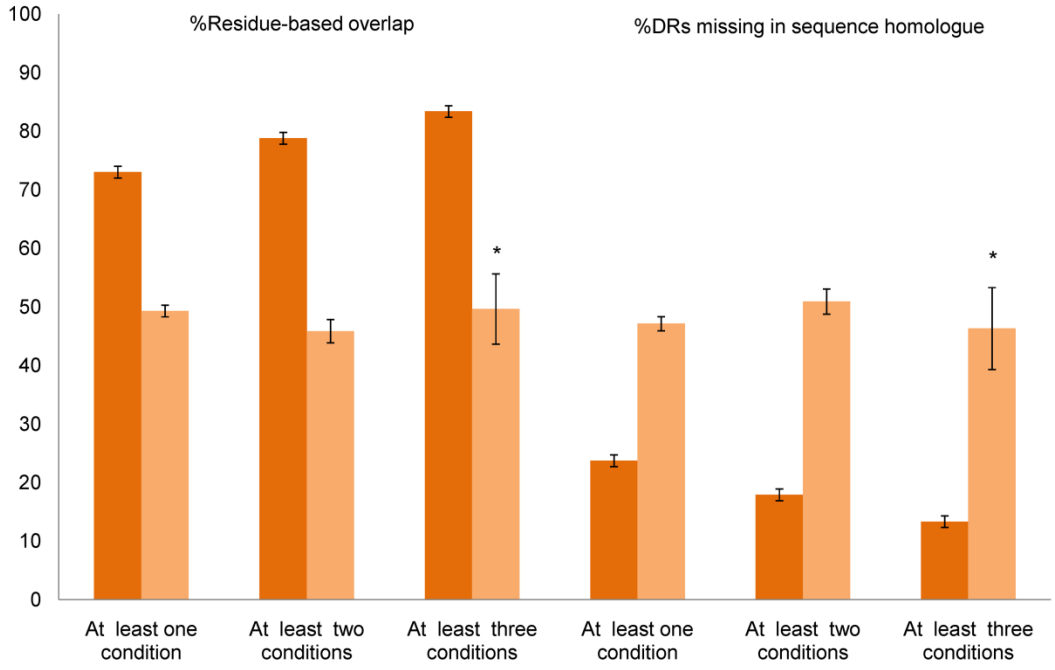

Figure 7: Consistency of disordered residues and regions as a function of experimental conditions.

In Table 2 we present complete results of the consistency measurements for both ordered and disordered regions for the pairs of chains with sequence identity ≥90%. Ordered regions from such pairs of proteins appeared as highly overlapping (>98%), which is a direct consequence of the unbalanced number of ordered and disordered residues in the non-redundant data set (14:1 ratio).

		At least one condition	At least two conditions	At least three conditions
Same	Number of proteins	4086	3488	852
Conditions	Mean D overlap	73.0	78.8	83.4
	Mean O overlap	98.7	99.0	99.1
	Mean accuracy	85.9	88.9	91.2
	DRs missing (%)	23.8	17.9	13.3
Different	Number of proteins	1427	440	42
Conditions	Mean D overlap	49.3	45.9	49.7
	Mean O overlap	98.0	98.4	98.8
	Mean accuracy	73.7	72.1	74.2
	DRs missing (%)	47.2	50.9	46.3

Table 2: Mean overlap for disordered (D) and ordered (O) regions for protein pairs with ≥90% sequence identity crystallized under similar and different experimental conditions.

We estimated the mean agreement of disordered residues using pairs of similar and identical protein sequences wherein experimental information at the time of pair generation was not considered. If identical protein pairs are considered, the mean overlap of disordered and ordered residues was 89% and 99%, respectively. When we considered disordered regions of length 30 or more, the mean overlap was 93% and 98%, respectively (Figure 8). Interestingly, all pairs from our analysis in which long disordered regions significantly differed belonged to dissimilar experimental classes thus strongly suggesting that the appearance of disordered regions is influenced by variations in experimental conditions (e.g. 1COT-B and 1S6P-B). Consideration of similar sequences resulted in a significant reduction in the mean overlap: 31% for all disordered regions and 35% for long disordered regions only. Note that the slightly smaller overlap of disordered

residues, compared to the one from Figure 8, is due to the influence of completely ordered proteins for which we were unable to extract experimental conditions and therefore were excluded from the analysis in Figure 8.

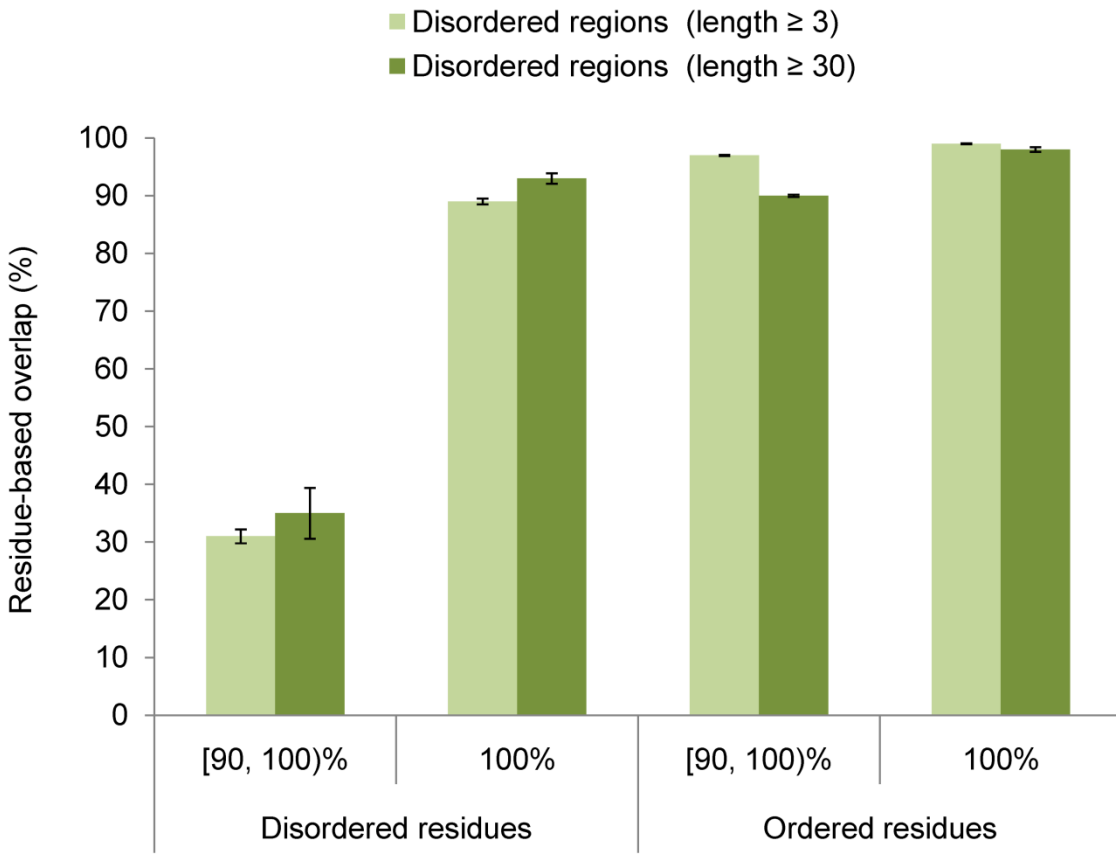

Figure 8: The mean observed agreement between ordered and disordered residues in similar and identical protein chains.

Amino acid compositions

Amino acid compositions for the proteins crystallized under various experimental

conditions are shown in Figure 9A. As can be observed in this figure, disordered residues crystallized under high temperature, salt, and pH conditions exhibit pronounced enrichment in tyrosine (Y), glutamine (Q), and aspartic acid (D) in comparison to those crystallized under low temperature, salt and pH conditions. Unlike these, cysteine (C) and asparagine (N) are depleted in disordered regions crystallized using this group of experimental conditions. Most remaining amino acids appear to be similarly abundant in any of the high or low experimental group, but show some preference for a particular experimental condition. Some clear examples include the enrichment of tryptophan (W) under high salt and temperature conditions and its depletion in disordered regions characterized under high pH conditions. Similarly, isoleucine (I) appears to be enriched in disordered regions crystallized under high salt and high temperature and has no preference for a particular class of pH condition.

Another compositional study compared the mean fractional content for each of the twenty basic amino acids in comparison to those obtained from a disorder data set, DO_{rep}.[128] This data set contains all disordered regions characterized using X-ray, NMR and CD, and has frequently been considered the equivalent of a representative set of intrinsically disordered proteins. Compositions of each of the twenty basic amino acids from disordered regions found in sequences crystallized under each experimental group was compared with those obtained using the representative disorder set (Figure 9B). Not surprisingly, composition profiles of disordered regions from each of the six experimental groups closely mimic the trends observed in the representative disordered data set. More specifically, nearly all amino acids on the right hand side of the plot (G, Q, S, N, P, E, K) are enriched in all six data sets. On the other hand, buried residues (W, F, I) appear to be

depleted in comparison to DO_{rep}. Interestingly, there is a clear difference prevailing between cysteine (C) and threonine (T) content in disordered segments crystallized in contrasting experimental conditions. Although cysteine seems to be depleted under high temperature, salt and pH conditions, threonine is enriched in data sets belonging to the same group of conditions. Also, disordered residues crystallized under low temperature, salt, and pH conditions have relatively higher concentrations of amino acids such as cysteine(C), asparagine (N), and proline (P) than a representative data set of disordered regions. A case-by-case look at some of the amino acid profiles suggests that higher pH conditions bring about decreased content of with tryptophan (W). Similarly, crystallization under high temperatures negatively affects the histidine (H) content in disordered regions. Finally, enrichment in comparison to representative contents of alanine (A) and glutamine (Q) in disordered proteins is apparent in disordered residues crystallized using high pH and high salt concentrations.

To summarize, disordered regions from homologous sequences crystallized under varying conditions exhibit preferential compositional profiles with respect to one another. Differences between high and low experimental classes are especially pronounced in the case of buried residues such as W, C, I and Y. Flexible residues, from disordered regions on the other hand seem much less affected by changes in experimental conditions.

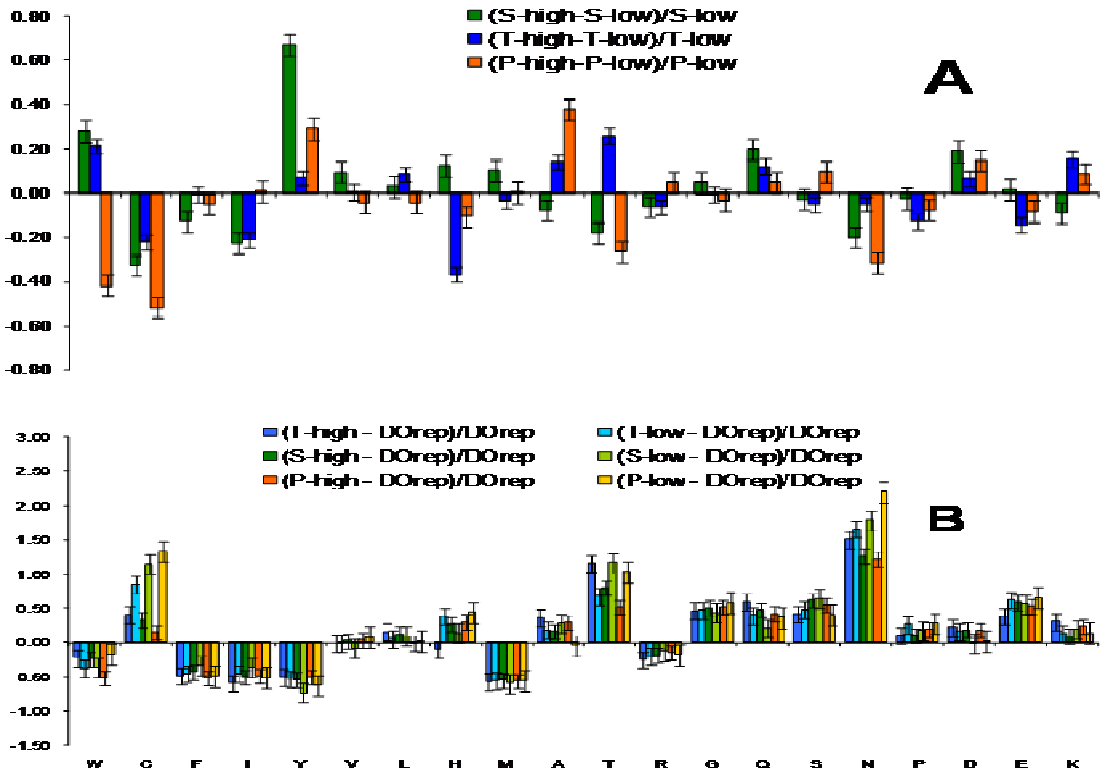

Figure 9: Fractional amino acid composition profiles of proteins crystallized using high temperature, salt and pH conditions with respect to (A) proteins crystallized at low temperature, salt and pH conditions and (B) a representative set of disordered proteins, DO_{rep}.

Prediction accuracies

With the goal of assessing whether a condition-specific predictor of intrinsically disordered regions is feasible and useful, we constructed a prediction model for each experimental group of data and evaluated it both on the data for that experimental group as well on the remaining groups. Within its own group, accuracy of each predictor was estimated using 10-fold cross-validation, while the performance across groups was estimated by training the model on the entire data set for one experimental group and testing it on the other group (out-of-sample testing). All proteins within one group and

across different groups were non-redundant, i.e., the global sequence identity between any two sequences within or across data sets was below 25%. Full details of the data set construction process have been provided previously in the Materials and Methods section.

Due to the small number of sequences in various subsets and especially due to the small number of disordered residues, we used only the six largest categories of experimental conditions. That is, $E_i \in \{T_l, T_h, P_l, P_h, S_l, S_h\}$, where $E_i = T_l$ indicates that each protein in data set D_i was crystallized under low temperature, while the remaining two variables (pH value, salt concentration) were allowed to have any value (low, high or unknown). As described earlier in the Materials and Methods sections, features for all predictors comprised of moving averages of 20 basic amino acid compositions over a window of length w as well as the sequence complexity, calculated using Shannon's entropy formula.[129] A support vector machine with linear, polynomial (quadratic) and Gaussian ($\sigma = 10^{-4}$) kernels with a default value for the capacity parameter was also constructed. Table 3 lists the estimated prediction accuracy (*acc*) and area under the ROC curve (*AUC*) for each case. It can be observed that each individual model achieved higher accuracy when evaluated on the proteins from its own experimental group, compared to the accuracy on a different experimental group. The difference in accuracy ranged from 1.2 percentage points (Gaussian kernel, $w = 21$) to 5.7 percentage points (polynomial kernel, $w = 21$), suggesting that there exist certain amino acid biases in each group that can be exploited by the machine learning models and that the optimal decision boundary is also condition specific. Similarly, the difference in the AUC values ranged from 0.9 percentage points to 6.9 percentage points.

As summarized in Table 3, we also find that on average the polynomial kernel achieved highest accuracies and AUCs (w = 21) when the training and test data sets belonged to similar experimental conditions. Interestingly, the same predictor also produced the lowest accuracies and AUCs (w = 41) when trained and tested on data sets belonged to different experimental groups.

		Mean Accuracy(AUC)	
	Window	training & testing datasets from similar experimental conditions	training & testing datasets from different experimental conditions
Linear	21	71.8(78.8)	70.4(77.5)
	31	71.2(77.5)	69.8(76.2)
	41	70.1(75.8)	68.6(74.5)
Polynomial	21	74.3(81.5)	69.3(75.6)
	31	74.1(81.1)	68.4(74.2)
	41	73.4(79.8)	67.4(73.0)
Gaussian	21	72.0(79.0)	70.8(78.1)
	31	71.6(78.0)	70.1(76.7)
	41	70.5(76.4)	68.9(75.0)

Table 3: 10-fold CV results for condition specific predictors of disorder over overlapping windows of length 21, 31 and 41 for linear, non-linear and Gaussian kernels.

In Table 4, the experimental scenario was slightly tweaked. First, in predicting disordered regions using similar conditions, we again split D_i into training ($90\% * |D_i|$) and test ($10\% * |D_i|$) sets and applied cross-validation. However, the training set was augmented by including all sequences from D, which are non-redundant (<25% sequence

identity) with respect to any sequence in D_i. In this way, we tested whether simply enlarging the data set, without considering experimental conditions, is likely to improve classification accuracy. The results in Table 4 show that for the different experimental conditions the addition of extra sequences (potentially also from the same experimental group) is beneficial (increase of about 2 percentage points). However, for the same experimental conditions the addition of extra sequences had either no significant effect, or it caused a decrease in predictor performance (2-3 percentage points for polynomial kernel). This indicates that a higher quality prediction of disordered regions is achievable when experimental conditions are accounted for and that the changes in disordered regions with respect to experimental conditions are predictable to some degree.

		Mean Accuracy	
	Window	training & testing datasets from similar experimental conditions	training & testing datasets from different experimental conditions
Linear	21	71.4(78.3)	71.0(78.1)
	31	70.6(76.9)	70.3(76.8)
	41	69.3(75.2)	69.1(75.1)
Polynomial	21	72.0(79.2)	71.0(77.9)
	31	71.4(78.5)	70.3(76.9)
	41	70.5(77.1)	69.3(75.8)
Gaussian	21	71.7(78.8)	71.3(78.3)
	31	71.0(77.6)	70.5(77.0)
	41	70.0(76.1)	69.5(75.6)

Table 4:10-fold CV results for new predictors of disorder over overlapping windows of length 21, 31 and 41 for linear, non-linear and Gaussian kernels.

CHAPTER FOUR: DISORDER-MEDIATED MOLECULAR RECOGNITION

Background

A previous work presented what is now known as the Molecular Recognition Feature or MoRF hypothesis to explain the process of disorder-mediated protein interactions.[44] This hypothesis proposed that disordered regions of relatively short lengths (e.g. length ≤ 70) undergo a disorder-to-order transition upon binding to their partner. Such regions have been estimated to be common in proteomes and especially enriched in those belonging to higher organisms. In 2005, Oldfield et al., [44] focused on a set of 15 disordered protein fragments from 13 PDB proteins that gained α-helical structure upon binding (also known as α-MoRFs) to their partners. A good example conforming to the MoRF hypothesis is the p53 α-MoRF (as shown in Figure 10). Another example of an α-MoRF although of length greater than 70 residues is, calmodulin (also known as CaM). CaM is an important transducer of calcium signals and is known to interact with multiple partners in eukaryotic cells.[130] The CaM molecule reportedly undergoes disorder-to-order transitions upon binding to each of its target proteins.[131] Here we propose the development of a comprehensive data set of Molecular Recognition Features (MoRFs) and their partners as available from the RCSB Protein Data Bank. We suggest using this data set to study MoRFs in greater detail to directly impact our understanding of MoRFs and MoRF-binding proteins, especially, their physiochemical and structural propensities, interaction surface properties besides improving our knowledge about disorder-to-order interactions (also known as coupled-folding and binding interactions. We further propose to use this data set to study the differences between other protein interaction surfaces with

respect to MoRF–interaction surfaces to examine the computational predictability of MoRF interaction surfaces.

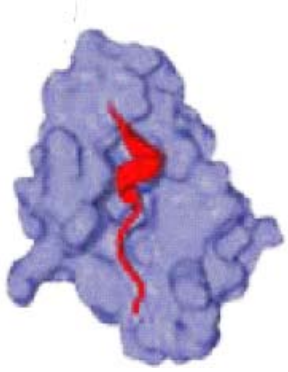

Figure 10: α-MoRF of p53 bound to MDM2 (PDB code 1YCR)

We anticipate that such a study will help us attempt answers to a variety of currently unknown aspects of disorder mediated protein interactions including the identification of other specialized classes of MoRFs, how MoRFs distinguish themselves from representative disordered and ordered sequences, a baseline set of characteristics corresponding to MoRF-binding partners, their compositional preferences and more.

Materials and Methods

A data set of MoRFs and their partners

An initial set of MoRFs was collected from PDB SEQRES file by selecting protein chains of less than 70 residues bound to other protein chains greater than 100 residues. The PDB SEQRES data set contains all the protein sequences available in the PDB along with the residues observed in protein crystals or in solution. Sequences in this

data set also include residues not present in the crystal model (e.g. disordered, lacking electron density, cloning artifacts, and His–tags). The choice for selecting protein chains with lengths less than 70 residues was ad-hoc and stemmed from the generally accepted notion that such proteins would be less likely to form a stable structure without the presence of a partner, especially, prior to interaction. In other words, such protein chains would less likely be able to develop significant buried surface area before participating in the molecular recognition event.

Using these constraints, a starting data set consisting of 2,512 protein chains was assembled, where upon these chains were reduced to give the final non-redundant MoRF data set. Information for each of these steps can be found in more detail in.[107] PDB files corresponding to the initial 2,512 proteins were downloaded to obtain sequences, secondary structure and information on Ramachandran's φ and ψ angles. Once the initial data set was gathered, the next step was to remove all chains containing ambiguous sequence information (e.g. sequences containing X designations instead of standard amino acid designations). Protein fragments with lengths ≤10 were also removed to facilitate mapping MoRF chains to their parent sequences. That is, many MoRF chains in the PDB are fragments of longer proteins and such short peptides may not be long enough to be specific to the parent protein sequence. At the end of this step, 1,261 chains remained. The next step was to remove sequence redundancy, which was done through application of length-dependent thresholds of sequence identity. This was necessary in order to overcome length variations and the overall short lengths of the MoRFs. It has been shown that pair-wise sequence identity alone is a poor definition of the twilight zone, which is the point where the inference of structural similarity from sequence

similarity becomes ambigous.[132] The use of length dependent cut-offs to ascertain degrees of similarity within the MoRF data set helps to correlate sequence alignments to actual structural similarity more strongly. Rost's formula[132] was used to dynamically calculate the sequence identity threshold based on each chain's length. Our final data set comprised of 372 clusters with a single representative MoRF for each cluster. Table 5 (modified from Mohan et al.,[107]) lists the number of MoRFs including the number of residues at each preprocessing step.

	Number of MoRFs
Total qualifying PDB fragments	2,512
Removal of fragments with ambigous sequence information	1,261
Filtering for redundancy	372

Table 5: Number of MoRFs after each data processing step.

The selected structures consisted of 336 X-ray structures, 23 NMR structures, and five cryo-electron microscopy structures. The average resolution of the X-ray structures was 2.41(±0.60) Å. The minimum number of members per cluster was 1 and the maximum number of members was 177 (Thrombin, Alpha-Thrombin). Analysis of the lengths for all MoRFs revealed that approximately two-thirds of the selected chains had lengths between 10 and 20 residues. All but 53 MoRFs were found with a mapping to longer sequences. For comparative analyses, three data sets of ordered proteins were also prepared, namely: ordered monomers (OM), ordered complexes (OC), and PDB select 25[133] (PDB_25). The OM set contained unique monomeric proteins from PDB X-ray structures with an average resolution of 2.04(±0.50) Å. The OC set represents chains

from PDB X-ray structures that are ordered prior to complex formation, with an average resolution of 1.86(±0.43) Å. The PDB_25 set is a non-redundant data set that is representative of all chains in the PDB, where no chain in this set has a resolution poorer than 3.5 Å. Secondary structure assignment for all MoRFs was determined using the DSSP program, which was designed to standardize protein secondary structure assignments.[134]

MoRF and MoRF-binding protein interfaces

A follow-up study done later in 2007[135] studied all of the binding partners in complex and not in complex with their respective 372 MoRFs. Since the main goal of this study was to analyze the interaction surfaces between MoRFs and their binding partners, external data sets from previous studies done by Jones and Thornton[136, 137] and Conte et al., [138] were used for comparison controls. Protein surfaces and interfaces were analyzed at the residue level using the Molecular Surface (MS) software package from Biohedron (http://www.biohedron.com). This package is an implementation of the Connolly surface algorithm[139] and was run for individual MoRF-MoRF partner chains to determine solvent accessible surface area for them.

Calculation of amino acid compositions

Twenty basic amino acid compositional profiles for MoRFs were derived using the fractional difference between MoRF compositions and PDB_25 compositions. The fractional difference was calculated as shown in the following equation:

$$\frac{C_{MoRF} - C_{order}}{C_{order}}$$

Equation 6: Fractional amino acid compositions of MoRFs with respect to PDB Select 25 data set.

Here, C_{MoRF} is the averaged amino acid composition of a MoRF data set, and C_{order} is the averaged amino acid composition in PDB_25. Standard errors for amino acid compositions were calculated from 200 bootstrap iterations.

A later update to this database using the October 2006 version of PDB SEQRES data set improved the original of number of MoRFs from 372 to 486 and 2,512 total fragments to 4,410 qualifying PDB fragments.

Results

Visual inspection of MoRFs

The structures of a few examples of MoRFs were visualized with respect to their residue-wise VLXT predictions in Figure 11.[107] This illustration provides an example of each of α, β, ι, and complex-MoRFs (e.g. Figure 11(a), (g), (c), and (d), respectively), and also provides examples of structural polymorphism in a MoRF bound to two different partners (Figure 11(b) and (c), (d) and (e)).

Tumor suppressor p53 is a protein well known for its pivotal role in the regulation of cellular division processes in response to mutations and DNA damage and is estimated to be present in more than half of all known human cancer occurences.[140] The upper plot in Figure 11 shows the four domains crucial to p53 function in context of the VLXT

prediction for p53. These four domains include: an N-terminal MoRF, the DNA binding domain (Figure 11, box 1), the tetramerization domain (Figure 11, box 2), and a C-terminal MoRF (Figure 11, both overlapping red boxes). Both the N and C-terminal MoRFs have been verified to be disordered in the absence of binding partners.[141, 142] In p53, the N-terminal fragment is an example of an α-MoRF and corresponds to the transactivation domain of p53 bound to MDM2 (Figure 11(a))[143] where this interaction inhibits p53′s transactivation activity and downstream cell cycle arrest.[144] The C-terminal region, on the other hand, contains a fragment that serves as a good example of an ι-MoRF and is shown interacting with the CDK2/cyclin A complex (Figure 11(b)). This interaction facilitates phosphorylation and subsequent activation of p53.[145] An overlapping region of p53 also interacts with S100ββ, an interaction that blocks oligomerization and phoshorylation[146, 147], in turn blocking activation, of p53, and forms an α-helix when bound (Figure 11(c)).[141] The C-terminal region of p53 represents a single MoRF that interacts with multiple partners; it is an example of the richness of function possible under the MoRF model.

