Molecular recognition of intrinsically disordered proteins: theory, predictions and applications

PhD Thesis

Bálint Mészáros

Supervisors:

Dr. Zsuzsanna Dosztányi, PhD

and

Prof. István Simon, PhD, DSc

Statistical Physics, Biological Physics and Physics of Quantum Systems
Doctoral Program, Physics Doctoral School

Institute of Physics, Eötvös Loránd University, Faculty of Science

Protein Structure Research Group, Institute of Enzymology

Research Centre for Natural Sciences, Hungarian Academy of Sciences

Budapest, Hungary

2012

Preface

This dissertation describes my work done between October 2007 and March 2012 in the Doctoral Program of Statistical Physics, Biological Physics and Physics of Quantum Systems in the Protein Structure Research Group of the Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, under the supervision of senior research fellow Dr. Zsuzsanna Dosztányi and group leader Prof. István Simon.

The dissertation is based on my results published in the following papers:

- Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* **5**(5):e1000376 (2009)
- Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**(20):2745-6 (2009)
- Dosztányi Z, Mészáros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform* **11**(2):225-43 (2010)
- Mészáros B, Simon I, Dosztányi Z. The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Phys Biol* **8**(3):035003 (2011)
- Mészáros B, Tóth J, Vértessy BG, Dosztányi Z, Simon I. Proteins with complex architecture as potential targets for drug design: a case study of Mycobacterium tuberculosis. *PLoS Comput Biol* **7**(7):e1002118 (2011)
- Pajkos M, Mészáros B, Simon I, Dosztányi Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst* **8**(1):296-307 (2012)
- Mészáros B, Dosztányi Z, Simon I. Disordered binding regions and linear motifs – two sides of the coin. *In preparation - 2012*

Acknowledgements

I am indebted to my supervisors, Zsuzsanna Dosztányi and István Simon for their support throughout my PhD studies. Without their guidance none of the work I present here could have been done.

I am grateful to all my past and present co-workers at the Protein Structure Research Group of the Institute of Enzymology. They provided an inspiring and pleasant environment that was a pleasure to work in.

I would like to thank Prof. László Budai, the head of the Institute of Enzymology for providing me the opportunity of working in the Institute.

I am thankful to my family and all my friends who supported and encouraged me in these past years.

Table of Contents

1. Introduction

The middle of the 20th century saw the advent of structural biology, a branch of modern biology that over the years undoubtedly had one of the largest effects on our view of life. Its importance is also illustrated by the fact that since the 1950's over a dozen Nobel Prizes were awarded to research directly linked to structural biology, including the structure determination of nucleic acids, proteins and their complexes and the ribosome. One of the key factors behind the long lasting success of the field is the successful integration of different scientific areas. The applied experiments, such as X-ray crystallography, NMR or electron-microscopy rely heavily on various fields of physics, including thermodynamics, statistical physics, quantum physics and electrodynamics. Parallelly, the theoretical description of molecular structures also has a strong physical background. However, the ultimate goal of every experiment is to deduce biologically meaningful statements ranging from the function of single molecules to a systems biology level. These statements on one hand shed light on how living organisms work, on the other hand provide means to translate this knowledge into practical applications, such as drug design. Over the past few decades, however, the amount of knowledge and data has reached a level unmanageable by manual methods. Accordingly, structural biology has been extensively computerized up to a point where much of the current research projects rely on bioinformatics methods, computer simulations and online databases.

The extensive review of structural biology – if at all possible – would fill volumes alone, hence clearly out of the scope of this dissertation. Instead, in the next few chapters I aim to give a very brief summary of the theoretical and practical background that is directly relevant to my work.

1.1. Protein folding

1.1.1. Levels of protein structure

Proteins are long, linear polymers, generally built up by the 20 standard amino acids. Sequencing experiments can provide the primary structure of protein chains by determining the order of its composing amino acids that are linked by covalent bonds. As the peptide bonds between consecutive residues are planar, the conformation of the main chain of a protein can be described by two angles (termed Φ and Ψ dihedral angles) per residue, describing the relative orientation of the two planes of two consecutive peptide bonds. Following the appearance of the first solved protein structures, it became clear that the distribution of dihedral angles in a protein is highly non-random. There are several preferred Φ-Ψ angle combinations that correspond to various local order in the structure. Consecutive residues with Φ-Ψ angles around (-60º,-45º) form α-helices which are stabilized by H-bonds between the main chain atoms of the i^{th} and $i+4^{th}$ residues. The other most common emerging local structure are given by Φ-Ψ angles around (-135º,135º) resulting in an extended β-strand