Another protein known as the Wiskott–Aldrich syndrome protein (WASP) plays an important role in Arp2/3-mediated modulation of the actin cytoskeleton.[148] Four domains important for WASP function are shown in the context of the WASP VLXT prediction (Figure 11, center plot), which are, from the N to C termini: the N-terminal WH1 domain (Figure 11, box 3), a complex-MoRF that corresponds to the GTPase binding domain (GBD; Figure 11, both overlapping red boxes), an α-MoRF corresponding to the WH2 domain, and the C-terminal VCA region (Figure 11, green box). Of these, only the GDB MoRF is currently found in our data set as the WH2-actin

complex structure (Figure 11(f))[149] was released after construction of the data set. In addition to this, the VCA-GDB complex (Figure 11(e))[150] is found as a single chimerical chain, which was discarded by the MoRF selection criteria. However, both the VCA and WH2 domain are consistent with MoRF criteria and are therefore considered here. The VCA domain interacts directly with the Arp2/3 complex and, together with the actin binding activity of the WH2 domain (Figure 11(f)),[149] stimulates polymer nucleation.[148] Interestingly, the Arp2/3 binding activity of the VCA MoRF is auto-inhibited by the GDB MoRF (Figure 11(e)). This auto-inhibitory interaction is interrupted by binding of the GDB MoRF to activated Cdc42 (Figure 11(d)), which releases the VCA MoRF to interact with Arp2/3. The two GDB MoRF complexes (Figure 11(d) and (e)) show radically different structures, which is an extreme example of multiple binding affinities through bound structure conformational heterogeneity.

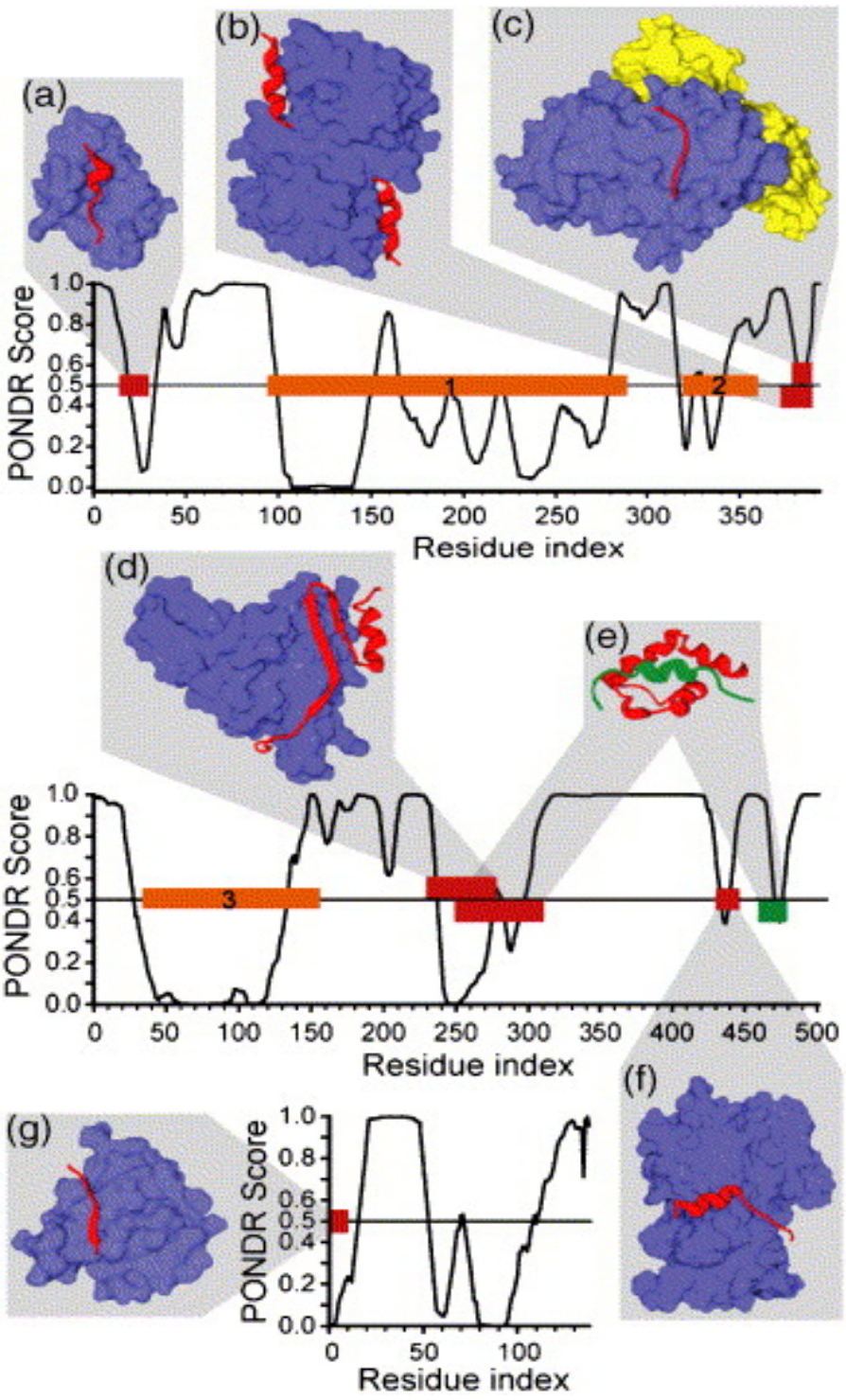

Figure 11: Examples of VLXT predictions of MoRF containing proteins and complexes between MoRFs and their binding partners.

Characteristics of MoRFs

Our study determined that there are primarily three classes of MoRFs based on their secondary structure types. These classes included 68 α-, 20 β-, and 176 ι-MoRFs respectively. In addition to these three classes, a fourth class comprising of 108 complex-MoRFs containing a combination of α-helical, β-structural and/or irregular structural elements, was also identified.

Secondary structure analysis revealed that 27% of the residues in the MoRF data set have an α-helical conformation, 12% have β-sheet like conformation and approximately 48% were residues of irregular structure. The remaining 13% were residues with missing coordinates in the corresponding PDB files suggesting their highly flexible (disordered) nature. We compared these results with those from the OM data set (Figure 12).

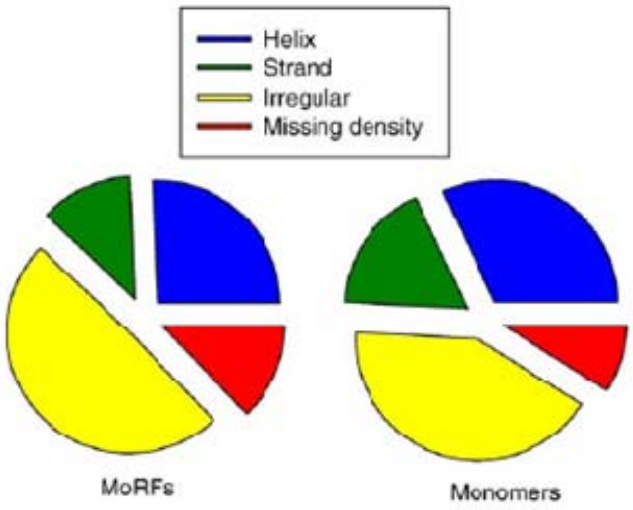

Figure 12: Secondary structure distribution of residues in the MoRF data set and in the OM data set.

The two distributions are found to be significantly different by a $\chi 2$ (chi-square) test (p-value $=4\times10^{-80}$), but the relative $\chi 2$ value (0.003) indicates that the difference although significant, is relatively minuscule. On an average, MoRFs have a higher content of irregular structure (+6%) than OMs. More importantly, in comparison to OM residues, a larger number of MoRF residues have missing density indicating a possible bias for intrinsic disorder. These observations are supplemented by a proportional decrease in the α-helix and β-strand content in MoRFs. Between OM and MoRF data sets, irregular structure is the most abundant secondary structure type followed by α-helix whereas β-strand is the least.

The effect of local and non-local interactions on MoRFs

Since the adoption of secondary structural configurations can be influenced by local as well as non-local interactions, we studied the relative roles of both these types of interactions in determining the secondary structures of MoRFs with respect to OMs. The role of non-local interaction, both inter-chain interactions and intra-chain interactions between residues distant in sequence, in the determination of local structure is contentious. Some authors have found local interactions to be dominant over long range interactions in determining local structure,[151] while others have shown long range interactions to have a direct effect on accuracy of predictions of local structure.[152] Here, we take the view that different proteins, and likely different regions in the same protein, vary in the relative contributions of local and non-local interaction to local structure, but we make no attempt to contribute to the debate over the relative degree of these

contributions. The role of non-local interactions was examined by comparing the secondary structure prediction accuracies, using the single sequence PHD predictor,[153] for the MoRF and OM data sets. In the single sequence mode, the PHD secondary structure prediction algorithm uses a series of neural networks applied over the local sequence only. As predictions consider only windows of an entire sequence and not the complete protein sequence itself, predicted secondary structure is assumed to be a good indicator of the secondary structural preferences of the local sequence, excluding influences from non-local interactions and bound partners.

Typically the PHD algorithm uses sequence profiles generated from multiple alignments for better accuracy. Using multiple alignments allows a local representation of information about non-local interactions. Since this was not an intended goal of this study, we limited ourselves to the use of a single-sequence (or non - multiple-sequence alignment) mode PHD algorithm for this analysis. In the non-MSA mode, differences between DSSP assigned secondary structure and predicted secondary structure can help determine the extent to which interactions between distance residues in a sequence (in the case of monomers) or binding partners (in the case of MoRFs), have an influence on the final protein conformation. We find that the overall prediction performance is consistent with the reported accuracy of PHD, with a single sequence prediction accuracy of 61% and a reduced accuracy of prediction for β-strands relative to α-helices and irregular structure (Table 6).

	α-helices (%)	β-strands (%)	irregular (%)	Missing density (%)
MoRFs	{74,9,17}	{11,55,34}	{21,15,64}	{18,10,72}
OM	{65,9,32}	{16,51,32}	{20,18,61}	{31,27,41}

Table 6: PHD secondary structure prediction accuracies for MoRFs and OM assigned secondary structure classes; Table entry legend: {predicted helix, predicted beta-strand, predicted irregular}

Between MoRFs and OMs, the accuracy of secondary structure predictions for MoRFs is better than that for OMs by 5%. Furthermore, prediction accuracy is better for MoRFs for all defined secondary structure types, where much of this difference is due to the prediction accuracy for α-helices (+9%) rather than for β-strand (+4%) or irregular structure (+3%). These data suggest that the local secondary structural propensity of MoRFs is somewhat better preserved in their bound state, especially for helical regions, than the local secondary structural inclination of OMs. The secondary structure predictions for regions of missing density are also revealing. We observe that missing density in MoRFs is predominantly predicted to be in an irregular conformation with much less of the missing density in OMs predicted to be irregular (+31%). As mentioned earlier in the Background section, all missing density cannot be treated as intrinsic disorder, since missing density may correspond to mobile, structured domains or even other artifacts of crystallization experiments. However, the lower content of predicted regular secondary structure in MoRFs, relative to OMs, may be an indication that the missing density in MoRFs is more likely to be disordered than the missing density in OMs. This provides further support to the idea that MoRFs occur in a disordered context, since the majority of missing density in these chains occurs in the N and C-terminal tails

of the crystallized fragments.

Structural types were further analyzed in terms of contiguous structural regions. The MoRF set was broken into 1,880 regions of sequence contiguous elements of secondary structure or missing density. Examination of the different structural types[107] revealed that 14% regions were α-helical while 20 were β-strands. The larger proportion of β-strand regions than α-helix regions can be reconciled with the larger number of α-helix residues than β-strand residues (Figure 12) by observing that α-helical regions are on average longer than β-strand regions, with average lengths of 10±8 and 3±2 residues, respectively. More than half of the total regions (~53%) were found to have an irregular conformation. The remaining 13% regions were disordered. The average lengths of irregular regions (5±5 residues) and missing density regions (5± 6 residues) are of intermediate length compared to α-helices and β-strands. These results have also been tabulated in Table 7 for the reader.

Region length	Missing density (%)	α-helices (%)	β-strands (%)	Irregular (%)
1 - 9	86	62	99	85
10 -19	11	28	1	13
20 -29	2	6	0	1
30 - 69	1	3	0	1

Table 7: Region wise distribution in different structural types of MoRFs.

Composition profiles, charge and aromatic content in MoRFs

It has been reported that local amino acid composition, flexibility, hydropathy, charge, coordination number and several other physiochemical properties of intrinsically

disordered protein regions are significantly different from the same characteristics derived using ordered protein regions.[10, 154] These properties have been examined for MoRFs, in comparison to ordered proteins, to explore the order/ disorder propensity of MoRF regions. For this analysis, PDB_25 was used, since this set has been well characterized in terms of composition relative to intrinsically disordered proteins.[10, 154] Previous research[31, 154] has shown that, amino acid profiles derived using intrinsically disordered proteins shows depletion of order-promoting residues, such as C, V, L, I, M, Y, F, and W, and the abundance of disorder-promoting residues, such as Q, S, P, E, K, G, and A, relative to ordered proteins. We observe a similar trend for the MoRF data set. More specifically, in comparison to the PDB_25 data set (Figure 13) MoRFs are enriched in many of the disorder promoting residues such as, R, G, S and P and depleted in many of the order promoting amino acids such as W, I, Y, V and L. These biases suggest that MoRFs are similar in composition to general intrinsically disordered proteins. Intriguingly, some other biases contradict this simple explanation. For instance, MoRFs are depleted or show similar composition to PDB_25 in charged residues except R, which are believed to be disorder promoting. It is likely that the lower charge density of R relative to K makes it less likely to maintain a twofold role in both ordered and disordered states. Another bias inconsistency between MoRFs and intrinsically disordered proteins is the enrichment of C and F in MoRFs. Cysteine and phenylalanine are well known to be order promoting and found depleted in disordered proteins. Since cysteine is important for the formation of disulfide bonds, its presence was further investigated. We find that of the 372 MoRFs, 36 contain at least one intra-chain disulfide bond, 18 contribute to at least one inter-chain disulfide bond, and 4 have at least one of

each intra and interchain disulfide bonds. The cysteine residues involved in these bonds account for 73% of the cysteine residues in the MoRF data set, which suggests that reduced cysteine is less prevalent feature of MoRF regions. The presence of intra-chain disulfide bonds in MoRF sequences has clear implications for the hypotheses that these sequences are disordered in the absence of binding partners, as disulfide bridges are well-known to stabilize proteins.[155] Since disulfide bonds potentially stabilize as many as 11% of MoRFs in this data set in the absence of their binding partners, we classify these fragments as pseudo-MoRFs.

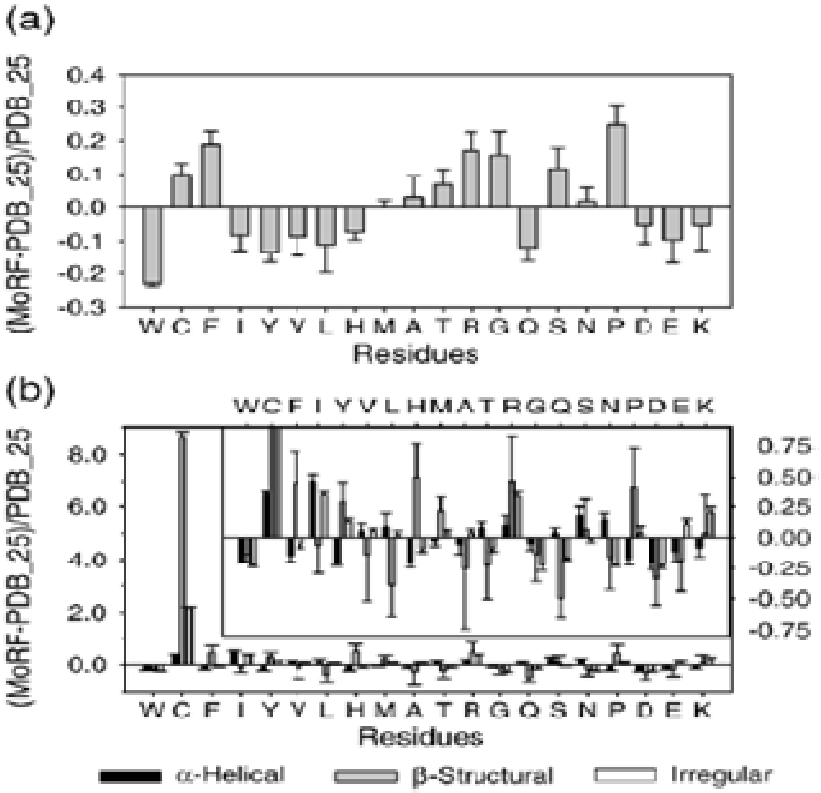

Figure 13: (a) Relative amino acid composition of MoRFs with respect to PDB_25. (b) Relative amino acid composition of different structural types (α-helical, β-structural, and irregular) of MoRFs with respect to the same structural types in PDB_ 25. Inset represents graph (b) with a reduced relative frequency range.

A comparison of the total charge (K+R+D+E), net charge (K+R–D–E), proline, and aromatic content (F+W+Y) of MoRF proteins and PDB_25 proteins is shown in Figure 14. Despite being depleted in lysine, aspartic acid, and glutamic acid, MoRFs demonstrate a higher net charge than the PDB_25 proteins. The enrichment in arginine in MoRFs is apparent from the positive net charge of MoRFs, compared to the negative net charge in PDB_25. This is similar to a previous description of intrinsically disordered proteins.[156] MoRFs also show lower proportions of aromatic amino acid residues in comparison with PDB_25 proteins, despite being enriched in phenylalanine. However, the vast majority of MoRF regions contained at least one aromatic residue, often phenylalanine. This is consistent with the molecular recognition function of MoRFs, since the side-chains of aromatic amino acids tend to make strong and specific interactions[157]. Finally, the proline content observed in MoRFs exceeds that found in PDB_25 proteins by nearly 50%. This high concentration of proline was further examined for the presence of polyproline II helices by Mohan et al in 2006[107] the results of which showed that while many MoRFs contained regions of polyproline II helix, this conformation does not occur as the predominant secondary structure for any of the examples found to date.[107]

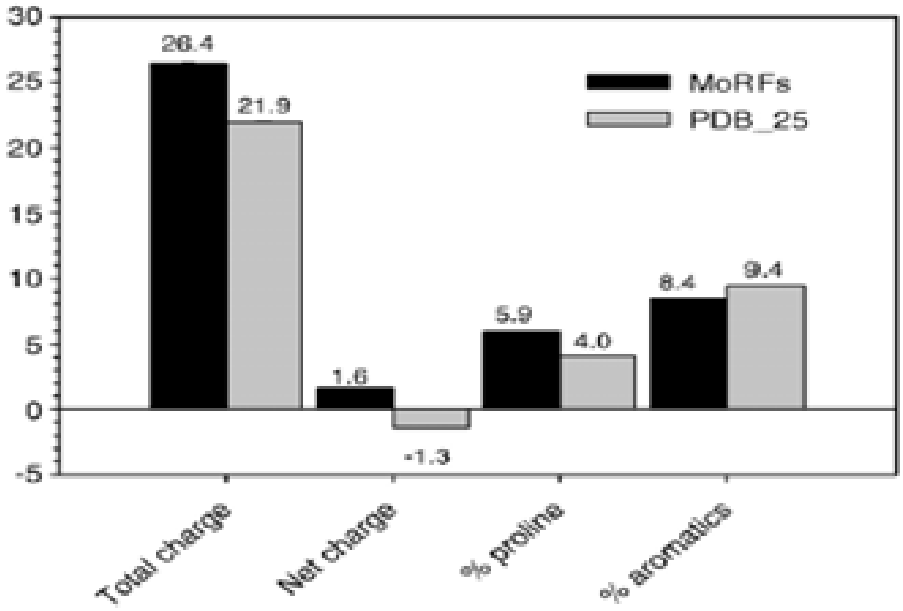

Figure 14: Total and net charge (calculated as charge per 100 residues) and the proportion of proline and aromatic amino acid residues in MoRFs and PDB_25. Error bars representing one standard error, calculated using 200 bootstrap iterations.

Predictions of Order-Disorder

Computational structure and sequence-based evaluations of order and disorder were performed to provide support for the idea that MoRFs are disordered in isolation and undergo a disorder-order transition upon binding to targets. Structure-based evaluations of disorder were performed using the criteria of Gunasekaran et al.,[158] who showed that the complexes of intrinsically disordered proteins have much larger interface and surface areas than those of ordered proteins. Sequence-based evaluations used prediction of disorder from sequence using both VLXT[62, 154, 159] and VL3.[160] The behavior of VLXT on MoRF containing proteins has been characterized on a small set of

validated MoRFs[44], whereas the behavior of VL3 has not been characterized in this respect.

Gunasekaran et al.,[158] have demonstrated that intrinsic disorder in the unbound state is reflected in structures of the bound state through relatively large surface and interface areas. A structural analysis of the bound structures of MoRFs in this data set was carried out using the previously characterized[158] OC data set as a negative control (Figure 15). Almost all MoRFs in the data set were above the order-disorder boundary suggested by Gunasekaran et al., which indicates that these regions are likely to be disordered in isolation, while all structured proteins were below this boundary, which indicates that these proteins are probably ordered in isolation. Only two of the β-MoRFs and one of the ι-MoRFs falls below the suggested boundary. This analysis should be viewed with some caution, since the data set used to derive the boundary was relatively small. Indeed, only a slight shift in the boundary would put all of the MoRFs above it. Thus, the boundary provides a strong indication that the MoRFs in this data set are indeed disordered in the absence of their binding partners and undergo a disorder-to-order transition upon complex formation. It should also be noted that disulfide bonds are not considered in this analysis, and so the indication that oxidized pseudo- MoRFs are disordered in the absence of their binding partners is likely to be in error. However, this analysis suggests that pseudo-MoRFs would probably be disordered in the absence of their binding partners and in the reduced state.

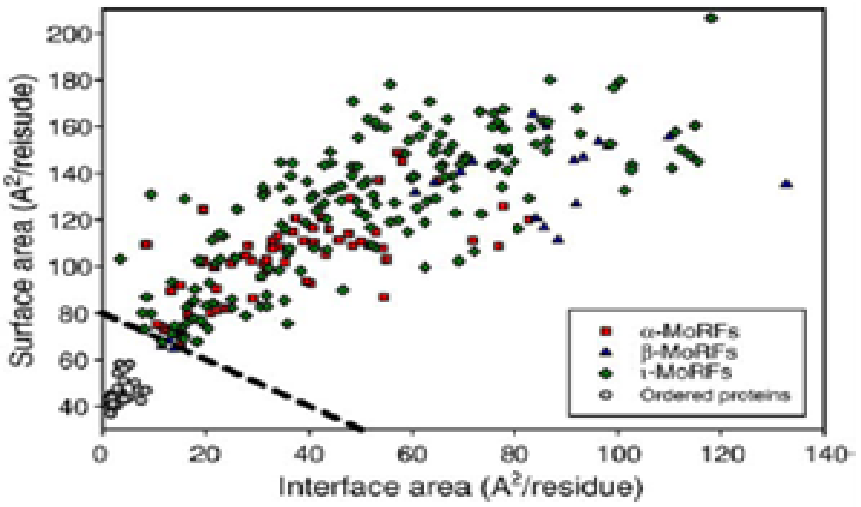

Figure 15: Surface and interface area normalized by the number of residues in each chain for the MoRF and the OC data sets.

Sequence based predictions of order/disorder, made with both the VLXT[62, 154, 159] and VL3[160] predictors, seem to contradict the structure based results. Specifically, predictions of disorder in MoRF regions (Figure 16(a)) suggest that, while many MoRFs are highly disordered, some MoRFs may be ordered. This is in part due to the large content of cysteine in these sequences, which is strongly correlated with prediction of order.[154] Also, it has been previously observed that disorder-to-order binding regions within larger disordered regions are often predicted to be ordered[44, 161] and our findings likely reflect these earlier observations.

The previously observed bias of disorder-to-order transition serves as a false indication of intrinsic order in many MoRF sequences. This bias is evident by the extreme behavior of disordered predictions for MoRFs (Figure 11(a)), where most MoRFs are predicted to be either highly disordered or highly ordered. Therefore, disorder

predictions were also examined for the entire sequences of proteins containing MoRFs and the sequence regions to the N and C sides of MoRFs in these sequences, in order to provide support for the idea that these regions occur in longer region of disorder. Disorder predictions for the full-length proteins that contain MoRFs (Figure 16(b)), relative to OM proteins (Figure 16(c)), suggest that many MoRF containing proteins are highly disordered. For the calculation of disorder in regions surrounding MoRFs, the fraction of residues predicted to be disordered was calculated over two windows of residues in the parent sequence of the MoRF, one on the C side and one on the N side of the MoRF. For ordered proteins, random sequence windows of equal size were taken from the OM data set. Similar to the entire sequence of proteins containing MoRFs, the sequence regions immediately surrounding MoRFs show a high content of predicted disordered residues, relative to OM proteins (Figure 17). This suggests that these MoRFs frequently occur in longer regions of predicted disorder.

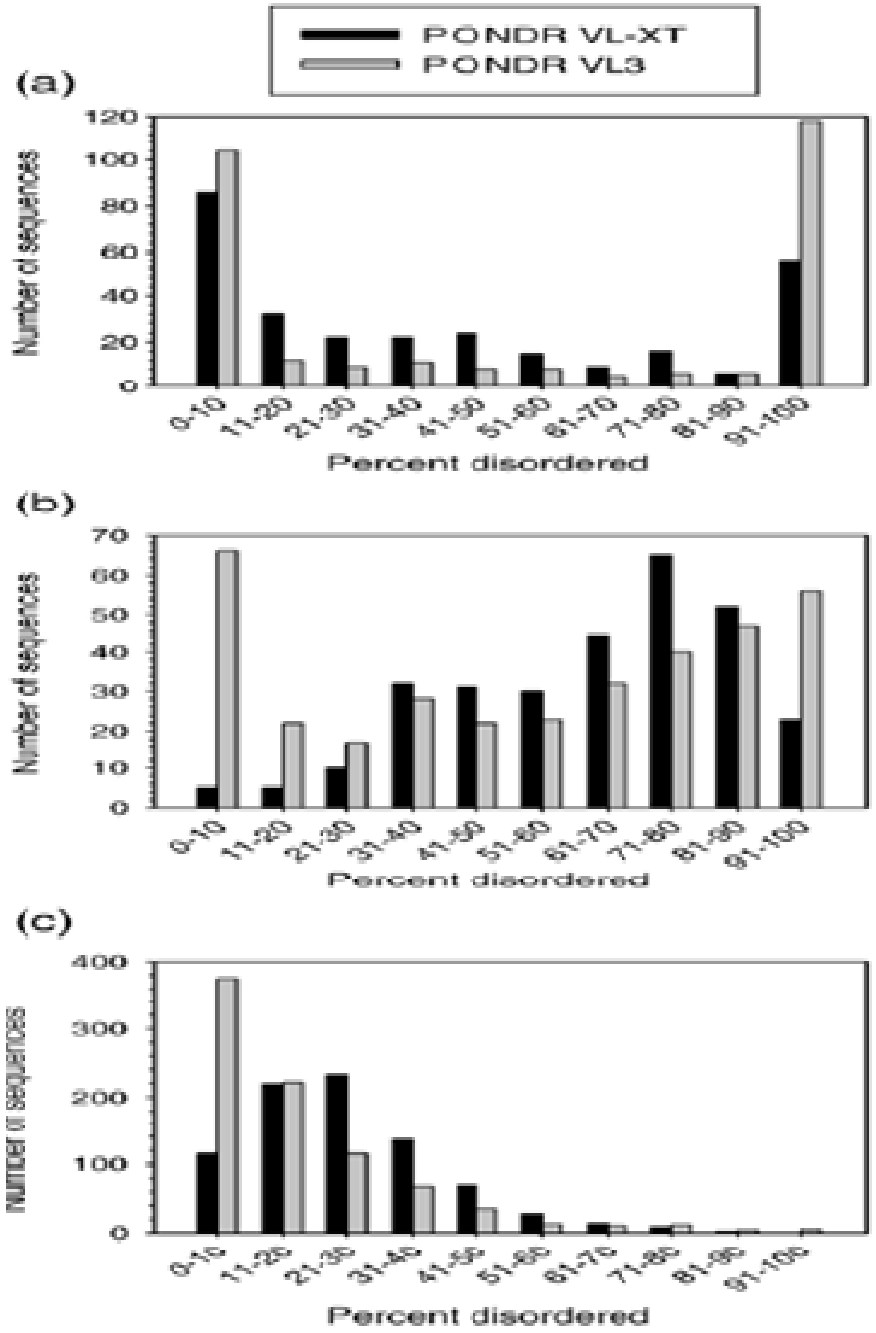

Figure 16: Disorder distribution in (a) MoRFs and (b) MoRF containing proteins and (c) OM proteins estimated by VLXT and VL3 predictors.

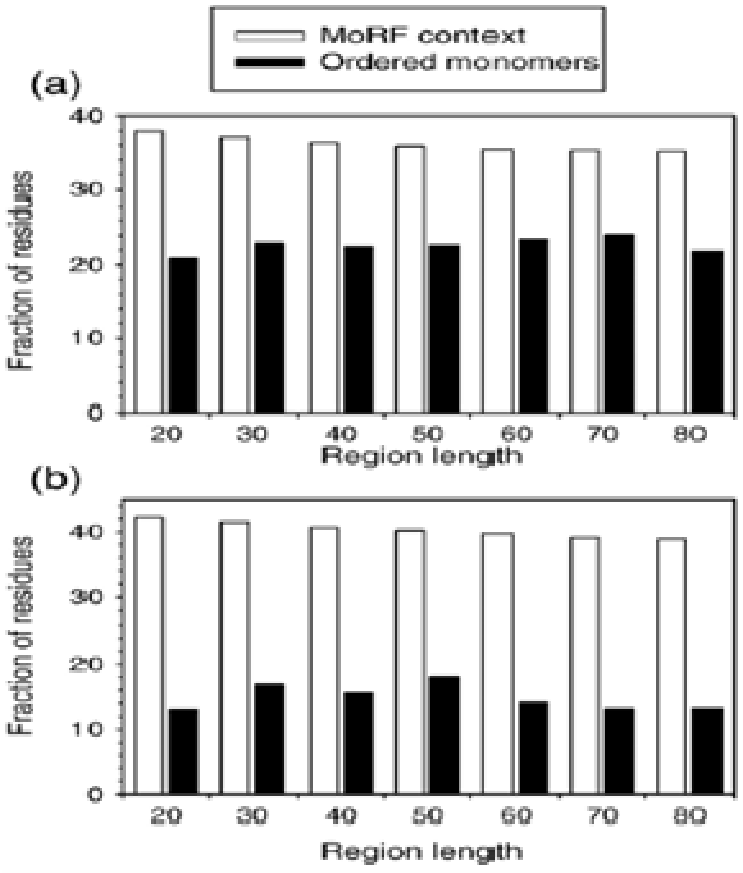

Figure 17: Fraction of residues predicted to be disordered for regions surrounding MoRFs and regions taken from ordered monomers using (a) VLXT and (b) VL3.

MoRF interface analysis

In another study[135] we investigated the specific properties of 62 α-, 20 β- and 176 ι-MoRF interfaces obtained using the Connolly surface algorithm.[139] Our results show that all MoRF interfaces (from MoRFs and MoRF-binding partners) are generally depleted in the six most exposed residues: N, D, Q, E, R and K and enriched in the six most highly buried residues: C, I, V, L, F and M. These trends are indicative of the propensity of these residues towards interaction. We also find that MoRF interfaces are

very different from the interfaces found using a non-redundant set of hetero-complexes earlier presented by Jones and Thornton.[136, 162] This suggests that MoRF interaction surfaces are distinct from those of other complexes.

Our investigations[135] on the predictability of different types of interaction surfaces using a combination of physicochemical properties and multiple geometric parameters such as accessible surface area, planarity etc applied to naïve Bayes classifiers revealed that it is possible to predict MoRF surfaces with balanced accuracies within the 84 – 94% interval. Surfaces on MoRF binding partners however are predictable with only 77 – 88% accuracy.

CHAPTER FIVE: APPLICABILITY OF DISORDER IN DETECTION OF POST-TRANSLATIONALLY MODIFIED PEPTIDES

Background

Until two decades ago most phosphorylation sites were identified by the application of standard knockout and/or mutation techniques to a residue in a protein of interest. The availability of mass spectrometry methods, eased access to high sensitivity means of phosphosite detection. To describe briefly, mass spectrometry (also referred to as MS) methods measure the mass-to-charge ratio (*m/z*) for peptides derived from an enzyme (e.g. trypsin) digested protein sample. Due to this method's ability to be able to identify component molecules at extremely low concentrations, it is often used to analyze organic compounds such as plasma and blood serum samples. An MS experiment typically consists of five parts (diagrammatically represented in Figure 18): sample introduction, ionization, mass analyses, ion detection (green box) and spectral data processing (blue box).

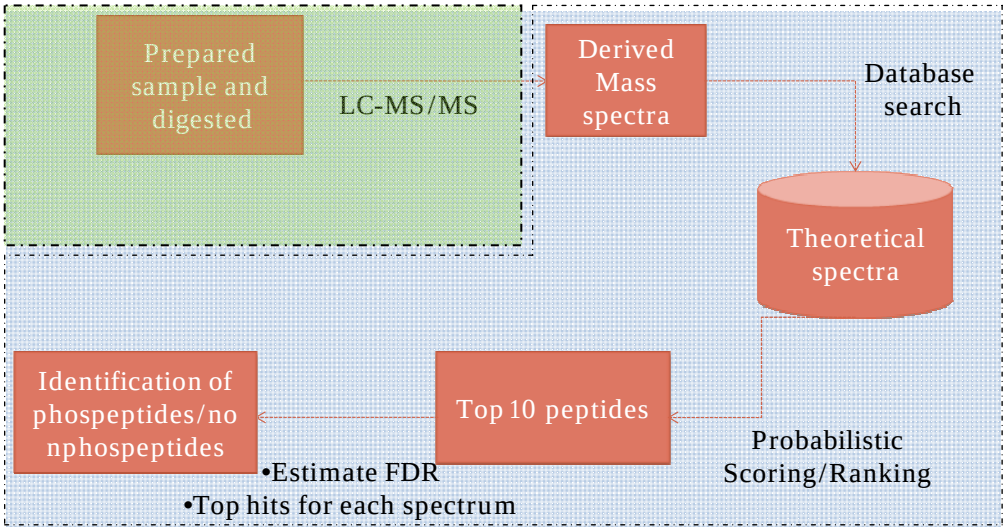

Figure 18: Standard mass spectrometry procedure for the identification of phospeptides

These days, a more accurate variant of MS, known as tandem mass spectrometry is being used for the analysis of post-translational modifications such as phosphorylation. Tandem mass spectrometry refers to multiple rounds of MS, also referred to as MS/MS (or MS^2) or MS/MS/MS (or MS^3). Due to the use of more than one mass analyzer, it is more specific nature and can identify the exact location of a modification on a peptide with better accuracy as compared to simple MS. Based on the type of mass analyzer used in MS, experiments can be customized. Occasionally the sample introduction step is preceded by an enrichment step to improve the chances of phospeptide identification by reducing the sample size while increasing the concentration of phospeptides in it. Some of the preferred methods for enrichment include immunoprecipitation with phospeptide specific antibodies, phosphoamidate chemistry and β-elimination.[163, 164] Depending on the type of mass analyzers used (e.g. time-of-flight (TOF), quadrapole time-of-flight (Q-TOF), orbitrap or linear ion trap) mass spectra with varying degrees of resolution can be acquired in large-scale proteomic studies. Standard outputs include a fragmentation pattern of peptides along with its corresponding molecular mass. Since post-translational modifications of proteins involve a change in their molecular masses, in addition to catering to the identification of the amino acid sequence of a protein, traditional MS methods have also served as a means to study and characterize PTMs in the past.[165]

Often, the combined use of liquid chromatography (LC) and tandem mass spectrometry is also applied in the process of phospeptide identification. This method is also popularly referred to as *shotgun proteomics*. In this approach, the sample is collectively digested into smaller peptides, selectively separated using liquid chromatography and then analyzed using standard mass spectrometry techniques. The

use of LC as a separation device in shotgun proteomics helps in increasing the number generated mass spectra. Each of the corresponding peptides or spectra is compared to a peptide database or theoretical spectra in order to determine the sequence composition of individual peptides. All identified peptides are finally used to identify their respective source proteins.

Database search algorithms used in phospeptide detection

Two of the leading search engines for database search in proteomics include Mascot[166] and Sequest.[167] Such tools strongly rely on the use of use of informal rules[168] or unified probabilistic models[166] to estimate the likelihood that a given spectrum was generated from its matching sequence returned by the database search. The highest scoring peptide corresponding to a given spectrum is assumed to be the best match for the spectrum. The success of all known database search algorithms depends on the assumption that the search database is complete and that the scoring algorithms used to draw peptide-spectrum matches are error-free. In other words, results of peptide identification are strongly influenced by the underlying scoring algorithms and database content employed during the search.[169] Although such methods are highly specific and serve as a promising technology in the study of cellular proteins especially those that are post-translationally modified, a typical experiment can identify only about 20 – 30% of the originally obtained spectra and can confidently map only a few peptides per protein. Besides this, the computational requirements for phospeptide identification continue to remain substantial, especially when peptides contain more than one modification. This

experience is further intensified when considering more than one protein modification. With high-confidence phospeptide detection as the underpinning of all phosphoproteomic initiatives, there is a pressing need to explore faster, more rigorous methods of phospeptide identification, both in terms of the number of identified peptides and their quantity in the sample.

One way to alleviate this search problem would be to reduce the number of non-phospeptide spectra and retain only those spectra that contain phosphorylation site information. This can be treated as a spectral filtering approach to improve the chances of identifying phospeptides by eliminating background noise before a database search. A recent article,[170] suggested the use of support vector machines to screen tandem-mass spectra with the goal of improving the chances of detecting phospeptide spectra. Although this method can effectively search 80% of the available tandem-mass spectra from rat brain to identify 95% of the total phospeptide spectra, we find that it has limited portability and expansion to non-+2 charge data sets. Another method to reduce the peptide search space can be to systematically filter all unlikely phospeptide candidates from the protein database. By doing so, comparisons with peptides that are unlikely to be identified in a phosphorylated form can be avoided, thereby significantly reducing the time required to perform a search for phospeptides. To this effect, we propose a new algorithm that can systematically filter peptide-spectrum matches thereby reducing the searched protein database used to identify phospeptides in MS/MS. Our proposed methodology exploits previously established concepts such as peptide detectability[171] and the fact that phosphorylation sites are closely correlated to the degree of intrinsic disorder in the parent protein. Peptide detectability is defined as the probability of detecting a

peptide given that its parent protein exists at standard quantity.[172] It has been shown that predicted peptide detectability can be successfully used in protein inference.[171] Our proposed algorithm is a supervised learning based one which can be used to estimate the probability of detecting a phospeptide in MS/MS and later towards prioritizing a database of peptides used for phospeptide search.

To summarize, while high-throughput mass spectrometers can capably generate a large number of spectra, the algorithms used to search and identify phospeptides from these spectra are comparably slower and have low sensitivity. This not only impedes the overall goal of improving our existing repositories of phosphorylation but also hampers our complete understanding of the role of phosphorylation in cells. Here we propose the development of a new, faster algorithm to search for phospeptides that uses the peptide detectability concept and scores from a disorder-based predictor of phosphorylation sites to learn and predict MS/MS identified phospeptides.

Materials and Methods

This analysis makes use of mouse liver tissue sample as provided by Quanhu Sheng at the Shanghai Institutes for Biological Sciences. Protein phosphorylation has been reported to play a crucial role in normal liver development and function.[173] Previously, sites determined using liver tissue have been reported to have assisted in increasing our understanding of phosphorylation-related liver conditions such as those related to aberrant glucose and lipids metabolism.[174]

Sample preparation

A linear ion trap/Orbitrap (LTQ-Orbitrap) hybrid mass spectrometer (ThermoFinnigan, San Jose, CA, USA) equipped with an NSI nanospray source was operated in data dependent mode to automatically switch between MS and MS/MS acquisition with ion transfer capillary of 200 °C and NSI voltage of 1.85 kV. Normalized collision energy was set at 35.0. According to the different detectors (Orbitrap and LTQ) used for the MS scan, two surveys of full scan mass spectra acquired modes were used to obtain final spectral data. The mass spectrometer was set such that, one full MS scan was acquired in the Orbitrap parallel to (or following) ten MS/MS scans in the linear ion trap on the ten most intense ions from the full MS spectrum with the following Dynamic Exclusion™ settings: repeat count 2, repeat duration 30 seconds, exclusion duration 90 seconds. The resolving power of the Orbitrap mass analyzer was set at 100,000 (m/Δm50% at m/z 400) for the precursor ion scans. To establish a benchmark for the number of phospeptides and non-phospeptides identified using this data set, all resultant spectra were searched using Mascot. Forward database was constructed using sequences corresponding to the source proteins returned for every peptide-spectrum match. By reversing each of the sequences from the forward database we generated our decoy database.

Sequence data sets

Our initial data set S contained 1,290,314 peptide-spectrum (230,340 unique spectra) pairs as returned by Mascot. This data set comprised of eight subsets, S_{+1}, S_{+2},

S_{+3}…S_{+8} based on the charge of a peptide from each peptide-spectrum pair. Individual statistics for each of the subsets have been provided in Table 8.

Data set	Number of peptide-spectrum matches
S_{+1}	21,305
S_{+2}	601,714
S_{+3}	647,129
S_{+4}	18,535
S_{+5}	1,463
S_{+6}	128
S_{+8}	40

Table 8: Distribution of peptide-spectrum matches returned by Mascot.

Since the majority of the peptides in *S* were either +2 or +3, all further analyses were restricted to S_{+2} and S_{+3}. S_{+2} and S_{+3} data sets were further split into two sets: $D_{STY\text{-}P}$ or the set of all peptide-spectrum matches containing at least one identified phosphorylation site and, $D_{STY\text{-}NP}$ consisting of all peptide-spectrum matches with at least one serine, threonine or tyrosine in the matched peptide and without any observed phosphorylation site. A standard Mascot search was performed on each data set $D_{STY\text{-}P}$ and $D_{STY\text{-}NP}$ to identify phospeptides and non-phospeptides at 1% and 5% false discovery rates. DisPhos[114] predictions were made on all proteins identified as the source of peptides belonging to S_{+2} and S_{+3}. This included both forward and reverse sequences. Mean and maximum DisPhos scores over all sites for each peptide from S_{+2} and S_{+3} were computed. Peptide detectability predictions were also made for each peptide in both data sets using an in-house detectability predictor.

Training data sets (+2 and +3) for phospeptide predictor were generated as follows. All peptides from S_{+2} with a modified serine, threonine or tyrosine and a mascot score $\geq$ 25 were considered in the construction of the positive data set. Remaining peptides from S_{+2} with at least one serine, threonine or tyrosine were considered in the construction of the negative data set. Finally, positive and negative data sets were filtered such that only unique peptides were used to train +2-phospredictor. A similar procedure using S_{+3} was followed to prepare positive and negative data sets for a +3 phospredictor however with a mascot score of 28 instead of 25. The choice of Mascot scores used to construct training data sets for +2 and +3 phospeptide predictors was based on scores returned from a standard Mascot search for phospeptides at 5% false discovery rate. All peptides common between either of the positive and negative data sets were retained only in the positive data set and eliminated from the negative data set.

Predictor development and evaluation

Our initial three predictors were constructed using data sets D_f^{+i} where $i \in \{+2, +3\}$ and feature $f \in \{disphos, peptide\ detectability, disphos \cup peptide\ detectavility\}$ data sets for each of these predictors were generated using five basic sequence features including length of a peptide, number of serines, threonines and tyrosines in a peptide and peptide mass. Additional features in all of the data sets included mean and maximum DisPhos score, peptide detectability, and mean DisPhos, maximum DisPhos in conjunction with peptide detectability scores, respectively. A control data set $D_{spectral}^{+i}$ was assembled using

only spectral features as described in Lu et al.,[170] These features included: number of peaks corresponding to neutral loss of 98 pairs, precursor "neutral loss of 98/base peak" intensity ratio, number of peaks corresponding to neutral loss of 49 pairs, number of peaks corresponding to neutral loss of 80 pairs and precursor "neutral loss of 49/base peak" intensity ratio. The purpose of the control data set was to measure our predictor's performance in comparison to a spectral-feature based predictor previously suggested to screen collision-induced dissociated tandem mass spectra prior to a database search for phospeptides.[170] The mean and standard deviation corresponding to each of these features were calculated for positive and negative data sets from both data sets. These statistics have been reported in Table 9.

		Charge							
	Feature	+2				+3			
		Phospeptides		Nonphospeptides		Phospeptides		Nonphospeptides	
	Sequence	mean	std	mean	std	mean	std	mean	std
1	Average disphos score	0.78	0.20	0.48	0.26	0.79	0.16	0.44	0.25
2	Max disphos score	0.85	0.19	0.56	0.27	0.88	0.10	0.54	0.27
3	Peptide detectability	0.60	0.39	0.56	0.40	0.53	0.40	0.54	0.40
4	Number of serines	2.39	1.46	1.16	1.16	2.55	1.62	1.44	1.38
5	Number of threonines	0.75	0.92	0.86	0.92	1.05	1.02	1.02	1.03
6	Number of tyrosines	0.27	0.61	0.45	0.71	0.31	0.63	0.51	0.74
7	Peptide mass	1836.22	413.48	1734.04	495.89	2545.86	562.39	2317.32	572.61
8	Length of peptide	16.92	4.01	15.56	4.54	23.59	5.87	20.86	5.50
	Spectral								
1	Peak-NL pairs(98)	2.00	0.00	2.00	0.00	3.00	0.00	3.00	0.00
2	NL/Base peak intensity ratio	8.69	2.98	5.63	2.33	3.99	2.60	4.95	2.46
3	Peak-NL pairs/+2	0.55	0.43	0.06	0.11	0.33	0.38	0.04	0.09
4	Peak-NL pairs(80)	2.37	1.86	1.27	1.28	8.44	3.65	4.55	2.43
5	H2OLoss/Base peak ratio	0.68	0.91	1.74	1.34	0.58	0.88	1.53	1.26
6	Percentage of ions with intensities above 1%	0.01	0.05	0.05	0.13	0.00	0.02	0.01	0.05
7	Intensity difference between the highest and second highest peaks	0.03	0.05	0.02	0.02	0.04	0.07	0.04	0.07

Table 9: Table of statistics for features used for training the positive and negative data sets.

Prior to training our predictors, we used paired t-tests to select the most significant features in each data set using a p-value threshold of 0.1. All selected features were normalized using the z-score approach before performing a principal component analysis (with 95% of retained variance) in order to further reduce the dimensionality and internal correlation within data sets.

SVM^{light} software[175] was used to predict phosphorylated peptides. We evaluated a linear and non-linear kernel, where non-linear kernel was gaussian radial basis ($\sigma = 10^{-4}$). The default value was used for capacity *c* for all experiments. 10-fold cross validation was applied on D_f^{+i} (training $=90\% * |D_f^{+i}|$, test $=10\% * |D_f^{+i}|$)) to evaluate prediction accuracy. Sensitivity (sn), specificity (sp), balanced-sample accuracy acc = ½ · (sn + sp), and area under the ROC curve (AUC) was estimated to evaluate each predictor's overall performance. Sensitivity is defined as the prediction accuracy on the phospeptides and specificity corresponds to the prediction accuracy on non-phospeptides. As sensitivities achieved by the linear kernel predictor were higher in comparison to those obtained using Gaussian kernel we decided to use the linear kernel predictor for all subsequent analyses.

Filtering peptide-spectrum matches before database search

For each filtering experiment, all peptides in data set S_{+i}, $i \in \{+2, +3\}$ were sorted based on feature f - predictor scores where, $f \in \{disphos, peptide\ detectability, disphos \cup peptide\ detectavility\}$. Standard Mascot search for phospeptides and non-phospeptides was repeated using all data sets after eliminating candidate phospeptides from bottom j^{th}

percentile $j \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. To ensure correctness of the algorithm we specifically included a final checkpoint step of eliminating all unlikely phospeptides from lowest 100^{th} percentile. A schematic representation of this algorithm has been provided in the following flowchart.

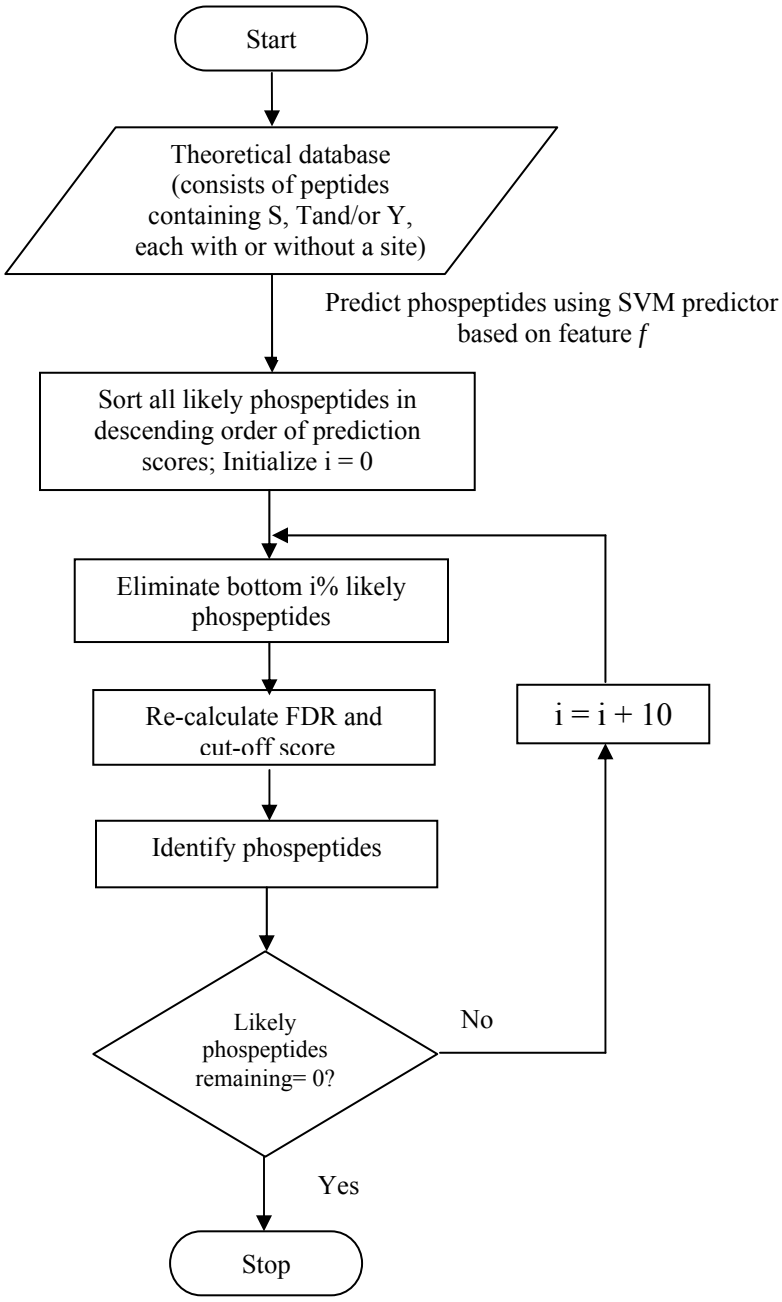

Figure 19: Flowchart of a novel algorithm for searching phospeptides in tandem mass-spectrometry

Results

Identification of phospeptides and non-phospeptides

By applying a standard Mascot search at 5% on data set D_{+2} we found 20,747 peptides including 2,754 phospeptides and 17,993 non-phospeptides. The total number of unique phospeptides and non-phospeptides was 498 and 3,726. A similar search using data set D_{+3} found 4,543 peptides including 212 phospeptides and 4,331 non-phospeptides. The total number of unique phospeptides and non-phospeptides was 78 and 1,525. Searches for identified phospeptides and non-phospeptides were also repeated at 1% FDR. Table 10 presents a summary of results from this exercise.

		Phospeptides			Nonphospeptides		Total peptides
FDR	Charge	Total	Unique	# Sites	Total	Unique	
1%	+2	2,219	387	2,524	15,476	3,150	17,695
	+3	138	41	153	3,772	1,278	3,910
5%	+2	2,754	498	3,252	17,993	3,726	20,747
	+3	212	78	239	4,331	1,525	4,543

Table 10: Identified phospeptides and non-phospeptides at 1%FDR and 5%FDR.

Peptide detectability and mean DisPhos score distribution

Figure 20 shows box plots of the peptide detectability (Figure 20A) and mean DisPhos (Figure 20B) scores corresponding to all non-redundant identified phospeptides, identified non-phospeptides and all unidentified peptides in data sets set D_{+2} and D_{+3} determined using a Mascot search. As seen in the figure, on an average, identified phospeptides have higher DisPhos scores in comparison to identified non-phospeptides as

well as unidentified peptides. This finding supports previous conclusions suggesting the incidence of phosphorylation sites within regions of intrinsic disorder in proteins. This figure also shows that identified phospeptides as well as identified non-phospeptides have much higher peptide detectability scores in comparison to unidentified peptides. These results clearly suggest that a disorder-based predictor of phosphorylation sites can be used to computationally learn the predictability of detectable and identified phospeptides in LC-MS/MS experiments. Additionally, peptide detectability also correlates well with identified peptides (both, phospeptides and non-phospeptides) and can be used to serve as a feature in predicting peptides likely to be determined via LC-MS/MS experiments.

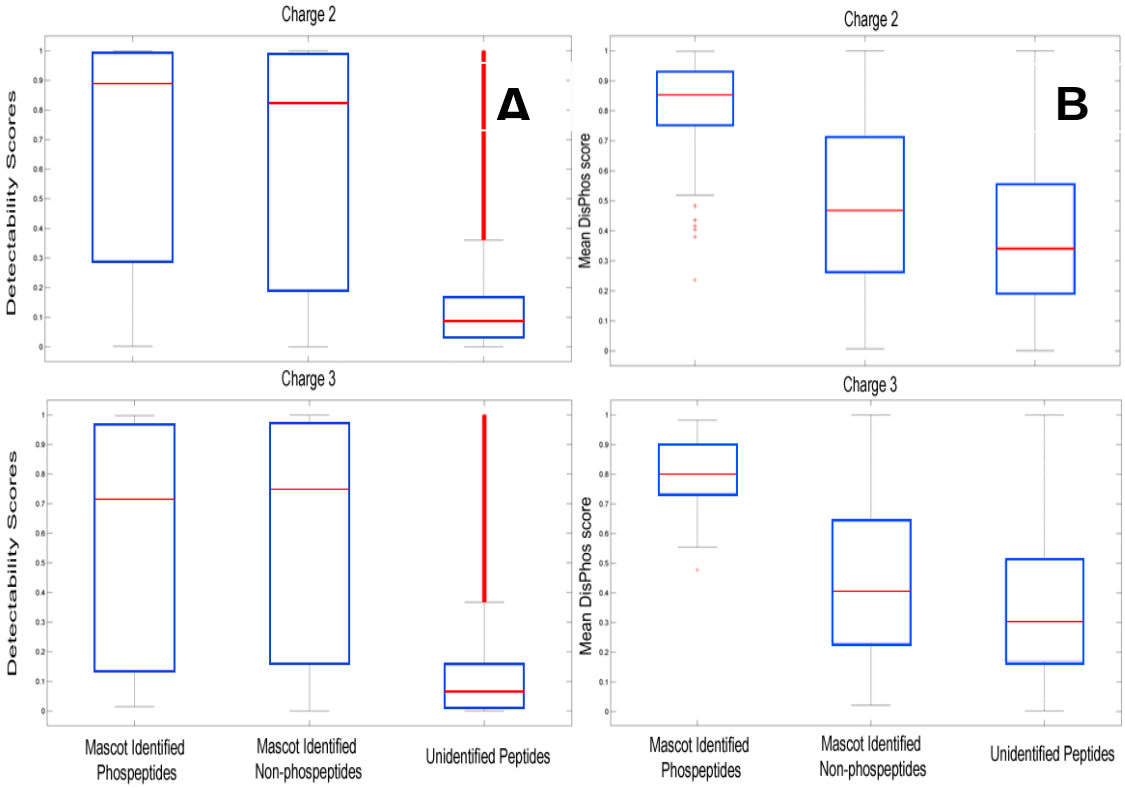

Figure 20: Boxplots depicting (A) peptide detectability distribution for +2 (top) and +3 (bottom) (B) mean DisPhos score distribution for +2 (top) and +3 (bottom) in identified phospeptides (left), identified non-phospeptides (center) and unidentified peptides (right).

Prediction of phospeptides

Our predictor trained on sequence features (as described above in Materials and Methods section) achieved 72% and 84% accuracy on +2 and +3 phospeptides respectively. Both predictors reached sensitivities of 62% and 73%. It is important to note that given the disproportionate sizes of positive and negative data sets we considered comparing sensitivity and the area under the receiver operator characteristic (ROC) curve achieved by this predictor and not accuracies alone. Our results show that by using simple sequence relevant features such as peptide detectability, mean DISPHOS scores, twenty basic amino acid compositions, serine, threonine and tyrosine counts, peptide mass and peptide length, it is possible to discriminate identified phospeptides from other peptides with high accuracy.

Since, to the best of our knowledge, no other sequence based predictors of phospeptides was available at the time of this study we compared the performance of our predictor with a previously developed SVM predictor, Colander, which makes use of spectral features in predicting phospeptide spectra.[170] We recreated Colander to compare its performance with our sequence-based predictor in terms of the phospeptide prediction accuracy. Accuracy, sensitivity and AUC for predictors constructed using sequence and spectral features have been presented in Table 11. As seen in Table 11, lower prediction scores were obtained using spectral features alone suggesting that while spectral features can be used to predict phospeptide spectra (and thereby phospeptides) they may not be as generalizable across MS data sets as sequence features may be.

Predictor	Charge	Accuracy (%)	AUC (%)	Sn (%)
$D_{disphos\ u\ detectability\ u\ AA}$	+2	73	78	62
	+3	81	87	83
$D_{spectral}$	+2	70	76	52
	+3	73	79	71

Table 11: Prediction accuracy, AUC and sensitivity for sequence feature based and spectral feature based predictors.

Feature analysis

A paired t-test was performed to compare the means of all attributes used in the construction of each of the five predictors. Among the top 10 significant attributes (p-value < 0.05) for $D_{sequence}^{+2}$ and $D_{sequence}^{+3}$though with varying degrees of contribution were peptide standard detectability, mean DisPhos score for a peptide, maximum DisPhos score for a peptide, number of threonines, number of tyrosines and amino acid compositions corresponding to aspartic acid (D), glutamic acid (E) , threonine (T), cysteine (C) and phenylalanine (F).

A novel algorithm for improved detection of phospeptides in LC-MS/MS

We observe that by eliminating 80% of the low scoring + 2-phospeptides at 1% false discovery rate we are able to improve the total count of +2-unique phospeptides identified by 18% and their corresponding phosphosites by 31%. Similarly, we find that by removing 80% of the poorly scoring + 3-phospeptides we gain nearly 44% more

unique + 3-phospeptides and as many as 23% more phosphorylation site. At 5%FDR, the number of +2 and + 3-phospeptides gained after eliminating 80% of poorly ranked phospeptides is slightly lower (13% and 22% respectively). The number of non-phospeptides identified remained unaffected by the process of elimination of less likely phospeptides for both data sets. To validate the correctness of our results we also calculated the overlap between phospeptides identified by a standard Mascot search and those identified by our method at each percentile step (as described in the Materials and Methods section). We find that by using only the top 20% scoring phospeptides our method was able to detect 97% of the phospeptides identified by the standard Mascot search.

Detailed results from this experiment, including the number of phospeptides and phosphorylation sites identified at 1% and 5% FDR have been illustrated in Figure 21 (top and bottom).

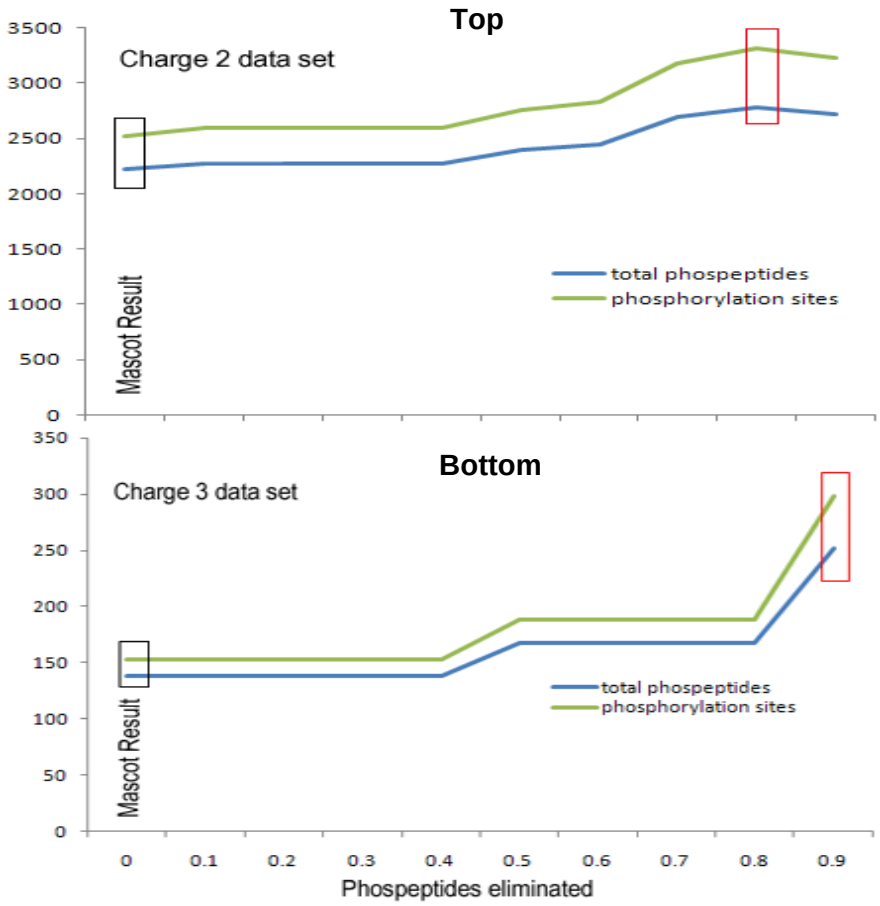

Figure 21:Number of phospeptides identified at (Top) 1%FDR and (Bottom) 5% FDR, after eliminating phospeptides from bottom 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% peptides

Our results show that our proposed method is not only able to search phospeptides by using a reduced protein database but can also help in identifying more phospeptides and phosphorylation sites compared to a Mascot search. The observation is consistent for both, +2 and +3 data sets.

Most functions in the methodologies described here and previously have been designed using Matlab v7.8 (R2009a) with the exception of a few scripts that were written in Perl or bash shell.

CHAPTER SIX: EVALUATING DISORDER IN PATHOGENIC APICOMPLEXANS VERSUS NON-PATHOGENS

Background

Pathogenic apicomplexans and their diseases

Parasitic protozoa belong to the phylum *apicomplexa* that have a significant impact on humans. This phylum includes anaerobic organisms such as *Giardia lamblia* and *Entamoeba histolytica*. *G. lamblia,* a diplomonad, is one of the most common intestinal protozoans that cause diarrhoea.[176] *E. histolytica*, causing colitis and liver abscesses, is the second leading cause of death from parasitic diseases in the world, killing up to 100,000 people a year.[177] The kinetoplastids include *Trypanosoma brucei* and *Trypanosoma cruzi*, which are the causative agents of African sleeping sickness and Chagas' disease respectively. An estimated 18 million persons are infected with *T. cruzi* in Latin America[178] and 300 000–500 000 cases of African sleeping sickness occur per year. Currently, there are very few treatment regimes available for the *Trypanosoma* species, some of which are highly toxic.

Also included in this group are the *Plasmodium* species (the primary causative agent of malaria), the *Cryptosporidium* species, (causative agent of an intestinal infection leading to substantial water-borne outbreaks resulting in serious strains on agricultural and medical resources[179]) and *Toxoplasma gondii,* (implicated in congenital birth defects and with known links to neurological disorders and behavioral anomalies in humans.)[180-182] Malaria is one of the most catastrophic infectious diseases of our times, having infected nearly 500 million people in 2002 and resulting in at least 1 million casualties, a

majority of which were children.[183, 184] *Cryptosporidium* and *Toxoplasma* are also known to cause fatal infections in immunocompromised individuals (as in the case of AIDS), [185] and have therefore been categorized by the National Institute of Allergy and Infectious Disease (NIAID) as *category B pathogens* that are relevant to bio-defense research.[186]

With genome sequences now readily accessible for a number of these protozoa, we now have the unique opportunity to explore phenomena that may help us in understanding the basic cell biology within pathogens and explain why some apicomplexan protozoan organisms cause disease while others do not. This, in turn, may ultimately lead us to the discovery and development of novel and effective therapies to manage diseases caused by pathogens. Having said that, there are several complications that hamper functional genomic studies in protozoan parasites.[187] For example, most protozoal genomes contain a high number of genes that lack reasonably confident orthologues in other organisms.[188]

Abundance of intrinsic disorder in *P.falciparum*

Several computational studies have estimated the abundance of intrinsic disorder in proteins of *P. falciparum*. For example, it has been reported that at least 35% of proteins encoded by genes on chromosomes 2 and 3 in this pathogen are predicted to contain long (up to 40 consecutive residues) disordered regions.[25] A later study claimed that this number was in fact an underestimate, and nearly 52–67% of the proteins are predicted to contain long disordered regions.[31] Furthermore, it was shown that, proteins expressed in the sporozoites of *P. falciparum* were more intrinsically disordered

compared to those expressed during other life cycle stages.[34]

Each of the above findings naturally pushes us into asking a few questions: Are low-complexity regions common amongst pathogenic organisms? Does the presence of low-complexity regions (and therefore likely disordered regions) have a role in the pathogenic behavior of such organisms? Does intrinsic disorder lend a functional advantage to apicomplexan pathogens? To attempt an answer to these, we propose a study that closely analyzes the disorder content in multiple pathogenic proteomes and compare it to those from non-pathogenic organisms. In the following pages, we describe methods to investigate the compositional, motif and charge-hydropathy preferences within proteins from pathogens as well as non-pathogens. We also compare the results from both these groups with a model eukaryote and a prokaryote.

Materials and Methods

Sequence data sets

Various online databases were used as sources for annotated genomes corresponding to the following species: *Plasmodium falciparum* (excluding mitochondrial and plastid proteins), *P. berghei*, *P. chabaudi*, *P. vivax*, *P. yoelii* (Release 3.4),[189, 190] *Toxoplasma gondii* (Release 4.1),[191] *Theileria parva* (http://www.tigr.org/),[192] *Cryptosporidium hominis* and *Cryptosporidium parvum* (http://cryptodb.org/cryptodb/),[193, 194] *Candida albicans* and *Candida glabrata*,[195] *Entamoeba histolytica* (http://www.tigr.org/), *Giardia lamblia*[196] and *Trypanosoma brucei* (http://www.tigr.org/). In addition to these, annotated data corresponding to the

non-pathogenic free-living protozoan *Tetrahymena thermophila* (http://www.tigr.org/), the slime mold *Dictyostelium discoideum* (http://dictybase.org/) and the yeast *Saccharomyces cerevisiae* (http://www.yeastgenome.org/) were obtained to serve as control organisms. *Caenorhabditis elegans* (http://www.wormbase.org/) and *Vibrio cholerae* (http://www.tigr.org/) were used as models for a multicellular eukaryote and prokaryote, respectively. All occurrences of ambiguous residues such as B, X, or Z in the data sets were replaced by alanine, due to its neutrality to order as well as disorder. The total numbers of sequences, mean sequence lengths, and number of ambiguous residues for each working data set have been summarized in Table 12.

Organism	Number of annotated sequences available and used	Average sequence length (rounded to the nearest integer)	Number ambiguous residues (replaced by alanine)
C. parvum	3806	597	355
C. hominis	3886	452	318
P. falciparum	5411	751	71
P. berghei	12235	245	2029
P. chabaudi	15007	194	2223
P. vivax	5352	682	368
P. yoelii	7861	433	8689
T. parva	4079	465	4
T. gondii	7793	720	10287
E. histolytica	9766	389	0
G. lamblia	9646	351	0
T. brucei	8758	502	8779
C. albicans	6125	479	128
C. glabrata	5271	502	204
D. discoideum	4032	668	4124
T. thermophila	27424	605	60746
S. cerevisiae	11081	435	1316
C. elegans	38398	465	718
V. cholerae	3887	299	20

Table 12: Summary of number of sequences, mean sequence length, and ambiguous residues in each of the 19 proteomes.

The choice of these organisms relies on their known pathogenicity in mammalian organisms as described here. *P. falciparum* causes the most dangerous form of malaria in

humans. *P. vivax* is the most frequent and widely distributed cause of recurring malaria though benign, in humans. *P. berghei*, *P. chabaudi*, and *P. yoelii* are three of the four malaria parasites of African murine rodents. *T. gondii* causes toxoplasmosis in warm-blooded vertebrates. *T. parva* is the causative agent of East Coast Fever (ECF), an acute, tick-borne disease causing high rates of morbidity and mortality in cattle. *Cryptosporidium* species cause diarrhoeal illness. *C. albicans* is a diploid fungus (a form of yeast) capable of causing opportunistic oral and genital infections in humans. *C. glabrata* is now recognized as a highly opportunistic pathogen of the urogenital tract as well as of the bloodstream in immunocompromised individuals. *E. histolytica* and *G. lamblia* are anaerobic protozoan parasites that infect the gastrointestinal tract. *T. brucei* is a parasitic protist that causes African trypanosomiasis (sleeping sickness) in humans and animals. *D. discoideum* (also known as slime mold) a non-pathogen, is a soil-living amoeba that exists in uni- and multi-cellular forms. *T. thermophila* is a non-pathogenic free-living ciliated protozoan. *Saccharomyces cerevisiae* is a species of the budding yeast. *C. elegans* is a freeliving nematode. *V. cholerae* is a gram negative bacterium that causes cholera in humans.

Compositional profiling

To gain an insight into the relationships between sequence and disorder, amino acid compositions from different data sets were compared using an approach recently developed for intrinsically disordered proteins.[197] To this end, the fractional difference in composition between a given set of proteins and a set of reference proteins (either a set of

ordered proteins[154], disordered proteins from DisProt database,[198] or proteins from *Tetrahymena thermophila*, *Caenorhabditis elegans* or *Vibrio cholerae*) was calculated for each amino acid residue. The fractional difference was calculated as,

$$\frac{C_x - C_{reference}}{C_{reference}}$$

Equation 7: Fractional amino-acid compositions for proteins from apicomplexan pathogens and non-pathogens.

where, C_X is the content of a given amino acid in a given protein (or protein set), and $C_{reference}$ is the corresponding content in a set of reference proteins and plotted for each amino acid. In corresponding plots, the amino acids were arranged from the most rigid to the most flexible according to the Vihinen's flexibility scale, which is based on the averaged B-factor values for the backbone atoms of each residue type as estimated from 92 proteins.[199]

Predictions of intrinsic disorder

Disorder predictions for proteins corresponding to each of the above listed organisms were made using VLXT[154, 159] and VSL2B.[91] Cumulative distribution function (CDF) curves[200] were generated for each data set using VLXT scores for each of the 19 organisms. CDF analysis discriminates between order and disorder by means of a boundary value. This value can be interpreted as a measure of proportion of residues with low and high disorder predictions. Additionally, charge-hydropathy (CH) distributions were also analyzed for these organisms using methods as described in Uversky et al., [156]

Predicting alpha-MoRFs

The predictor of α-helix forming Molecular Recognition Features, α-MoRF, is based on observations that predictions of order in otherwise highly disordered proteins corresponds to protein regions that mediate interaction with other proteins or DNA. This predictor focuses on short binding regions within long regions of disorder that are likely to form helical structure upon binding.[27] It uses a stacked architecture, where VLXT is used to identify short predictions of order within long predictions of disorder, and then a second level predictor determines whether the order prediction is likely to be a binding site based on attributes of both the predicted ordered region and the predicted surrounding disordered region. An α-MoRF prediction indicates the presence of a relatively short (20 residues), loosely-structured helical region within a largely disordered sequence.[27] Such regions gain functionality upon a disorder-to-order transition induced by binding to partner.[107, 135]

We also made use of a protein–protein interaction map from *P. falciparum* published recently.[201] This map contains 19,979 interactions involving 2,321 proteins. This map was generated using logistic regression methods to interpret protein–protein interactions involved in conserved protein interactions, their underlying domain interactions and listed supplemental experimental data.[201] Our goal for working with this map was to compare the connectivity of *P. falciparum* proteins (i.e. how many interactions a given protein participates in) and their extent of intrinsic disorder.

Results

Amino acid composition profiles

Amino acid compositions of fifteen early-branching eukaryotic organisms (the nine apicomplexa *P. falciparum, P. berghei, P. chabaudi, P. vivax, P. yoelii, T. gondii, T. parva, C. hominis*, and *C. parvum*, *C. albicans* and *C. glabrata*, *T. brucei*, amoebozoa *D. discoideum* and *E. histolytica*, and *G. lamblia*) were compared with compositions of proteins from a representative disordered (Figure 22A) and ordered (Figure 22B) data set. Compositions of all organisms except *D.Discoideum* were compared to the freeliving non-pathogenic protozoan *T. thermophila* (Figure 23A), *D. Discoideum* (Figure 23B) and *S. cerevisiae* (Figure 23C). In addition to these, profiles have been plotted in comparison to *C. elegans* (Figure 24) and *V. cholerae* (Figure 25).

These figures depict fractional relative compositions, with the amino acids arranged from left to right in increasing order of surface accessibility in globular proteins (also known as the Vihinen flexibility scale). Several trends emerge in these figures. For instance, parasitic protozoan data sets are significantly depleted in tryptophan (W) and enriched in lysine (K), in comparison to ordered sequences (Figure 22A). However, in comparison to the disordered data set, most of the protein sets are depleted in tryptophan (W) (Figure 22B). It is interesting to note that in comparison to ordered as well as disordered sequences, majority of these fourteen organisms are enriched in phenylalanine (F), isoleucine (I) and tyrosine (Y). Comparison of Figure 22A and B suggests that early-branching eukaryotes represent a unique group whose proteomes are compositionally different when compared to typical ordered and disordered proteins.

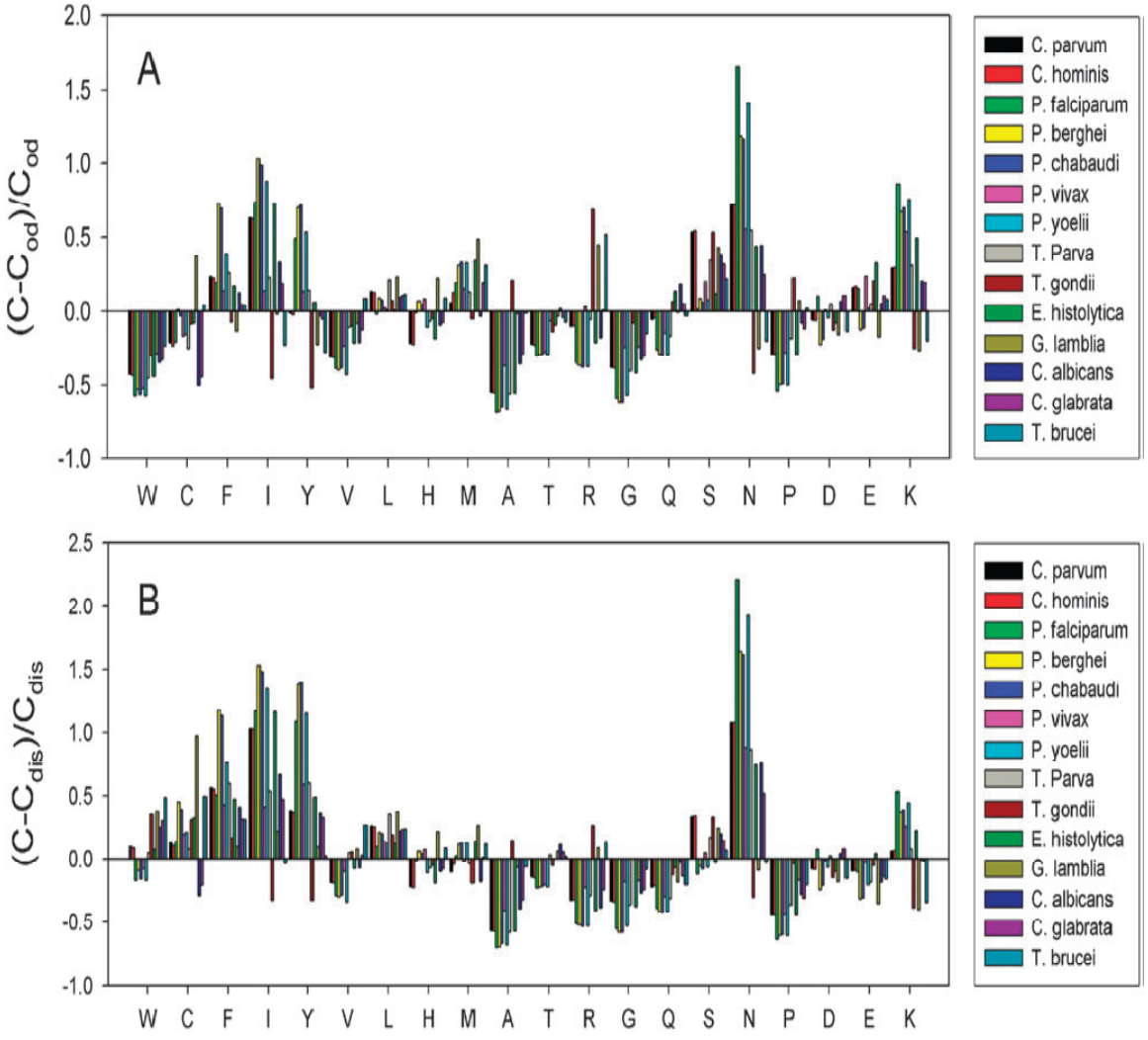

Figure 22: Compositional profiling of early-branching eukaryotes in comparison with a (A) set of ordered and (B) experimentally characterized disordered proteins.

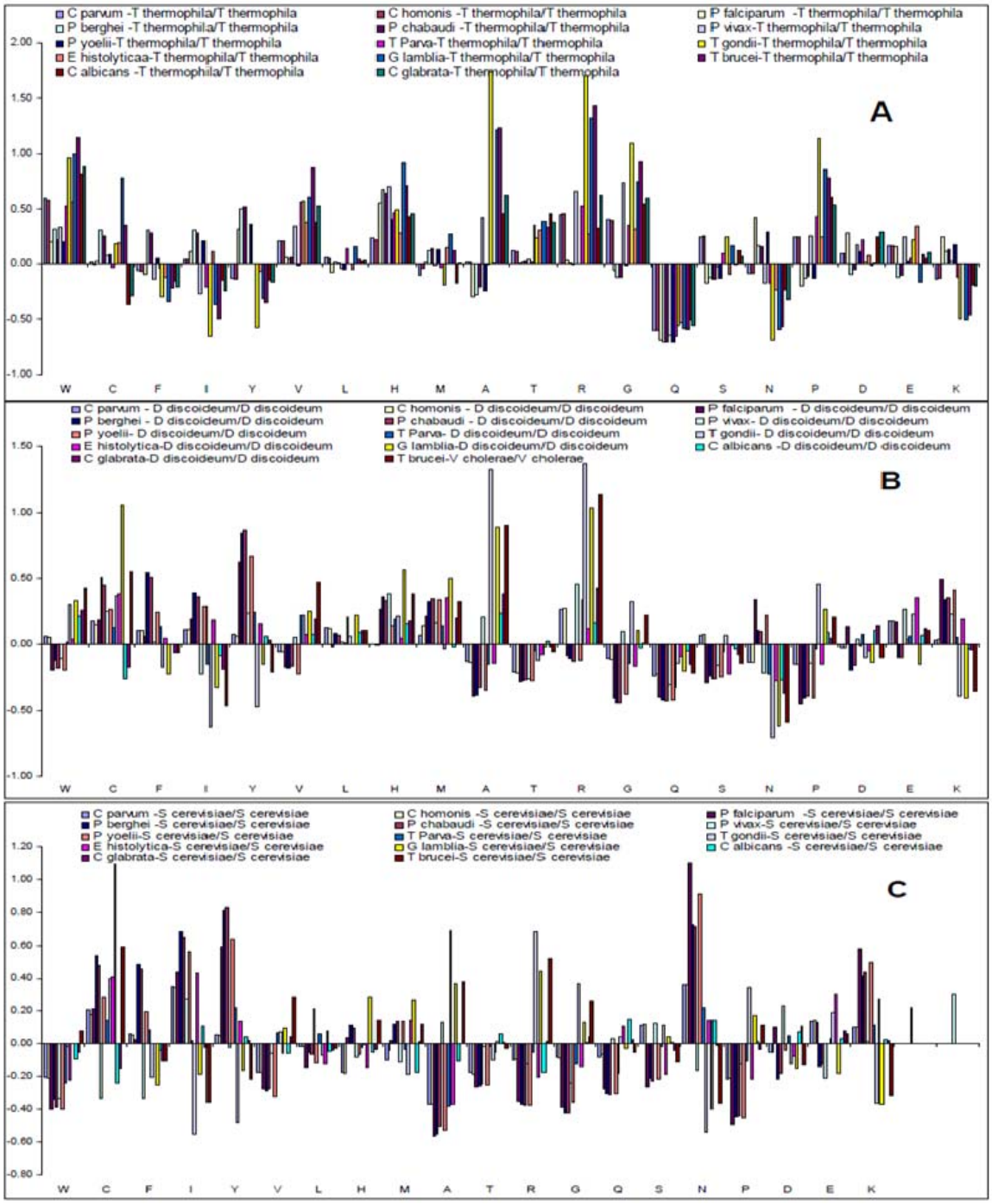

Figure 23: Amino acid compositions of pathogenic early-branching eukaryotic proteomes in comparison to three non-pathogens, *Tetrahymena thermophila* (A), *Dictyostelium discoideum* (B) and *Sacchromyces cerevisiae* (C).

The high abundance of phenylalanine and tyrosine residues might be related to peculiarities of protein folding and/or functionality.[202] Several intrinsically disordered proteins have shown to be enriched in these residues. For example, multiple tyrosine residues were shown to be essential for the function of the Ewings sarcoma (EWS) fusion proteins (EFPs). EFPs are potent transcriptional activators and reportedly interact with other proteins required for mRNA biogenesis. A characteristic functionality of EFPs is associated with the EWS activation domain (EAD), containing multiple degenerate hexapeptide repeats (consensus SYGQQS) with a conserved tyrosine residue. This intrinsically disordered domain was shown to be responsible for transcriptional activation and cellular transformation.[203] Furthermore, these multiply conserved tyrosines were shown to be essential for the EAD function. Intriguingly, they can be effectively substituted by phenylalanine, showing that an aromatic ring can confer EAD function in the absence of tyrosine phosphorylation.[203] Other examples include a set of phenylalanine–glycine repeat-containing nucleoporins (FG-Nups), specific proteins from nuclear pore complexes (NPCs) that are embedded in the nuclear envelope of eukaryotic cells. There are 13 such proteins in the *Saccharomyces cerevisiae* NPC. These are known to bind to karyopherins and facilitate the transport of karyopherin–cargo complexes. All these proteins were found to be intrinsically disordered and the FG repeat regions of Nups were shown to form a meshwork of random coils at the NPC through which nuclear transport proceeds. Another example is the immunoreceptor tyrosine-based activation motif (ITAM)-containing cytoplasmic domains of many immune receptors, which were recently shown to represent a novel class of intrinsically disordered proteins.[204, 205]

In comparison to their free-living, non-pathogenic counterpart (i.e., *Tetrahymena thermophila*), pathogenic early-branching eukaryotes are significantly enriched in aspartic acid (D), proline (P) and valine (V) along with polar residues such as tryptophan (W) and histidine (H). However, depletion of the polar residue glutamine (Q) appears to be common across all species in comparison to *T. thermophila* (Figure 23A). Compared to *C. elegans* or *V. cholerae*, depletion of tryptophan (W) and valine (V), both order-promoting residues, is apparent in the microbes analyzed (Figure 24, Figure 25).

Although many other amino acids are also depleted in various proteomes, W and V are the only two residues with consistent behavior across all species in comparison to *C. elegans* and *V. cholerae*. These figures also show evidence for a pronounced lysine (K) content amongst most parasites.

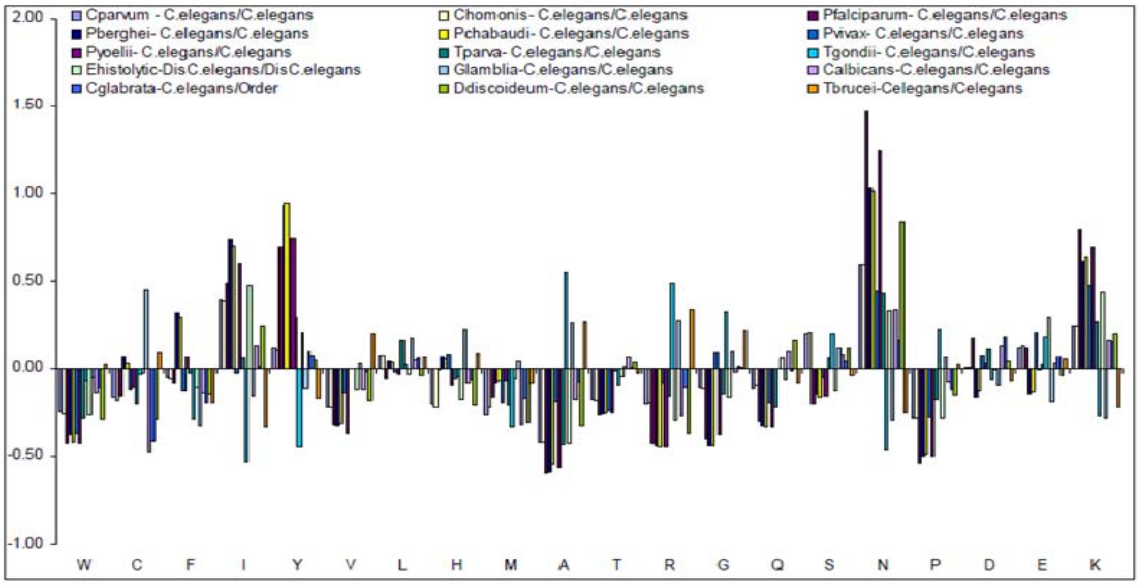

Figure 24: Amino acid compositions of pathogenic early-branching eukaryotic proteomes in comparison to a multicellular eukaryote, *Caenorhabditis elegans*.

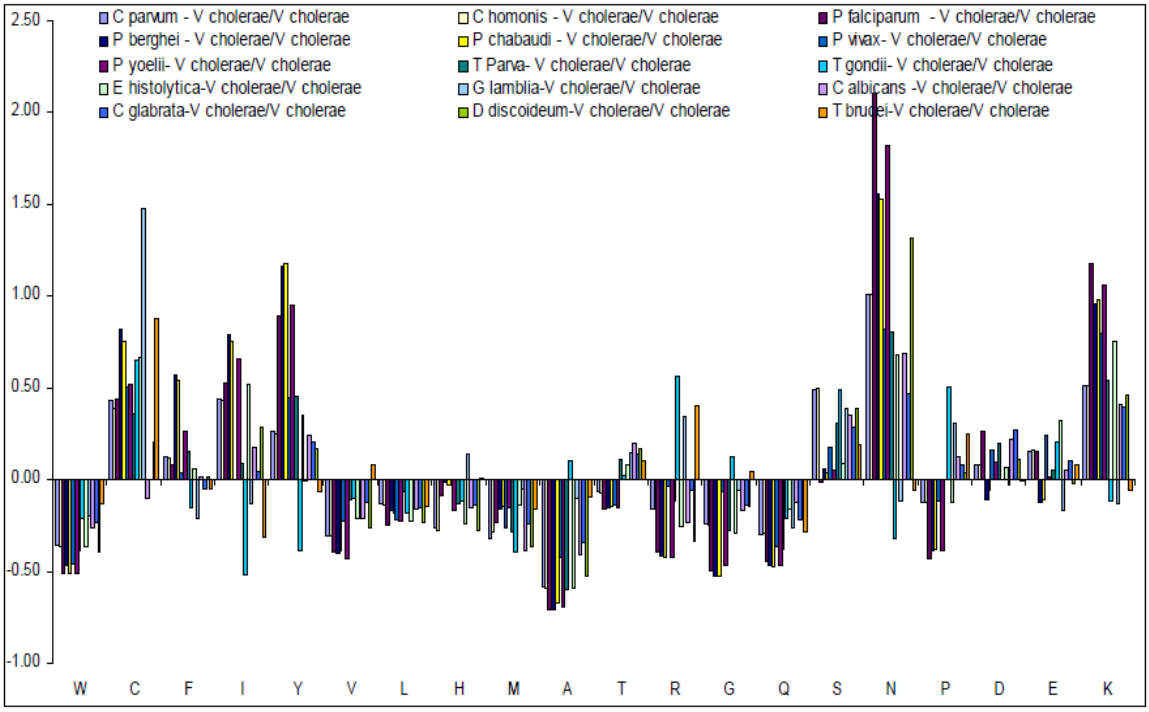

Figure 25: Amino acid compositions of pathogenic early-branching eukaryotic proteomes in comparison to a model prokaryote, *Vibrio cholerae*.

CDF and CH-plot analyses

The sequences of protozoan proteins were also used to predict whether these proteins are likely to be predominantly disordered using two binary algorithms of intrinsic disorder: the charge-hydropathy plot (CH-plot) algorithm[156, 200] and the cumulative distribution function approach (CDF analysis).[200] Both these methods classify whole proteins as either (a) mostly disordered or (b) mostly ordered. Here, the outcome of 'mostly ordered' suggests that proteins contain more ordered residues than disordered residues. Similarly, the outcome 'mostly disordered' indicates proteins that proteins contain more disordered residues than ordered residues.[200] A simultaneous observation of

low mean hydropathy and relatively high net charge is typical for the ‘‘natively unfolded’’ proteins, which are characterized by absence of a compact, collapsed structure.[156] Therefore, ordered and disordered proteins plotted in CH 2-dimensional space can be separated by a linear boundary, with proteins located above this boundary line being natively unfolded and with proteins lying below the boundary line being ordered. CDF analysis, on the other hand, summarizes the per-residue disorder predictions by plotting scores against their cumulative frequency. This allows ordered and disordered proteins to be separated based on the distribution of disorder prediction scores alone. In this study, order–disorder classification is based on whether a CDF curve is above or below a majority of boundary points: proteins with high scores will have CDF curves that have low cumulative values over most of the CDF curve, and proteins with low scores will have CDF curves that have high cumulative values over most of the CDF curve.[200] The individual results of CH-plot and CDF analyses for each of the 19 organisms are shown in Figure 26A-B. Figure 27 and 28 show CH-CDF analyses results for (A) 13 pathogens and (B) 3 non-pathogens in our data set.

Table 15 shows that there is a reasonable discrepancy between these two methods and the level of disorder predicted by CDF is on average 1.25-fold higher than that predicted by CH-plots. The difference between these two methods in the magnitude of predicted disorder supports previously published data[206]. This difference has been attributed to the fact that, a CH-plot is a linear classifier that considers only two properties of a particular sequence— net charge and hydropathy, whereas the results of a CDF analysis is strongly tied to the output of the nonlinear neural network based VLXT predictor. This predictor has been trained to learn order and disorder by using a much

larger feature space besides net charge and hydropathy. Owing to these methodological differences, CH-analysis learns to differentiate proteins with extended disorder (random coils and pre-molten globules) from proteins with globular conformations (molten globule-like and rigid well-structured proteins). On the other hand, VLXT based CDF analysis serves to segregate all disordered conformations including molten globules from rigid well-folded proteins. We believe that exploiting this difference in learning approaches of CDF and CH-plot can provide a computational tool to discriminate ‘‘natively unfolded’’ proteins in the apicomplexan phylum from native molten globules, that are predicted to be disordered by CDF, but compact by CH-plot. This model is consistent with the behavior of several intrinsically disordered proteins. Work is currently in progress to analyze the generality of this approach. Particularly in the context of protozoan proteins, this implies that some of them are predicted as extended, whereas others can possess molten globule-like properties.

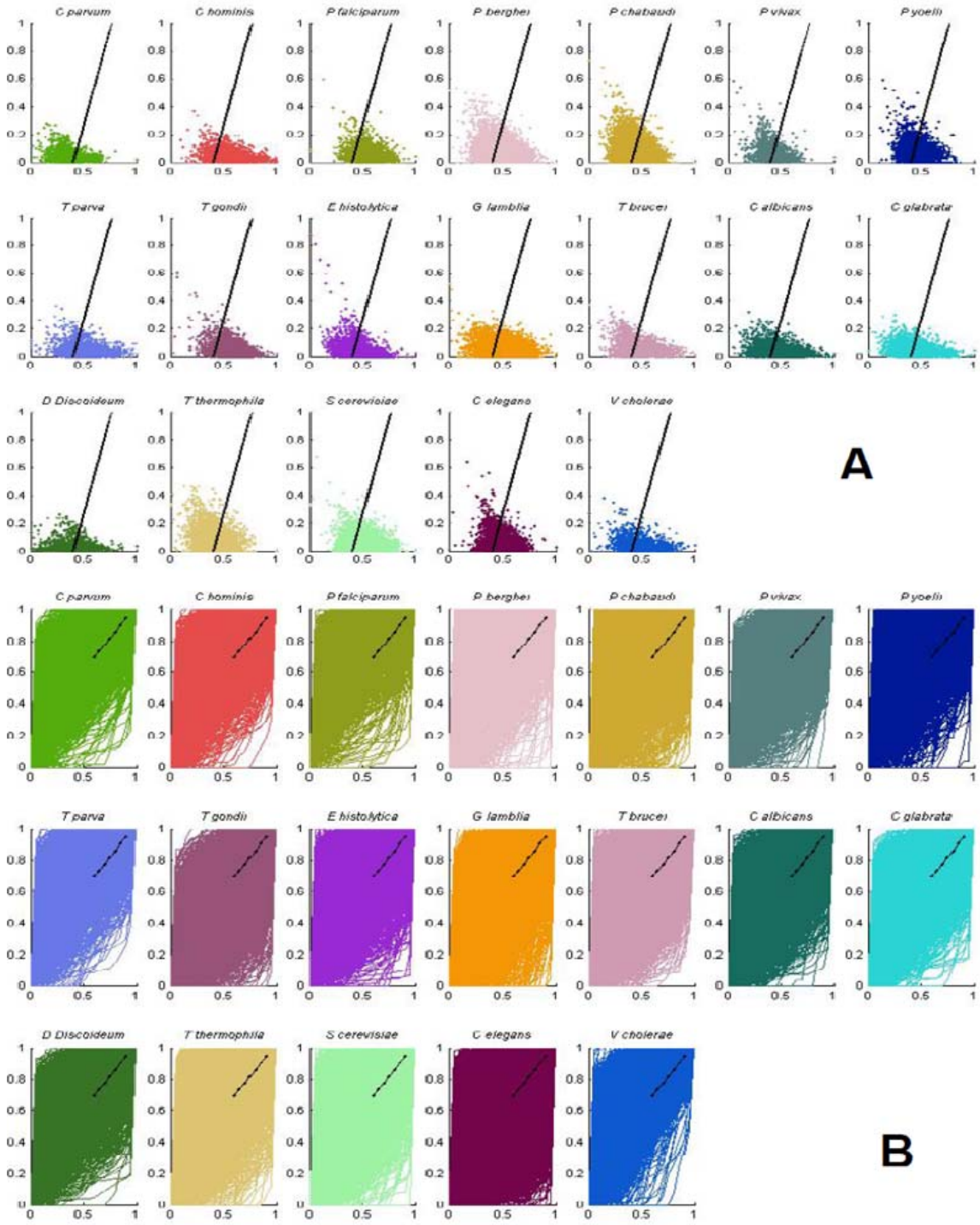

Figure 26: (A) Charge-Hydropathy plots (X-axis: Mean normalized hydropathy, Y-axis: Absolute mean net charge) (B) Cumulative distribution function curves (X-axis: Score, Y-axis: Cumulative fraction of residues) for all 19 organisms.

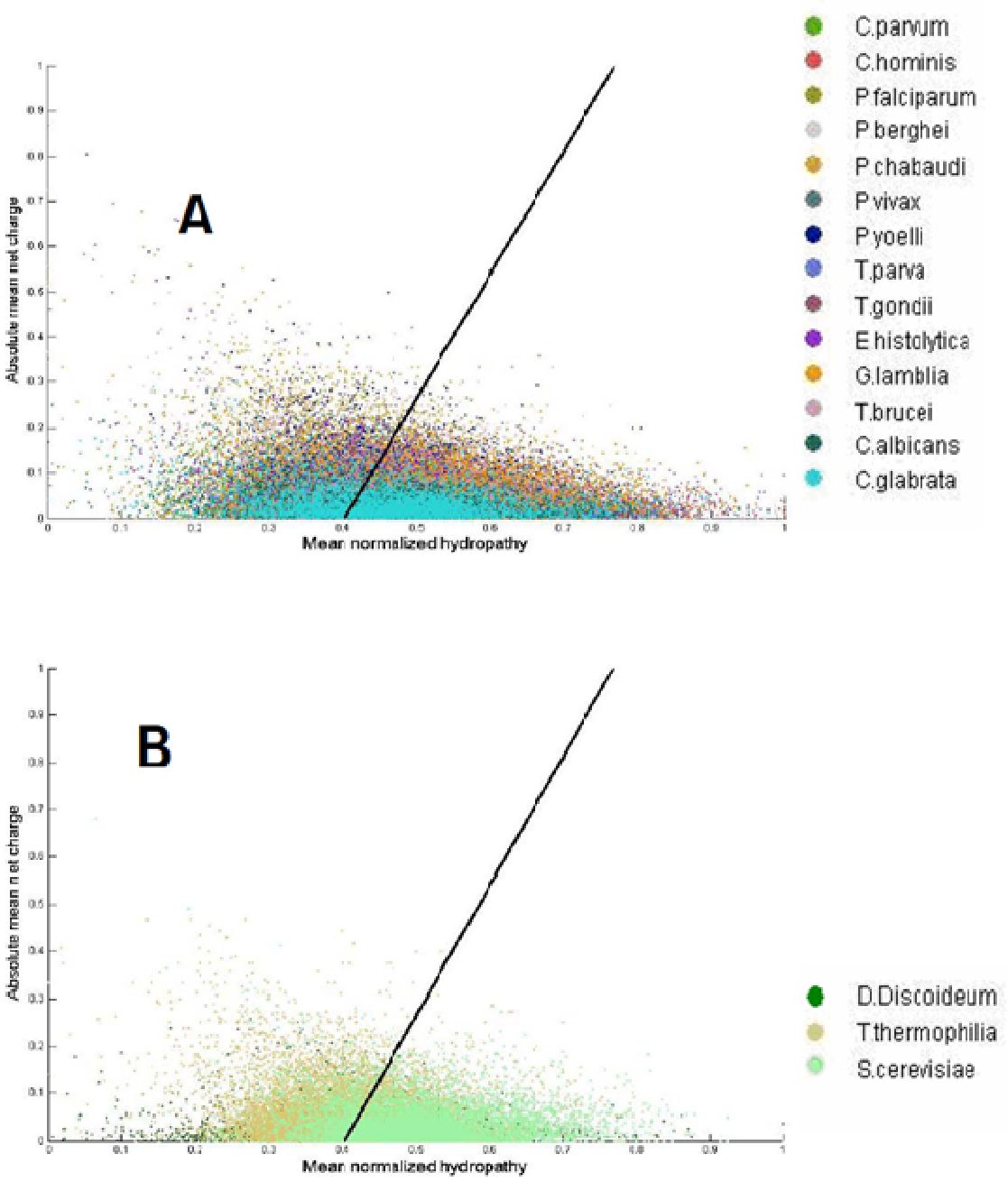

Figure 27: Charge-hydropathy plots corresponding to (A) 14 pathogens and (B) 3 non-pathogens as listed in Table 12.

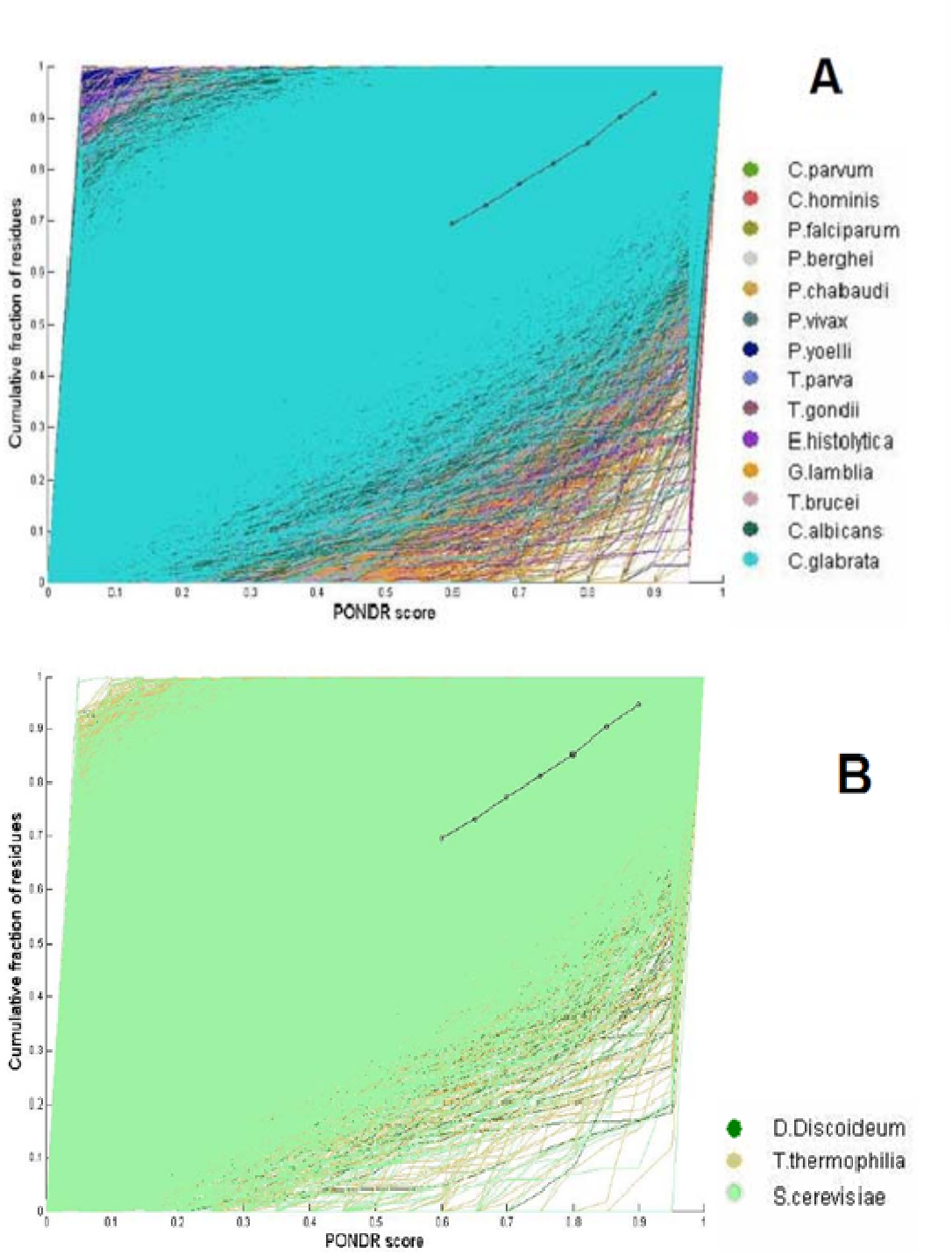

Figure 28: (A): Cumulative distribution function curves corresponding to (A) 14 pathogens and (B) 3 non-pathogens as listed in Table 12.

Organism	Total proteins used for predictions	Number of proteins with one or more α-MoRFs	Number of α-MoRFs	Number of proteins predicted disordered by both CH and CDF	Number of proteins predicted disordered only by CH	Number of proteins predicted disordered only by CDF	Number of proteins predicted ordered by both CH and CDF
C. parvum	3801	810	1375	367	238	357	2839
C. hominis	3884	720	1165	438	258	365	2823
P. falciparum	5400	1348	2757	509	1854	69	2968
P. berghei	10459	787	1204	745	2262	188	7264
P. chabaudi	12968	1133	1614	1235	2748	245	8740
P. vivax	5330	1878	4832	1287	440	897	2706
P. yoelii	7238	1155	2083	660	1729	159	4690
T. parva	4070	595	866	463	232	305	3070
T. gondii	7793	3761	11889	1300	73	3969	2451
E. histolytica	9766	991	1389	1113	705	487	7461
G. lamblia	9646	1242	1993	792	239	2025	6590
T. brucei	8758	2016	3530	944	146	2221	5447
C. albicans	6068	1472	2738	949	255	809	4055
C. glabrata	5271	1292	2390	763	210	849	3449
D. discoideum	4031	1153	2305	666	189	552	2624
T. thermophila	26212	5873	12121	2562	5703	726	17221
S. cerevisiae	10868	2302	4247	1448	433	1552	7435
C. elegans	38336	8379	15002	4339	751	6890	26356
V. cholerae	3829	72	76	92	60	327	3350

Table 13: CH, CDF and α-MoRF prediction results for all 19 organisms.

Figure 29 compares the results of the CH-plot and CDF analyses by showing the distributions of proteins in each proteome within the CH–CDF phase space. In these plots, each colored data point represents a single protein whose spatial coordinates are calculated as a distance of this protein from the boundary in the corresponding CH-plot (y-coordinate) and an averaged distance of the corresponding CDF curve from the boundary (x-coordinate). Positive and negative y values correspond to proteins, which, according to CH-plot analysis, are predicted to be natively unfolded or compact, respectively. On the other hand, positive and negative x values are assigned to proteins that, by the CDF analysis, are predicted as ordered or intrinsically disordered, respectively. Therefore, each plot contains four quadrants (see an explanatory panel in the low right corner of Figure 29): (-, -) contains proteins predicted to be disordered by CDF, but compact by CH-plot (i.e., proteins possibly with molten globule-like properties); (-,

+) includes proteins predicted to be disordered by both methods (i.e., random coils and pre-molten globules); (+, -) contains ordered proteins; (+, +) includes proteins predicted to be disordered by CH-plot but ordered by the CDF analysis. The sharp cut-off at the right side of each plot is due to the upper limit of a difference between the CDF curve (which has a maximum value of 1.0) and the boundary separating IDPs and ordered proteins in CDF plots. Analysis of the (-, -) and (-, +) quadrants in Figure 29 shows that the majority of the wholly disordered proteins from *C. elegans*, *S. cerevisiae*, and *V. cholerae* likely possess molten globule-like properties. In contrast, protozoan proteomes are generally characterized by a more balanced distribution between compact and extended disordered proteins. This balance is also observed in the case of *C. albicans* and *C. glabrata* proteomes demonstrating some prevalence for the extended disordered proteins.

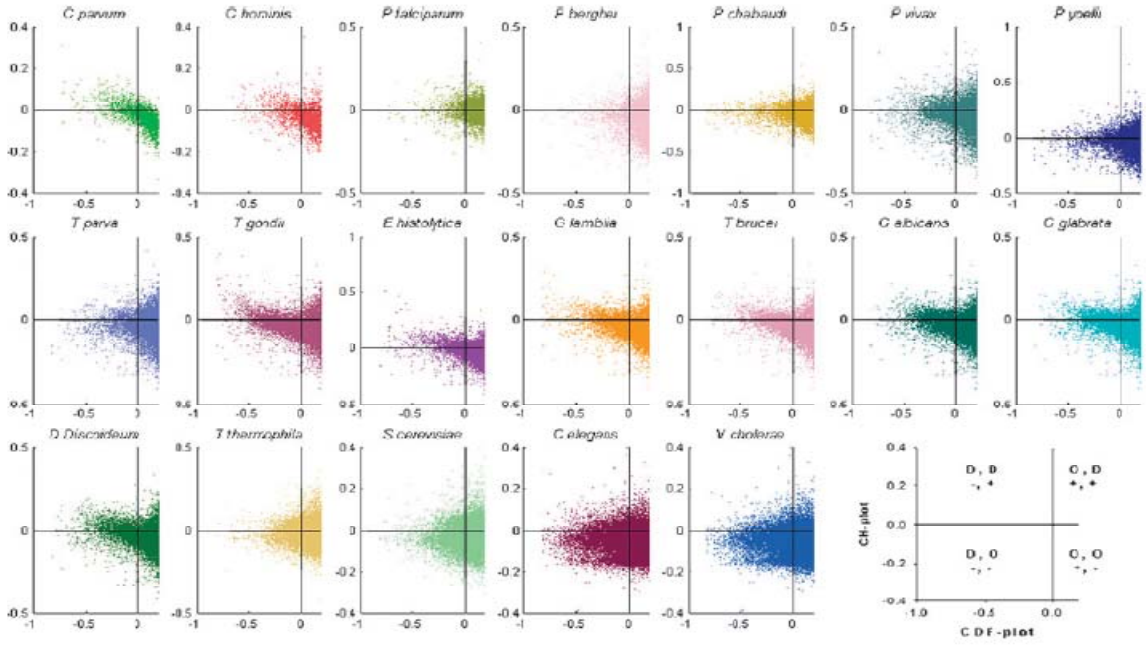

Figure 29: Comparison of the CDF and CH-plot analyses of whole protein order-disorder via distributions of proteins in each proteome within the CH-CDF phase space.

Prediction of disorder

We used two more approaches to further assess the presence of intrinsic disorder in each of the early-branching eukaryotic proteomes. First, the abundance of predicted intrinsic disorder in various organisms was estimated by calculating the fractions of proteins containing predicted disordered regions of a given length (e.g. ≥30, ≥40). This approach has earlier been used to show the presence of intrinsic disorder in signaling and cancer-associated proteins[43] and in proteins involved in cardiovascular diseases. [206] Figure 30A[202] shows that intrinsic disorder is predicted to be relatively abundant in early-branching eukaryotes. The percentages of proteins with 30 or more consecutive residues predicted to be disordered by VSL2B and VLXT (corresponding numbers are shown in brackets) were 87.8% (89.8%) for *T. gondii*, 80.3% (82.5%) for *P. vivax*, 79.0% (81.0%) for *P. falciparum*, 75.3% (76.8%) for *D. discoideum*, 73.8% (75.1%) for *C. parvum*, 72.4% (74.1%) for *C. albicans*, 71.9% (73.1%) for *C. glabrata*, 71.4% (72.4%) for *T. thermophila*, 70.4% (72.0%) for *T. brucei*, 69.7% (70.9%) for *C. hominis*, 67.5% (68.9%) for *T. parva*, 63.0% (64.5%) for *P. yoelii*, 62.6% (64.1%) for *S. cerevisiae*, 63.0% (64.3%) for *C. elegans*, 58.2% (59.5%) for *E. histolytica*, 52.1% (53.2%) for *G. lamblia*, 42.5% (43.4%) for *P. berghei*, 40.3% (41.3%) for *P. chabaudi* and 24.9% (25.1%) for *V. cholerae*.

A previous study using VLXT showed a set of eukaryotic proteins from Swiss-Prot and a set of ordered proteins from PDB Select 25, contained 47(±4)% and 13(±4)% proteins with 30 or more consecutive residues predicted to be disordered.[43] Therefore, in comparison to a set of ordered proteins, microbial proteomes were found enriched in proteins containing long disordered regions. Furthermore, the vast majority of the early-

branching eukaryotic organisms (except for *G. lamblia*, *P. berghei*, and *P. chabaudi*) contained more proteins with long disordered regions than a set of representative eukaryotic proteins from Swiss-Prot.

In the second approach to assess the prevalence of intrinsic disorder in early-branching eukaryotes, we compared the percentages of residues in long disordered regions (30 or more consecutive residues) as predicted by VSL2B (VLXT). These percentages were as follows (Figure 30B[202]): 58.3% (36.1%) *T. gondii*, 42.7% (7.3%) for *P. falciparum*, 41.5% (21.2%) for *P. vivax*, 37.9% (6.5%) for *T. thermophila*, 37.1% (6.9%) for *P. yoelii*, 34.4% (16.4%) for *D. discoideum*, 29.6% (6.1%) for *P. chabaudi*, 29.5% (13.7%) for *C.albicans*, 29.2% (4.2%) for *P. berghei*, 28.5% (14.5%) for *C. glabrata*, 27.7% (15.6%) for *C. elegans*, 27.6% (13.5%) for *S. cerevisiae*, 27.5% (17.8%) for *T. brucei*, 27.2% (10.4%) for *C. hominis*, 26.2% (9.4%) for *C. parvum*, 24.6% (12.5%) for *G. lamblia*, 23.5% (9.4%) for *T. parva*, 19.6% (6.6%) for *E. histolytica*, and 6.2% (6.7%) for *V. cholerae*. According to previous VLXT estimates, there were 6.5(±0.5)% and 1.48(±0.45)% residues in long disordered regions of eukaryotic proteins from Swiss-Prot and of non-homologous ordered proteins from PDB, respectively.[43] The data presented here suggests that sequences from early-branching eukaryotes, contain more disordered residues than eukaryotic Swiss-Prot proteins and ordered PDB proteins.

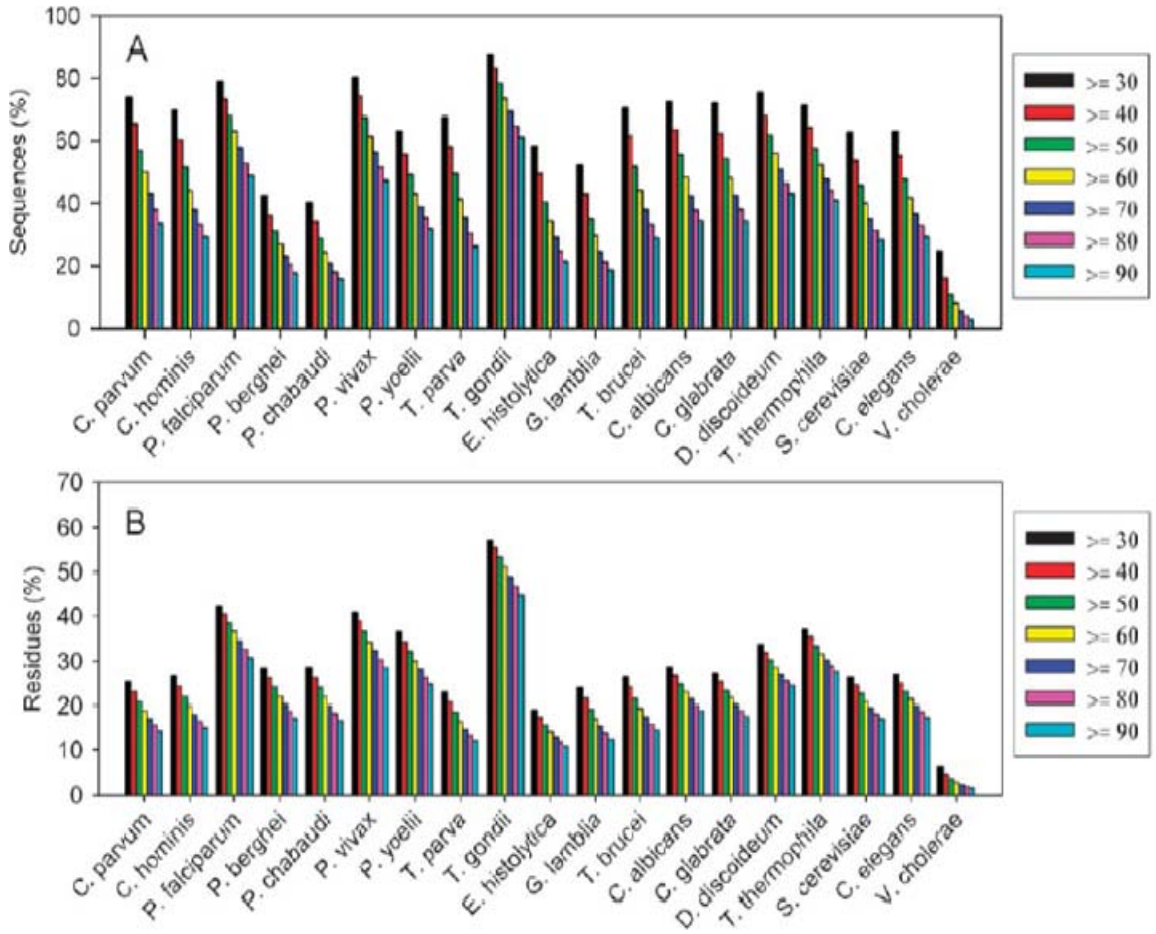

Figure 30: VSL2B disorder prediction results on 19 proteomes: *C. parvum, C. hominis, P. falciparum, P. berghei, P. chabaudi, P. vivax, P. yoelii, T. parva, T. gondii, E. histolytica, G. lamblia, T. brucei, C. albicans, C. glabrata, D. discoideum, T. thermophila, S. cerevisiae, C. elegans,* and *V. cholerae.* (A) Percentages of proteins in the 19 proteomes with ≥30 to ≥90 consecutive residues predicted to be disordered. (B) Percentages of residues in these 19 proteomes predicted to be disordered within segments of length ≥30 to ≥90.

Predictions of alpha-MoRFs

Table 13 also shows that, on average, nearly 20% of protozoan proteins contain α-MoRFs, ranging from 7.5% in *P. berghei* to 48.3% in *T. gondii*. The number of a-MoRF-

containing proteins in the prokaryotic representative *V. cholerae* is relatively smaller (1.9%). Importantly, in each proteome some long, highly disordered proteins have multiply predicted a-MoRF regions (Supplementary Table S1, Mohan et al.,[202]) that may potentially serve as binding sites for multiple proteins. For example, *C. elegans* protein CE25234 (4900 amino acid residues) has 49 predicted α-MoRFs. Similarly, *T. gondii* proteins 44.m02695 (putative protein phosphatase 2C, 3966 amino acids) and 42.m03467 (mediator complex subunit SOH1-related, 4253 amino acids) contain 24 and 22 predicted a-MoRFs respectively.

Analysis of *Plasmodium falciparum* protein–protein interaction map

The goal of this analysis was to study a published interaction map of *P. falciparum* while paying special attention to the degree of intrinsic disorder in the network. This map includes 2,321 proteins involved in 19,979 protein–protein interactions.[200] A log–log plot of the number of proteins versus the number of interactions shows that the published interaction map closely mimicked the properties of a scale-free network (Figure 31). Such networks are characterized by the presence of a few proteins participating in a high number of interactions (also known as *hubs*) and a large number of proteins having few or no interactions. This finding is further supported by a regression analysis of the data using the least squares method that showed that the data fits a linear equation with a negative slope. The fact that the R^2 value (0.9) is close to 1 is indicative of a reasonably good fit.

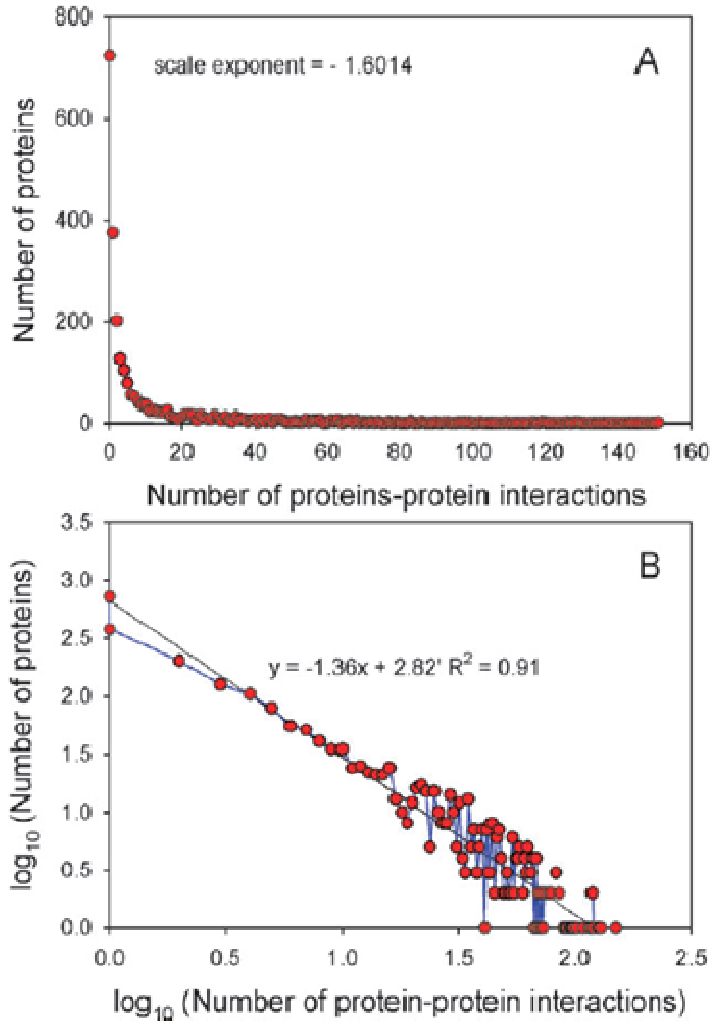

Figure 31: Analysis of *P. falciparum* interaction map. (A) Number of protein–protein interactions (x-axis) vs. number of proteins (y-axis) based on *P. falciparum* interaction map published in Wuchty and Ipsaro, 2007 (B) Log–log plot obtained using data from Figure 31A.

VSL2B predictions of proteins from the interaction map show that there is ≈45% disorder in this map. This number is marginally higher than the overall amount of disorder present in all annotated proteins from *P. falciparum* (41.6%) and is significantly higher than the level of intrinsic disorder in the *C. elegans* proteome (35.9%), the *V. cholera* proteome (22.2%), as well as the all early-branching eukaryotes (39.0%, see

Table 14). We also found that the correlation score between per protein VLXT score and the number of interactions by it was only 0.13 (p-value = 0.0001). Such a low correlation score is indicative of a weak association between intrinsic disorder and connectivity in the *P. falciparum* protein–protein interaction map.

α-MoRF predictions for these data reveal that, of the 529 putative hub proteins (i.e., proteins involved in 10 or more protein–protein interactions), 134 contain one or more predicted α-MoRF regions (25.3%).[202] In comparison to this, 600 of the 1792 likely non-hub proteins had a corresponding α-MoRF prediction (33.5%). Both these numbers are higher than the average number of eukaryotic proteins with predicted α-MoRFs (≈23%) and are significantly higher than a number of MoRF-containing proteins in bacteria (≈3%) and archaea (≈2.5%). In other words, both protein sets are highly enriched in disordered segments that are potentially involved in molecular recognition and that undergo disorder-to-order transitions upon interaction with their binding partners. Interestingly, non-hub proteins on average contain more α-MoRFs than hubs. On the other hand, the VLXT scores of 24.5% and 19.7% characterize hub and non-hub proteins respectively. This apparent discrepancy can be explained by the fact that MoRFs are short ordered regions (around 20 residues) located within long disordered regions. Therefore, higher MoRF content should correspond to lower overall disorder score.

CHAPTER SEVEN: DISCUSSION

A practical limit to intrinsic disorder prediction

Our study on the reproducibility of intrinsic disorder in Chapter Three addresses the relationship between intrinsically disordered protein regions and crystallographic structure determination. We find that the experimental reproducibility of disordered regions between highly similar proteins (≥90% global sequence identity) is strongly dependent on the parameters applied to a crystallization experiment, such as temperature, pH, and salt concentration. For the highly similar proteins crystallized under the agreement of all experimental parameters, the reproducibility of disordered regions was about 81%, while for completely different experimental conditions this reproducibility dropped to 40%. We believe that other extraneous factors such as the presence/absence of ligands is less likely to influence experimental factors. We also propose that experimental reproducibility of disordered regions can be used as a good indicator of an upper bound for the predictability of disordered regions. Given the continued use of current crystallization methods, we estimate that a standard computational experiment based on the crystallized proteins from PDB can achieve about 80% accuracy on average. If experimental conditions are taken into consideration, this accuracy may reach about 90%, while in the case when experimental conditions are different; this accuracy drops to 69%.

Since the estimated reproducibility of disordered regions reflects the overall likelihood of a protein residue to be disordered, we also constructed prototype predictors of disordered regions when experimental conditions are taken into consideration. The results of accuracy estimation show that a smaller sample of non-redundant proteins used

during training, but from the same class of experimental conditions, is either sufficient or even better for prediction accuracy than when a larger sample of proteins from all classes of experimental conditions are considered. This is an interesting finding because it suggests that the relationship between disordered regions and experimental conditions of structure determination is non-random and predictable, to a certain degree, just by using amino acid compositions from such proteins. One possible way of interpreting these results is that there is still room for improving prediction of intrinsically disordered regions. It is possible that the limits suggested here may not be achieved by the use of sequence based predictors alone. Therefore, methods that can exploit tertiary interactions as well as experimental conditions may be able to narrow the gap. Although it does not include experimental condition-specific features, a recent work presented a sequence-based and structure-based method for prediction of disordered regions.[207] We believe that the analyses in this study not only provides a quantitative view of the crystallographic inaccuracy, especially with respect to modeling protein dynamics using a static view, but also provides further clues with respect to driving crystallographic experiments through the incorporation of experimental conditions.

Molecular recognition by MoRFs involve disorder-to-order transitions

The purpose of manually examining a few examples of MoRFs in Chapter Four was to visualize a few instances of proteins or protein fragments that envelop their respective protein partners and participate in molecular recognition by a disorder-to-order transition. We have shown that MoRFs are unstructured in their unbound form via amino

acid compositional profiles, secondary structure predictions, interaction surface analyses and predictions. Each of these experiments has shown that MoRFs mimic the general properties of intrinsically disordered proteins in isolation of their binding partners and therefore conform to the MoRF theory. The disordered state of MoRFs allows them to bind to specific partners via a disorder-to-order transition.

Protein binding via disorder-to-order transition can be treated as a special type of protein folding mechanism. Conventional protein folding involves the formation of tertiary structure that stabilizes secondary structure elements. In disorder-to-order binding, formation of contacts between a MoRF and its binding partner stabilizes the secondary structural elements on the MoRF. We suggested two mechanisms in which MoRFs gain structure. The first mechanism is the inherent-structure mechanism which reflects the predominance of a specific local secondary structure type among the highly fluctuating conformations of the unbound MoRF. In this case, the structure of the MoRF is not entirely random and shows some features that are later stabilized in the bound state. In the second mechanism or the induced-structure mechanism, the MoRF is entirely disordered before binding and makes initial intra- and inter-chain contacts with its partner randomly. These contact points serve as nucleation sites for the subsequent folding and formation of secondary structure under the influence of subsequent contacts with the partner molecule. In such a mechanism, the inherent conformational preferences of the intrinsically disordered protein itself are overridden by interactions with the partner. The inherent-structure mechanism has been substantiated by comparing experimental and predicted secondary structure of MoRFs. The second mechanism or the induced-structure mechanism has been supported by presenting the accuracy of predicted structures of

MoRFs in comparison to monomers. It is equally likely that a combination of both of these mechanisms is at play in MoRF mediated interactions.

Identification of phospeptides in LC-MS/MS can be improved with application of DisPhos and peptide detectability scores

Although machine learning approaches have been applied to proteomics research in the past most of these approaches have largely concentrated on either the preprocessing of the tandem-mass spectra or the post-processing the peptide identification results by the use of traditional tools such as Mascot and Sequest. In Chapter Five, we described a novel approach that combines functional-residue predictors based on intrinsic disorder and sequence properties, with LC-MS/MS proteomics algorithms to improve identification of phospeptides. Our decision to develop such an approach stems from a number of previous articles that have noted the presence of phosphorylation sites in regions of intrinsic disorder. In addition to this, our method also incorporates a previously proposed concept known as 'peptide detectability' which is the probability of observing a peptide in a standard sample analyzed by a proteomics experiment. Our linear support vector machine predictor based on DisPhos (a logistic-regression based predictor of phosphorylation sites) generated features, peptide detectability, twenty basic amino acid compositions, the number of serines, threonines and tyrosines in a peptide, the length and mass of a peptide, can predict +2 and +3 LC-MS/MS peptides containing phosphorylation sites can be predicted with 73 – 81% accuracy. In addition to this, we also suggested an algorithm where only phospeptides

with top *n*% prediction scores were used for a Mascot search and compared it with the standard Mascot process which utilizes all likely phospeptides. Our results show that the proposed approach is able to identify all phospeptides identified by a standard Mascot search by using a reduced protein database. In addition to being able to effectively select the correct phospeptides, our algorithm is also able to gain as many as 26% more + 2 phospeptides (+31% phosphorylation sites) and nearly doubles +3 phospeptides at 1% FDR by removing unlikely phospeptide candidates, thereby maximizing the efficient use of computational resources. This gain was achieved by using only the top 20% and 10% scoring phospeptides respectively.

Pathogenic organisms have increased intrinsic disorder content in comparison to non-pathogenic organisms

Our study probing the degree of intrinsic disorder in pathogenic organisms in Chapter Six provides new insights into the evolution of intrinsic disorder in the context of adapting to a parasitic lifestyle. We describe and discuss a systematic bioinformatics approach that was used for the discovery and analysis of unfoldomes, the complement of intrinsically disordered proteins in a given proteome) of early-branching eukaryotes. Our results suggest that sequences from early-branching eukaryotes are predisposed to a higher degree of unfoldedness than eukaryotic Swiss-Prot proteins and ordered PDB proteins. We have also established that many protozoan sequences (20 - 60% depending on the organism) contain long disordered regions, disordered regions (lengths ≥ 90 consecutive residues). This corresponds to a 7-fold increase in comparison to the number

of similar regions from a PDB Select 25 data set. Therefore we believe that early-branching eukaryotic proteins are significantly enriched in predicted disorder in comparison to representative eukaryotic proteins from Swiss-Prot and ordered proteins from the Protein Data Bank. Finally, we proposed that regions of intrinsic disorder in pathogenic protozoa provide a flexible means to facilitate host cell invasion and overcome immune response. Our results stress upon the need for continued research in this direction to ascertain the contribution of intrinsically disordered proteins in the cellular physiology of parasitic organisms.

CHAPTER EIGHT: SUMMARY AND FUTURE WORK

Summary of dissertation

This dissertation aims to address the delicate balance that exists between intrinsically disordered regions and the linear sequence of amino acids harboring such regions in two important contexts: (1) the influence of the sequence environment on the presence or lack of such regions and (2) the role of such regions on the functionality embedded within complex biological processes as well as systems.

Our investigations on the former front convey that the existence, position, and length of disordered regions in highly similar proteins are strongly dependent on the variations in amino acid sequence as well as the parameters of crystallographic experiments, such as temperature, pH, and salt concentration. We find that for identical protein sequences, a majority of the observed modulations in the crystal lattice can be explained by variations applied to experimental conditions at the time of crystallization. For highly similar chains, both experimental conditions and the intrinsic change of protein structure were significant factors. At this time, we are hesitant to assign relative importance to these factors since the observed sequence differences in PDB are likely to be non-random (for example, mutations with functional or phenotypic significance are frequently of interest for structure determination). Having said that, the effect of chemical ligands on our analysis was limited, thereby making them less significant in the overall placement of disordered regions. The presence of a disordered region under one set of experimental conditions and absence under another can be understood through the framework of the probabilistic theory of protein folding. At any time instant, a protein

can be assigned a probability of any particular conformation based on its energy landscape.[208, 209] For ordered proteins, such energy landscapes are characterized by single (or a small number of) deep minima with high probabilities associated with the corresponding conformations. Since the number of conformations in the high energy states is enormous and the barriers for moving away from the dominant conformation are relatively large, the energy landscape has a shape of a funnel.[208] This minimum energy state is often associated with protein function and is called the native state. On the other hand, the energy landscapes for disordered proteins are shallower, typically characterized by flat and rugged valleys, i.e. they contain a large number of energy minima with relatively small barriers for transitioning between distinct conformations.[209] Consequently, the probability of each conformation corresponding to an energy minimum is relatively low. The absence of a high probability conformation eventually leads to missing electron density during crystallographic experiments. Thus, the variability in structures of identical proteins solved under different experimental conditions is caused by the environment-driven changes of the energy landscape (Figure 32). The altered probability distribution over the space of allowed tertiary structures ultimately results in a population shift between ensembles of pre-existing conformational isomers.[208-210]

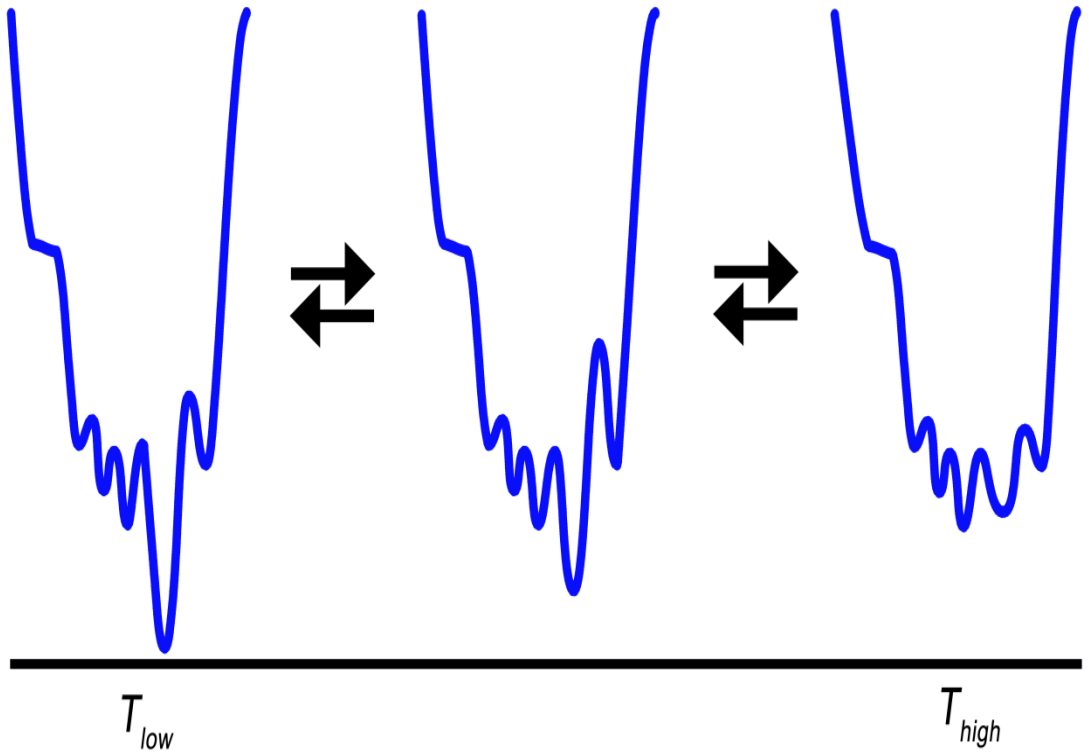

Figure 32: Stylized depiction of the energy landscape as a function of the environment.

In general, our work provides evidence that disordered protein regions are very sensitive to changes in amino acid sequence and experimental conditions of crystallographic experiments. The success of such crystallographic experiments depends on the complexity of a protein's structure and also on a number of experimental or environmental factors including purity of the protein sample, temperature, ionic strength, pH, and precipitants such as ammonium sulfate or polyethylene glycol.[118] Undoubtedly, there are a number of factors that distinguish crystallization conditions from physiological conditions, but there is also a body of evidence that supports that protein structures often correspond to their native states.[118] Therefore, it is reasonable to speculate that a wide range of intracellular and extracellular conditions may have similar

effects on the dynamics of protein 3D structure in vivo. The habitats for many living organisms vary from acidic to cold or hot, with various species being able to tolerate wide ranges of environmental conditions. The summary from our analysis suggests that, any similar variations in cellular environments could have profound effects on protein structure, dynamics, and function. Sensitivity to sequence changes, on the other hand, may facilitate the evolution of function, especially for proteins with the same fold classification.

We also examined the role of intrinsic disorder in molecular recognition, post-translational modifications and pathogenesis to address our second objective for this dissertation.

Our section presenting multiple examples of intrinsically disordered molecular recognition features or MoRFs such as tumor suppressor p53, Wiskott-Aldrich syndrome protein (WASP), the VCA and WH2 domains highlighted a novel way in which disorder mediates protein-protein interactions via the process of molecular recognition. MoRFs bind to their specific partners through a disorder-to-order transition. We studied the occurrence of comparably short fragments (< 70 residues), loosely structured protein regions within longer, largely disordered sequences that were characterized as bound to larger proteins also known as molecular recognition features (MoRFs). We show that, upon binding to their partner(s), MoRFs undergo disorder-to-order transitions.

Through extensive use of available computational tools for a bioinformatics analysis, we demonstrated that there is indeed an abundance of intrinsic disorder in the proteomes of early-branching eukaryotes, many that are pathogenic. While our analysis of a published *P. falciparum* interactome revealed a weak correlation between disorder

and the proclivity to engage in protein–protein interactions, many more similar networks must be evaluated before reaching a definitive conclusion on this front. Our study of fourteen apicomplexan pathogens in comparison to a non-pathogen apicomplexan, a model eukaryote as well as a model prokaryote suggests that pathogens in general have much higher content of intrinsically disordered proteins in comparison to their contemporaries. This could indicate that pathogenic organisms have evolved to retain larger fractions of low complexity regions than other organisms, especially non-pathogens, perhaps to bypass the host system's immune response at the time of invasion. Several other interesting patterns in the amino acid compositions, α-MoRF predictions, charge-hydropathy scores and cumulative distribution frequencies were also discovered in pathogenic organisms. Given the high degree and unusual nature of the intrinsically disordered regions we have analyzed here, it is clear that further steps to elucidate their biological roles in the context of parasite physiology and pathogenesis will be well worth the effort.

In summary, the conclusions and arguments presented in this dissertation emphasize the renewed need to explore the mechanics underlying the still unknown behavior of cellular systems must be initiated with special care being paid this time on intrinsically disordered proteins and their contribution to overall cellular physiology. Finally, given the well-established links between intrinsic disorder, cancer and neurodegenerative diseases we believe the work discussed and presented here will have positive ramifications in areas such as protein engineering and synthetic biology with emphasis on cancer therapeutics and discovery of preventive care for neurodegenerative conditions.

Future Research

With positive correlation established between intrinsic disorder and cancer[43], neurodegenerative[211], cardiovascular diseases[206] and pathogenic organisms[202] there is a pressing need to develop fast and accurate methods and tools that can predict a protein's structural and functional propensities. These can help establish the likelihood of an organism being susceptible to such diseases. An important step towards achieving this objective is trying to overcome the suggested upper limit to the prediction of intrinsic disorder by exploring novel methods that include structure and crystallization features. Our observation that disordered regions are responsive to environmental parameter perturbations motivates further studies probing environmental factors that affect protein function. We may thus be able to gain insights into the evolution of such regions and proteins. Our findings also have a direct impact on the ability to make educated estimates about experimental conditions for future structure characterization projects.

The discussion on MoRF examples presented in Chapter Four (Figure 11) suggests the possibility of a positive correlation between VLXT predictions and the sequence location of MoRFs in proteins. This finding has direct repercussions to our currently limited understanding of protein-protein interactions especially those involving structural disorder. By exploiting this new information presented on MoRFs, we can approach the problem of protein binding site predictors with a renewed perspective. Aside from this, the availability of a relatively larger number of MoRFs can also guide the development of more sophisticated MoRF predictors.

Previously significant correlations have been drawn between phosphorylation sites and cancer-associated proteins. [212] With the development of a new method that can

improve the identification of phosphorylation sites in high-throughput experiments, detecting phosphorylation sites in proteins found in tumorigenic tissues can be expedited. Any new sites discovered can potentially play a role in modulating tumor cell survival rates thereby controlling the length and success of anti-cancer drug clinical trials.

Lastly, our observation that a high degree of intrinsic disorder exists in apicomplexan pathogen proteomes suggests that further steps to elucidate the biological roles of disordered regions in the context of parasite physiology and pathogenesis would be effort well spent. Some of the interesting questions that can be asked here include, whether changes in pH, temperature, salt or other environmental conditions in host cells affect the survivability of a pathogen? Can docking of pathogens onto host cell proteins be disrupted by stabilizing disordered regions in them? By answering such questions and more, we can advance our understanding of host-pathogen interactions and learn to predict the progress of evolving diseases caused by pathogens besides identifying suitable anti-infective therapeutics and vaccination strategies, potentially before the onset of epidemics or pandemics.

To summarize, the results and analyses presented in this dissertation challenge the traditional, three-dimensional structure-based approach towards understanding the functionality and cellular physiology of proteins. The novel aspects of intrinsically disordered proteins presented here highlight the possibility of an alternative evolutionary niche occupied by disordered proteins thereby compelling one to re-think the evolution of protein function with a renewed perspective.

REFERENCES

1. Dyson, H. J.; Wright, P. E., Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **2005,** 6, (3), 197-208.
2. Uversky, V. N., Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **2002,** 11, (4), 739-756.
3. Uversky, V. N., Cracking the folding code: why do some proteins adopt partially folded conformations , whereas others don't? *Science* **2002,** Submitted.
4. Uversky, V. N., What does it mean to be natively unfolded? *Eur. J. Biochem.* **2002,** 269, (1), 2-12.
5. Uversky, V. N.; Oldfield, C. J.; Dunker, A. K., Showing your ID: intrinsic disorder as an ID for regcognition, regulation, and cell signaling. *J. Mol. Recogntion* **2005,** 18, (5), 343-384.
6. Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z., Intrinsic disorder and protein function. *Biochemistry* **2002,** 41, (21), 6573-6582.
7. Dunker, A. K.; Brown, C. J.; Obradovic, Z., Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* **2002,** 62, 25-49.
8. Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N., Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* **2005,** 272, (20), 5129-48.
9. Daughdrill, G. W.; Pielak, G. J.; Uversky, V. N.; Cortese, M. S.; Dunker, A. K., Natively disordered protein. In *Protein Folding Handbook*, Buchner, J.; Kiefhaber, T., Eds. Wiley-VCH: Verlag GmbH & Co. KGaA: Weinheim, 2005; pp 271-353.
10. Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z., Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001,** 19, (1), 26-59.
11. Dunker AK., O., Z., Romero, P., Garner, E. C., and Brown, C. J. , , Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform* **2000,** 11,161-171. .
12. Uversky, V. N.; Fink, A. L., Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim Biophys Acta* **2004,** 1698, (2), 131-53.
13. Uversky, V. N.; Gillespie, J. R.; Fink, A. L., Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000,** 41, (3), 415-427.
14. Uversky, V. N.; Gillespie, J. R.; Millett, I. S.; Khodyakova, A. V.; Vasilenko, R. N.; Vasiliev, A. M.; Rodionov, I. L.; Kozlovskaya, G. D.; Dolgikh, D. A.; Fink, A. L.; Doniach, S.; Permyakov, E. A.; Abramov, V. M., Zn(2+)-mediated structure formation and compaction of the "natively unfolded" human prothymosin alpha. *Biochem. Biophys. Res. Commun.* **2000,** 267, (2), 663-8.
15. Uversky, V. N.; Gillespie, J. R.; Millett, I. S.; Khodyakova, A. V.; Vasiliev, A. M.; Chernovskaya, T. V.; Vasilenko, R. N.; Kozlovskaya, G. D.; Dolgikh, D. A.; Fink, A. L.; Doniach, S.; Abramov, V. M., Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH. *Biochemistry* **1999,** 38, 15009-15016.

16. Uversky, V. N.; Kirkitadze, M. D.; Narizhneva, N. V.; Potekhin, S. A.; Tomashevski, A., Structural properties of alpha-fetoprotein from human cord serum: the protein molecule at low pH possesses all the properties of the molten globule. *FEBS Lett.* **1995,** 364, (2), 165-7.
17. Uversky, V. N.; Li, J.; Fink, A. L., Trimethylamine-N-oxide-induced folding of alpha-synuclein. *FEBS Lett.*. **2001,** 509, (1), 31-35.
18. Uversky, V. N.; Li, J.; Fink, A. L., Evidence for a partially folded intermediate in alpha-synuclein fibril formation. *J. Biol. Chem.* **2001,** 276, (14), 10737-10744.
19. Uversky, V. N.; Li, J.; Souillac, P.; Jakes, R.; Goedert, M.; Fink, A. L., Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma- synucleins. *J. Biol. Chem.* **2002,** 25, 25.
20. Uversky, V. N.; Narizhneva, N. V.; Ivanova, T. V.; Kirkitadze, M. D.; Yu, A., Ligand-free form of human alpha-fetoprotein: Evidence for the Molten Globule State. *FEBS Lett.* **1997,** 410, 280-284.
21. Uversky, V. N.; Ptitsyn, O. B., All-or-none solvent-induced transitions between native, molten globule and unfolded states in globular proteins. *Fold Des* **1996,** 1, (2), 117-22.
22. Uversky, V. N.; Yamin, G.; Souillac, P. O.; Goers, J.; Glaser, C. B.; Fink, A. L., Methionine oxidation inhibits fibrillation of human α-synuclein *in vitro*. *manuscript* **2002**.
23. Tompa, P., Intrinsically unstructured proteins. *Trends Biochem Sci* **2002,** 27, (527-533).
24. Dedmon, M. M.; Patel, C. N.; Young, G. B.; Pielak, G. J., FlgM gains structure in living cells. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** 99, (20), 12681-12684.
25. Dunker, A. K.; Obradovic, Z.; Romero, P.; Garner, E. C.; Brown, C. J., Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **2000,** 11, 161-171.
26. Uversky, V. N., Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol Life Sci* **2003,** 60, (9), 1852-71.
27. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K., Comparing and Combining Predictors of Mostly Disordered Proteins,Ä†. *Biochemistry* **2005,** 44, (6), 1989-2000.
28. Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T., Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **2004,** 337, 635-645.
29. Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, Protein disorder prediction: implications for structural proteomics. . *Structure. 11:1453-1459* **2003**.
30. Liu, J.; Rost, B., Comparing function and structure between entire proteomes. *Protein Sci.* **2001,** 10, (10), 1970-1979.
31. Vucetic, S.; Brown, C. J.; Dunker, A. K.; Obradovic, Z., Flavors of protein disorder. *Proteins* **2003,** 52, 573-584.
32. Aravind, L.; Iyer, L. M.; Wellems, T. E.; Miller, L. H., Plasmodium Biology: Genomic Gleanings. *Cell* **2003,** 115, (7), 771-785.

33. Pizzi, E.; Frontali, C., Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Res.*. **2001,** 11, (2), 218-229.
34. Feng, Z.-P.; Zhang, X.; Han, P.; Arora, N.; Anders, R. F.; Norton, R. S., Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. *Molecular and Biochemical Parasitology* **2006,** 150, (2), 256-267.
35. Uversky V.N., O., C., Dunker, A.K. , Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signalling. . *J. Mol. Recognition 18 (5) 343-384.* **2005**.
36. Schulz, G. E., Nucleotide Binding Proteins. In *Molecular Mechanism of Biological Recognition*, Balaban, M., Ed. Elsevier/North-Holland Biomedical Press: New York, 1979; pp 79-94.
37. Alber, T.; Gilbert, W. A.; Ponzi, D. R.; Petsko, G. A., The role of mobility in the substrate binding and catalytic machinery of enzymes. *Ciba Found. Symp.* **1982,** 93, 4-24.
38. Spolar, R. S.; Record II, M. T., Coupling of local folding to site-specific binding of proteins to DNA. *Science* **1994,** 263, 777-784.
39. Lewis, M.; Chang, G.; Horton, N. C.; Kercher, M. A.; Pace, H. C.; Schumacher, M. A.; Brennan, R. G.; Lu, P., Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **1996,** 271, 1247-1254.
40. Dunker, A. K.; Obradovic, Z.; Romero, P.; Kissinger, C.; Villafranca, E. J., On the importance of being disordered. *Protein Data Bank Quarterly Newsletter* **1997,** 81, 3-5.
41. Dunker, A. K.; Obradovic, Z., The protein trinity - linking function and disorder. *Nat. Biotechnol.* **2001,** 19, (9), 805-806.
42. Wright P. E., a. D., H. J. , Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol. ,* **1999**, (293), 321-331.
43. Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradovic, Z.; Dunker, A. K., Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **2002,** 323, 573-584.
44. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Romero, P.; Uversky, V. N.; Dunker, A. K., Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **2005,** 44, (37), 12454-70.
45. Dyson, H. J.; Wright, P. E., Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **2002,** 12, (1), 54-60.
46. Romero, P. R.; Zaidi, S.; Fang, Y. Y.; Uversky, V. N.; Radivojac, P.; Oldfield, C. J.; Cortese, M. S.; Sickmeier, M.; LeGall, T.; Obradovic, Z.; Dunker, A. K., Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences* **2006,** 103, (22), 8390-8395.
47. Christopher, T. W.; Sylvie, G.-T.; Gregory, J. G., Jr., Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. *Angewandte Chemie International Edition* **2005,** 44, (45), 7342-7372.
48. Mersfelder, E. L.; Parthun, M. R., The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucl. Acids Res.* **2006,** 34, (9), 2653-2662.

49. Cohen, P., The regulation of protein function by multisite phosphorylation - a 25 year update. *Trends in Biochemical Sciences* **2000,** 25, (12), 596-601.
50. Tyers, M.; Jorgensen, P., Proteolysis and the cell cycle: with this RING I do thee destroy. *Current Opinion in Genetics & Development* **2000,** 10, (1), 54-64.
51. Scheffner M, N. U., Huibregtse JM., Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. *Nature. 1995 Jan 5;373(6509):81-3* **1995**.
52. Punga, T.; Bengoechea-Alonso, M. T.; Ericsson, J., Phosphorylation and Ubiquitination of the Transcription Factor Sterol Regulatory Element-binding Protein-1 in Response to DNA Binding. *Journal of Biological Chemistry* **2006,** 281, (35), 25278-25286.
53. Wright, P. E.; Dyson, H. J., Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999,** 293, 321-331.
54. Fink, A. L., Natively unfolded proteins. *Curr Opin Struct Biol* **2005,** 15, (1), 35-41.
55. Radivojac, P.; Iakoucheva, L. M.; Oldfield, C. J.; Obradovic, Z.; Uversky, V. N.; Dunker, A. K., Intrinsic disorder and functional proteomics. *Biophys J* **2007,** 92, (5), 1439-56.
56. Rose, G. D., *Unfolded Proteins*. Academic Press: New York, 2002; Vol. 62.
57. Vucetic, S.; Xie, H.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; Uversky, V. N., Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* **2007,** 6, (5), 1899-916.
58. Xie, H.; Vucetic, S.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Obradovic, Z.; Uversky, V. N., Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* **2007,** 6, (5), 1917-32.
59. Xie, H.; Vucetic, S.; Iakoucheva, L. M.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N.; Obradovic, Z., Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* **2007,** 6, (5), 1882-98.
60. Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P., Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **2004,** 338, (5), 1015-26.
61. Hilser, V. J.; Thompson, E. B., Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc Natl Acad Sci U S A* **2007,** 104, (20), 8311-5.
62. Romero, P.; Obradovic, Z.; Kissinger, C. R.; Villafranca, J. E.; Dunker, A. K., Identifying disordered regions in proteins from amino acid sequences. *IEEE Int. Conf. Neural Netw.* **1997,** 1, 90-95.
63. Romero, P.; Obradovic, Z.; Kissinger, C. R.; Villafranca, J. E.; Guilliot, S.; Garner, E.; Dunker, A. K., Thousands of proteins likely to have long disordered regions. *Pac. Symp. Biocomput.* **1998,** 3, 437-448.
64. Ferron, F.; Longhi, S.; Canard, B.; Karlin, D., A practical overview of protein disorder prediction methods. *Proteins* **2006,** 65, (1), 1-14.

65. Huber, R., Flexibility and rigidity, requirements for the function of proteins and protein pigment complexes. Eleventh Keilin memorial lecture. *Biochem. Soc. Trans.* **1987,** 15, (6), 1009-20.
66. Huber, R.; Bennett, W. S., Jr., Functional significance of flexibility in proteins. *Biopolymers* **1983,** 22, (1), 261-279.
67. Douzou, P.; Petsko, G. A., Proteins at work: "stop-action" pictures at subzero temperatures. *Adv. Protein Chem.* **1984,** 36, 245-361.
68. Aviles, F. J.; Chapman, G. E.; Kneale, G. G.; Crane-Robinson, C.; Bradbury, E. M., The conformation of histone H5. Isolation and characterisation of the globular segment. *Eur. J. Biochem.* **1978,** 88, 363-371.
69. Muchmore, S. W.; Sattler, M.; Liang, H.; Meadows, R. P.; Harlan, J. E.; Yoon, H. S.; Nettesheim, D.; Chang, B. S.; Thompson, C. B.; Wong, S. L.; Ng, S. L.; Fesik, S. W., X-ray and NMR structure of human Bcl-x_L, an inhibitor of programmed cell death. *Nature* **1996,** 381, 335-341.
70. Chang, B. S.; Minn, A. J.; Muchmore, S. W.; Fesik, S. W.; Thompson, C. B., Identification of a novel regulatory domain in Bcl-X(L) and Bcl-2. *EMBO J.*. **1997,** 16, (5), 968-977.
71. Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E., Structural studies of $p21^{Waf1/Cip1/Sdi1}$ in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U. S. A.* **1996,** 93, 11504-11509.
72. Ishima, R.; Torchia, D. A., Protein dynamics from NMR. *Nat. Struct. Biol.* **2000,** 7, (9), 740-743.
73. Bracken, C., NMR spin relaxation methods for characterization of disorder and folding in proteins. *J. Mol. Graph. Model.* **2001,** 19, (1), 3-12.
74. Smith, L. J.; Dobson, C. M.; van Gunsteren, W. F., Side-chain conformational disorder in a molten globule: molecular dynamics simulations of the A-state of human alpha-lactalbumin. *J. Mol. Biol.* **1999,** 286, (5), 1567-1580.
75. Fasman, G. D., *Circular dichroism and the conformational analysis of biomolecules*. Plenem Press: New York, 1996.
76. Dolgikh, D. A.; Gilmanshin, R. I.; Brazhnikov, E. V.; Bychkova, V. E.; Semisotnov, G. V.; Venyaminov, S.; Ptitsyn, O. B., Alpha-Lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett.*. **1981,** 136, 311-315.
77. Ohgushi, M.; Wada, A., 'Molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Lett.*. **1983,** 164, 21-24.
78. Kuwajima, K., A folding model of alpha-lactalbumin deduced from the three-state denaturation mechanism. *J. Mol. Biol.* **1977,** 114, (2), 241-258.
79. Hubbard, S. J.; Beynon, R. J.; Thornton, J. M., Assessment of conformational parameters as predictors of limited proteolytic sites in native protein structures. *Protein Eng.*. **1998,** 11, 349-359.
80. Fontana, A.; de Laureto, P. P.; de Filippis, V.; Scaramella, E.; Zambonin, M., Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.* **1997,** 2, R17-R26.
81. Fontana, A.; Zambonin, M.; Polverino de Laureto, P.; De Filippis, V.; Clementi, A.; Scaramella, E., Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.* **1997,** 266, 223-230.

82. Kissinger, C. R.; Parge, H. E.; Knighton, D. R.; Lewis, C. T.; Pelletier, L. A.; Tempczyk, A.; Kalish, V. J.; Tucker, K. D.; Showalter, R. E.; Moomaw, E. W.; Gastinel, L. N.; Habuka, N.; Chen, X.; Maldanado, F.; Barker, J. E.; Bacquet, R.; Villafranca, J. E., Crystal structures of human calcineurin and the human FKBP12-FK506- calcineurin complex. *Nature* **1995,** 378, 641-644.
83. Manalan, A. S.; Klee, C. B., Activation of calcineurin by limited proteolysis. *Proc. Natl. Acad. Sci. U. S. A.* **1983,** 80, 4291-4295.
84. Kriwacki, R. W.; Wu, J.; Tennant, L.; Wright, P. E.; Siuzdak, G., Probing protein structure using biochemical and biophysical methods. Proteolysis, matrix-assisted laser desorption/ionization mass spectrometry, high-performance liquid chromatography and size-exclusion chromatography of $p21^{Waf1/Cip1/Sdi1}$. *J. Chromatogr. A* **1997,** 777, 23-30.
85. Weinreb, P. H.; Zhen, W.; Poon, A. W.; Conway, K. A.; Lansbury, P. T., Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry* **1996,** 35, 13709-13715.
86. Adler, A. J.; Greenfield, N. J.; Fasman, G. D., Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol* **1973,** 27, 675-735.
87. Smyth, E.; Syme, C. D.; Blanch, E. W.; Hecht, L.; Vasak, M.; Barron, L. D., Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers* **2001,** 58, (2), 138-51.
88. Uversky, V. N., A multiparametric approach to studies of self-organization of globular proteins. *Biochemistry (Mosc)* **1999,** 64, (3), 250-66.
89. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker. . , Predicting intrinsic disorder from amino acid sequence. . *Proteins. 53:566-572* **2003**.
90. Radivojac, P., Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker, Protein flexibility and intrinsic disorder. *Protein Sci. 13:71-80* **2004**.
91. Peng, K., P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics. 7:208* **2006**.
92. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A. K., Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **2005,** 61 Suppl 7, 176-82.
93. Linding, R., R. B. Russell, V. Neduva, and T. J. Gibson. , GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res. 31:3701-3708* **2003**.
94. Jones, D. T., and J. J. Ward, Prediction of disordered regions in proteins from position specific score matrices. *Proteins. 53:566-572* **2003**.
95. Liu, J., and B. Rost. . , NORSp: predictions of long regions without regular secondary structure. . *Nucleic Acids Res. 31:3833-3835* **2003**.
96. Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; J., G. T.; Russell, R. B., Protein disorder prediction: implications for structural proteomics. *Structure* **2003,** 11, (11), 1453-1459.
97. Dosztanyi, Z., V. Csizmok, P. Tompa, and I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. . *Bioinformatics. 21:3433-3434* **2005**.

98. Yang, Z. R., R. Thomson, P. McNeil, and R. M. Esnouf. . . , RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics. 21:3369-3376* **2005**.
99. Prilusky, J., C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman, FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics. 21:3435-3438* **2005**.
100. Cheng, J.; Sweredoski, M. J.; Baldi, P., Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* **2005,** 11, (3), 213-222.
101. Coeytaux, K., and A. Poupon Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics. 21:1891-1900* **2005**.
102. Gu, J., M. Gribskov, and P. E. Bourne, Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput. Biol. 2:e90* **2006**.
103. Melamud, E.; Moult, J., Evaluation of disorder predictions in CASP5. *Proteins* **2003,** 53 Suppl 6, 561-5.
104. Jin, Y.; Dunbrack, R. L., Jr., Assessment of disorder predictions in CASP6. *Proteins* **2005,** 61 Suppl 7, 167-75.
105. Lorenza Bordoli, F. K. T. S., Assessment of disorder predictions in CASP7. *Proteins: Structure, Function, and Bioinformatics* **2007,** 69, (S8), 129-136.
106. Bordoli, L.; Kiefer, F.; Schwede, T., Assessment of disorder predictions in CASP7. *Proteins* **2007,** 69 Suppl 8, 129-36.
107. Mohan, A.; Oldfield, C. J.; Radivojac, P.; Vacic, V.; Cortese, M. S.; Dunker, A. K.; Uversky, V. N., Analysis of molecular recognition features (MoRFs). *J Mol Biol* **2006,** 362, (5), 1043-59.
108. Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G., Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2000,** 97, (16), 8868-73.
109. Strong M, S. M. R. W. S., Philips M. , Cascio D, Eisenberg D, Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *PNAS* **2006,** 103:8060-8065.
110. Predrag, R.; Vladimir, V.; Chad, H.; Ross, R. C.; Amrita, M.; Joshua, W. H.; Mark, G. G.; Lilia, M. I., Identification, analysis, and prediction of protein ubiquitination sites. *Proteins: Structure, Function, and Bioinformatics* **2009,** 9999, (9999), NA.
111. Tompa, P.; Prilusky, J.; Silman, I.; Sussman, J. L., Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins: Structure, Function, and Bioinformatics* **2008,** 71, (2), 903-909.
112. Gsponer, J.; Futschik, M. E.; Teichmann, S. A.; Babu, M. M., Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science* **2008,** 322, (5906), 1365-1368.
113. Prakash, S.; Tian, L.; Ratliff, K. S.; Lehotzky, R. E.; Matouschek, A., An unstructured initiation site is required for efficient proteasome-mediated degradation. *Nat Struct Mol Biol* **2004,** 11, (9), 830-837.
114. Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Dunker, A. K., The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research* **2004,** 32, (3), 1037-1049.

115. Collins, M. O.; Yu, L.; Campuzano, I.; Grant, S. G. N.; Choudhary, J. S., Phosphoproteomic Analysis of the Mouse Brain Cytosol Reveals a Predominance of Protein Phosphorylation in Regions of Intrinsic Sequence Disorder. *Mol Cell Proteomics* **2008,** 7, (7), 1331-1348.
116. L, B., Low-complexity regions in Plasmodium proteins: in search of a function. *Genome Research* **2001,** 11(2), 195-197.
117. Mehlin, C.; Boni, E.; Buckner, F. S.; Engel, L.; Feist, T.; Gelb, M. H.; Haji, L.; Kim, D.; Liu, C.; Mueller, N.; Myler, P. J.; Reddy, J. T.; Sampson, J. N.; Subramanian, E.; Van Voorhis, W. C.; Worthey, E.; Zucker, F.; Hol, W. G. J., Heterologous expression of proteins from Plasmodium falciparum: Results from 1000 genes. *Molecular and Biochemical Parasitology* **2006,** 148, (2), 144-160.
118. Rhodes, G., *Crystallography made crystal clear: a guide for users of macromolecular models*. Academic Press: San Diego, 1993.
119. Blundell, T. L.; Johnson, L. N., *Protein Crystallography*. Academic Press: New York, 1976; p xiv, 565 p. : ill. ; 24 cm.
120. Zurdo, J.; Gonzalez, C.; Sanz, J. M.; Rico, M.; Remacha, M.; Ballesta, J. P., Structural differences between Saccharomyces cerevisiae ribosomal stalk proteins P1 and P2 support their functional diversity. *Biochemistry* **2000,** 39, (30), 8935-8943.
121. Palaninathan, S. K.; Mohamedmohaideen, N. N.; Snee, W. C.; Kelly, J. W.; Sacchettini, J. C., Structural insight into pH-induced conformational changes within the native human transthyretin tetramer. *J Mol Biol* **2008,** 382, (5), 1157-67.
122. Zhang, Y.; Stec, B.; Godzik, A., Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure* **2007,** 15, (9), 1141-7.
123. Eyal, E.; Gerzon, S.; Potapov, V.; Edelman, M.; Sobolev, V., The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J Mol Biol* **2005,** 351, (2), 431-42.
124. Mohan, A.; Uversky, V. N.; Radivojac, P., Influence of Sequence Changes and Environment on Intrinsically Disordered Proteins. *PLoS Comput Biol* **2009,** 5, (9), e1000497.
125. Sander, C.; Schneider, R., Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991,** 9, (1), 56-68.
126. Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O.; Ofran, Y., Automatic prediction of protein function. *Cell Mol Life Sci* **2003,** 60, (12), 2637-2650.
127. Joachims, T., *Learning to classify text using support vector machines: methods, theory, and algorithms*. Kluwer Academic Publishers: 2002.
128. Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K., Sequence complexity of disordered protein. *Proteins* **2001,** 42, 38-48.
129. Wootton, J. C., Statistic of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **1993,** 17, 149-163.
130. Van Eldik, L. J.; Watterson, D. M., *Calmodulin and signal transduction* Academic Press: San Diego, 1998.
131. Predrag, R.; Slobodan, V.; Timothy, R. O. C.; Vladimir, N. U.; Zoran, O.; Dunker, A. K., Calmodulin signaling: Analysis and prediction of a disorder-dependent molecular recognition. *Proteins: Structure, Function, and Bioinformatics* **2006,** 63, (2), 398-410.

132. Rost, B., Twilight zone of protein sequence alignments. *Protein Eng.* **1999,** 12, (2), 85-94.
133. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank. *Nucleic Acids Res.* **2000,** 28, (1), 235-242.
134. Kabsch, W.; Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983,** 22, (12), 2577-2637.
135. Vacic, V.; Oldfield, C. J.; Mohan, A.; Radivojac, P.; Cortese, M. S.; Uversky, V. N.; Dunker, A. K., Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* **2007,** 6, (6), 2351-66.
136. Jones, S.; Thornton, J. M., Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol* **1997,** 272, (1), 133-43.
137. Jones, S.; Thornton, J. M., Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **1997,** 272, (1), 121-32.
138. Conte, L. L.; Chothia, C.; Janin, J., The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology* **1999,** 285, (5), 2177-2198.
139. Connolly, M. L., The molecular surface package. *J. Mol. Graph.* **1993,** 11, (2), 139-141.
140. Ko, L. J.; Prives, C., p53: puzzle and paradigm. *Genes Dev* **1996,** 10, (9), 1054-1072.
141. Rustandi, R. R.; Baldisseri, D. M.; Weber, D. J., Structure of the negative regulatory domain of p53 bound to S100B(ββ). *Nat. Struct. Biol.* **2000,** 7, (7), 570-574.
142. Lee, H.; Mok, K. H.; Muhandiram, R.; Park, K. H.; Suk, J. E.; Kim, D. H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K. H., Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J Biol Chem* **2000,** 275, (38), 29426-32.
143. Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P., Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain [comment]. *Science* **1996,** 274, (5289), 948-953.
144. Oliner, J. D., Pietenpol, J. A., Thiagalingam, S.,Gyuris, J., Kinzler, K. W. & Vogelstein, B. , Oncoprotein MDM2 conceals the activation domain of tumour suppressor p53. . *Nature* **1993,** 362, 857–860.
145. Schulman, B. A., Lindstrom, D. L. & Harlow, E., Substrate recruitment to cyclin-dependent kinase 2 by a multipurpose docking site on cyclin A. *Proc. Natl Acad. Sci.* **1998,** 95, 10453–10458.
146. Baudier, J., Delphin, C., Grunwald, D., Khochbin, S.& Lawrence, J. J. , Characterization of the tumor suppressor protein p53 as a protein kinase C substrate and a S100b-binding protein. . *Proc. Natl Acad. Sci. USA,* **1992,** 89, 11627–11631.
147. Wilder, P. T., Rustandi, R. R., Drohat, A. C. & Weber,D. J. , S100B(betabeta) inhibits the protein kinase C-dependent phosphorylation of a peptide derived from p53 in a Ca2+-dependent manner. . *Protein Sci.* **1998,** 7, 794–798.
148. Machesky, L. M. I., R. H. , Signaling to actin dynamics. . *J. Cell Biol.* **1999,** 146, 267–272.

149. Chereau, D., Kerff, F., Graceffa, P., Grabarek, Z.,Langsetmo, K. & Dominguez, R. , Actin-bound structures of Wiskott-Aldrich syndrome protein (WASP)-homology domain 2 and the implications for filament assembly. *Proc. Natl Acad. Sci. USA,* **2005,** 102, 16644–16649.
150. Kim, A. S.; Kakalis, L. T.; Abdul-Manan, N.; Liu, G. A.; Rosen, M. K., Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature* **2000,** 404, (6774), 151-158.
151. Fiser A, D. Z., Simon I., The role of long-range interactions in defining the secondary structure of proteins is overestimated. *Comput Appl Biosci.* **1997** Jun;13(3):297-301.
152. Kihara, D., The effect of long-range interactions on the secondary structure formation of proteins. *Protein Science* **2005,** 14, (8), 1955-1963.
153. Rost, B., PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **1996,** 266, 525-539.
154. Romero, P., Z. Obradovic, X. Li, E. Garner, C. Brown, and A. K. Dunker, Sequence complexity of disordered protein. *Proteins: Struct. Funct. Gen* **2001,** 2001, 42:38-48.
155. Thornton, J. M., Disulphide bridges in globular proteins. *Journal of Molecular Biology* **1981,** 151, (2), 261-287.
156. Uversky, V.; Gillespie, J.; A., F., Why are "natively unfolded" proteins unstructured under physiological conditions? . *Proteins: Struct. Funct. Gen., 41(3): 415-427.* **2000**.
157. Burley, S. K.; Petsko, G. A., Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **1985,** 229, 23-28.
158. Gunasekaran, K.; Tsai, C.-J.; Nussinov, R., Analysis of Ordered and Disordered Protein Complexes Reveals Structural Features Discriminating Between Stable and Unstable Monomers. *Journal of Molecular Biology* **2004,** 341, (5), 1327-1341.
159. Li, X.; Romero, P.; Rani, M.; Dunker, A. K.; Obradovic, Z., Predicting protein disorder for N-, C-, and internal regions. *Genome Inform Ser Workshop Genome Inform* **1999,** 10, 30-40.
160. Romero, P.; Obradovic, Z.; AK., D., Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics* **1997,** 8, 110-124.
161. Garner, E.; Romero, P.; Dunker, A. K.; Brown, C.; Obradovic, Z., Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.* **1999,** 10, 41-50.
162. Jones, S.; Thornton, J., Analysis of protein-proteins interaction sites using surface patches. *J Molecular Biology* **1997,** 272, 121 - 132.
163. Oda, Y.; Nagasu, T.; Chait, B. T., Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **2001,** 19, (4), 379-382.
164. Zhou, H.; Watts, J. D.; Aebersold, R., A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **2001,** 19, (4), 375-378.
165. Roepstorff, P., Mass spectrometry in protein studies from genome to function. *Current Opinion in Biotechnology* **1997,** 8, (1), 6-13.

166. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-67.
167. Eng, J. K.; McCormack, A. L.; Yates Iii, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **1994,** 5, (11), 976-989.
168. Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **1995,** 67, (8), 1426-36.
169. Yates, J. R., 3rd, Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct* **2004,** 33, 297-316.
170. Lu, B.; Ruse, C.; Yates, J., Colander: A Probability-Based Support Vector Machine Algorithm for Automatic Screening for CID Spectra of Phosphopeptides Prior to Database Search. *J. Proteome Res.* **2008**.
171. Alves, P.; Arnold, R. J.; Novotny, M. V.; Radivojac, P.; Reilly, J. P.; Tang, H., Advancements in protein identification from shotgun proteomics using predicted peptide detectability. *Pac Symp Biocomput* **2007,** 12, 409-420.
172. Tang, H.; Arnold, R. J.; Alves, P.; Xun, Z.; Clemmer, D. E.; Novotny, M. V.; Reilly, J. P.; Radivojac, P., A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **2006,** 22, (14), e481-e488.
173. Villén J, B. S. A., Gerber S.A, Gygi S.P., Large-scale phosphorylation analysis of mouse liver. *PNAS* **2007** 104:1488-1493.
174. Du, K.; Herzig, S.; Kulkarni, R. N.; Montminy, M., TRB3: A tribbles Homolog That Inhibits Akt/PKB Activation by Insulin in Liver. *Science* **2003,** 300, (5625), 1574-1577.
175. Joachims, T. In *A support vector method for multivariate performance measures*, International Coference on Machine Learning (ICML), Bonn, Germany, 2005; Bonn, Germany, 2005.
176. Adam, R. D., Biology of Giardia lamblia. *Clin. Microbiol. Rev.* **2001,** 14, (3), 447-475.
177. Stanley, J. S. L., Amoebiasis. *The Lancet* **2003,** 361, (9362), 1025-1034.
178. Teixeira, A. R. L.; Nitz, N.; Guimaro, M. C.; Gomes, C.; Santos-Buch, C. A., Chagas disease. *Postgrad Med J* **2006,** 82, (974), 788-798.
179. Hoxie, N. J.; Davis, J. P.; Vergeront, J. M.; Nashold, R. D.; Blair, K. A., Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin. *Am J Public Health* **1997,** 87, (12), 2032-2035.
180. Flegr, J., Effects of Tozoplasma on human behavior. . *Schizophr Bull.* **2007,** doi:10.1093/schbul/sbl074.
181. Rorman, E.; Zamir, C. S.; Rilkis, I.; Ben-David, H., Congenital toxoplasmosis--prenatal aspects of Toxoplasma gondii infection. *Reproductive Toxicology* **2006,** 21, (4), 458-472.
182. Yolken, R. H.; Rouslanova, I.; Lillehoj, E.; Ford, G.; Torrey, E. F.; Bachmann, S.; Schroeder, J., Antibodies to Toxoplasma gondii in Individuals with First-Episode Schizophrenia. *Clinical Infectious Diseases* **2001,** 32, (5), 842-844.

183. Guo, X. H.; Zhao, N. M.; Chen, S. H.; Teixeira, J., Small-angle neutron scattering study of the structure of protein/detergent complexes. *Biopolymers* **1990,** 29, (2), 335-46.
184. Snow, R. W.; Craig, M. H.; Deichmann, U.; le Sueur, D., A Preliminary Continental Risk Map for Malaria Mortality among African Children. *Parasitology Today* **1999,** 15, (3), 99-104.
185. S. Y. Wong and J. S. Remington, Biology of Toxoplasma gondii . . *AIDS,* **1993, ,** 7, 299.
186. Slifko, T. R.; Smith, H. V.; Rose, J. B., Emerging parasite zoonoses associated with water and food. *International Journal for Parasitology* **2000,** 30, (12-13), 1379-1393.
187. Meissner. M., A.-N. C. a. S., Jr W. J. , Molecular tools for analysis of gene function in parasitic microorganisms. *Applied Microbiology & Biotechnology* **2007,** 75, (5), 963-975.
188. Gardner, M. J.; Hall, N., Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **2002,** 419, (6906), 498.
189. Bahl, A.; Brunk, B.; Coppel, R. L.; Crabtree, J.; Diskin, S. J.; Fraunholz, M. J.; Grant, G. R.; Gupta, D.; Huestis, R. L.; Kissinger, J. C.; Labo, P.; Li, L.; McWeeney, S. K.; Milgram, A. J.; Roos, D. S.; Schug, J.; Stoeckert, C. J., Jr, PlasmoDB: the Plasmodium genome resource. An integrated database providing tools for accessing, analyzing and mapping expression and sequence data (both finished and unfinished). *Nucl. Acids Res.* **2002,** 30, (1), 87-90.
190. Bahl, A., PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* **2003,** 31, (1), 212-215
191. Kissinger, J.; Gajria, B.; Li, L.; Paulsen, I. T.; Roos, D. S., ToxoDB: accessing the Toxoplasma gondii genome. *Nucleic Acids Research* **2003,** 31, (1), 234-236.
192. Gardner MJ, B. R., Shah T, de Villiers EP, Carlton JM, Hall N, Ren Q, Paulsen IT, Pain A, Berriman M, Wilson RJ, Sato S, Ralph SA, Mann DJ, Xiong Z, Shallom SJ, Weidman J, Jiang L, Lynn J, Weaver B, Shoaibi A, Domingo AR, Wasawo D, Crabtree J, Wortman JR, Haas B, Angiuoli SV, Creasy TH, Lu C, Suh B, Silva JC, Utterback TR, Feldblyum TV, Pertea M, Allen J, Nierman WC, Taracha EL, Salzberg SL, White OR, Fitzhugh HA, Morzaria S, Venter JC, Fraser CM, Nene V. , Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes.
. *Science.* **2005 Jul 1,** ;309(5731):134-7.
193. Xu P, W. G., Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA.; hominis., T. g. o. C., *Nature.* **2004 Oct 28;.**
, 431(7012):1107-12.
194. Abrahamsen MS, T. T., Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck GA, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V., Complete genome sequence of the apicomplexan, Cryptosporidium parvum. *Science.* **2004 Apr 16,** 304(5669):441-5.
195. Geer, R. C. a. S., E.W. , Entrez: Making use of its power. *Briefings in Bioinformatics.* **2003,** June;4(2):1779-184.

196. McArthur, A.; HG, M.; JE, N.; NQ, P.; U, K.; G, H.; MK, C.; ME, H.; R, F.; CI, R.; GE, O.; SB, A.; RD, A.; FD, G.; ML, S., The Giardia lamblia Genome Database In 2000.
197. Vacic, V.; Uversky, V.; Dunker, A. K.; Lonardi, S., Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* **2007,** 8, (1), 211.
198. Vucetic, S.; Obradovic, Z.; Vacic, V.; Radivojac, P.; Peng, K.; Iakoucheva, L. M.; Cortese, M. S.; Lawson, J. D.; Brown, C. J.; Sikes, J. G.; Newton, C. D.; Dunker, A. K., DisProt: a database of protein disorder. *Bioinformatics* **2005,** 21, (1), 137-140.
199. Vihinen, M., Relationship of protein flexibility to thermostability. *Protein Eng.* **1987,** 1, (6), 477-80.
200. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K., Comparing and Combining Predictors of Mostly Disordered Proteins. *Biochemistry* **2005,** 44, (6), 1989-2000.
201. Wuchty, S.; Ipsaro, J. J., A Draft of Protein Interactions in the Malaria Parasite P. falciparum. *J. Proteome Res.* **2007,** 6, (4), 1461-1470.
202. Mohan, A.; Jr, W. J. S.; Radivojac, P.; Dunker, A. K.; Uversky, V. N., Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Molecular BioSystems* **2008,** 4, (4), 328-340.
203. K. P. Ng, G. P., R. O. Savene, C. T. Denny, V. N. Uversky; and K. A. Lee, Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of EWS family oncoproteins. *Proc. Natl. Acad. Sci. U. S. A.,* **2007,** 104, 479.
204. Sigalov, A.; Aivazian, D.; Stern, L., Homooligomerization of the Cytoplasmic Domain of the T Cell Receptor Œ∂ Chain and of Other Proteins Containing the Immunoreceptor Tyrosine-Based Activation Motif‚Ä†. *Biochemistry* **2004,** 43, (7), 2049-2061.
205. Sigalov, A. B., Immune cell signaling: a novel mechanistic model reveals new therapeutic targets. *Trends in Pharmacological Sciences* **2006,** 27, (10), 518-524.
206. Cheng, Y.; LeGall, T.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N., Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* **2006,** 45, (35), 10448-60.
207. McGuffin, L. J., Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics* **2008,** 24, (16), 1798-804.
208. Tsai CJ, K. S., Ma B, Nussinov R Folding funnels, binding funnels, and protein function. . *Protein Sci* **1999,** 8: 1181–1190.
209. Ma B, N. R., Regulating highly dynamic unstructured proteins and their coding mRNAs. . *Genome Biol* **2009,** 10: 204.
210. Ma B, K. S., Tsai CJ, Nussinov R Folding funnels and binding mechanisms. . *Protein Eng* **1999,** 12: 713–720.
211. Uversky, V. N.; Cooper, E. M.; Bower, K. S.; Li, J.; Fink, A. L., Accelerated α–synuclein fibrillation in crowded milieu. *FEBS Lett.* **2001,** 515, 99-103.
212. Radivojac, P.; Baenziger, P. H.; Kann, M. G.; Mort, M. E.; Hahn, M. W.; Mooney, S. D., Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* **2008,** 24, (16), i241-7.

CURRICULUM VITAE

Amrita Mohan
162 W. 56th Street, #804.
New York, NY 10019 amrita.mohan@gmail.com

Education

'05 – '09	**Ph.D., Informatics**, Indiana University, USA
'03 – '05	**M.S, Bioinformatics**, Indiana University, USA
'99 – '03	**Bachelor of Information Technology**, University of Delhi, India

Honors/Awards

'07 – '09	Lilly Informatics Fellowship, Eli Lilly and Company Foundation.
'07	NSF (National Science Foundation) Travel Award
'07	Teaching Excellence Award, School of Informatics, Indiana University
'06 – '09	Iota Phi Nu (National Informatics Honors Society) member of alpha chapter.

Poster Presentations

Jun '08	*56th ASMS Conference on Mass Spectrometry and Allied Topics*: An novel method for computational phosphopeptide identification
Oct '07	*First Annual Midwest Symposium on Computational Biology and Bioinformatics*
May '07	*Fourth Annual Indiana Bioinformatics Conference:* The upper-bound of intrinsic disorder prediction
Mar '07	*Biophysical Society Meeting:* Structural variability of MoRFs/ Analysis of molecular recognition feature complexes
Sep '06	*20th Symposium of the Protein Society:* Sequence-Based Prediction of Hubs in Protein-Protein Interaction Networks
May '06	*Third Annual Indiana Bioinformatics Conference:* Has disorder prediction accuracy reached its limit? Almost / Molecular Recognition Features, MoRFs
Mar '06	*Biophysical Society Meeting:* Molecular Recognition Features, MoRFs
Jan '06	*Pacific Symposium on Biocomputing:* Intrinsically Disordered Hubs in C. elegans interactome
Sep '05	*Biochemistry and Molecular Biology Research Day:* eMoRFs – Experimentally derived Molecular Recognition Features; Visualization and analysis of MoRF binding surfaces
May '04	*First Annual Indiana Bioinformatics Conference:* MoREs: Their Predictability & Relationships with NORs

Work Experience

Summer '05 **Rosetta Inpharmatics (Merck & Co. Subsidiary), Seattle, US**
- *Intern*

Summer '02 **Institute of Advanced Biosciences - 'E-Cell Lab', Tsuruoka City, Japan**
- *Intern*

Jun '01 - Aug '02	**Institute of Genomics and Integrative Biology (IGIB), Delhi, India** • *Project Trainee*
Aug '05 -Aug '09	**School of Informatics and Com., Indiana University, Bloomington** • *Eli Lilly Discovery Informatics Research Fellow/Research Assistant/Associate Instructor*
Aug '03 - Aug '05	**Center for Computational Biology & Bioinformatics, Indiana Univ. Purdue Univ., Indianapolis** • *Research Assistant*
Jan – Jul '03	**Washington State University, Pullman, Washington** • *Project Trainee*

Teaching Experience

Spring '09	*Associate Instructor:* I210 – Information Infrastructure I
Fall '08	*Associate Instructor:* I500 – Fundamental Computer Concepts of Informatics
Spring '07	*Associate Instructor:* I211 – Information Infrastructure II
Fall '05,'06,	*Associate Instructor:* I500 – Fundamental Computer Concepts of Informatics

Journal Publications

'09	**A. Mohan**, V. Uversky, P. Radivojac: *Influence of Sequence Changes and Environment on Intrinsically Disordered Proteins* (*Plos CB*)
'09	P. Radivojac, V. Vacic , C. Haynes, R. R.Cocklin, **A. Mohan**, J. W. Heyen, M. G. Goebl, L.M. Iakoucheva: *Identification, Analysis and Prediction of Protein Ubiquitination Sites* (*Proteins*)
Unpublished	**A. Mohan,** R.J. Arnold, H. Tang, Q. Sheng, P. Radivojac**:** *A novel method for computational phosphopeptide identification*
Unpublished	W.T.Clark, **A. Mohan**, P. Radivojac: *Automated protein function annotation from primary structure*
'08	**A. Mohan,** W.J. Sullivan, P. Radivojac, A.K. Dunker, V. Uversky**:** *Intrinsic disorder in pathogenic and non-pathogenic microbes: Discovering and analyzing the unfoldomes of early-branching eukaryotes (Mol. BioSystems)*
'07	P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, **A. Mohan**, S. M. Boyle, S. D. Mooney*: An integrated approach to inferring gene-disease associations in humans.(Proteins: Structure, Function, and Bioinformatics)*
'07	V. Vacic, C. J. Oldfield, **A. Mohan,** P. Radivojac, M. S. Cortese, V. N. Uversky, A. K. Dunker: *Characterization of molecular recognition features, MoRFs, and their binding partners. (Journal of Proteome Research)*
'06	**A. Mohan**, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, V. N. Uversky: *Analysis of molecular recognition features (MoRFs). (Journal of Molecular Biology)*

Technical Skillset

- **Operating Systems:** WINDOWS, Mac OS X, LINUX/UNIX
- **Databases:** Oracle 8, MySQL
- **Programming/Scripting Languages:** Perl, Python, C, C++, C#
- **Other Software Tools:** MATLAB, SAS, Deep View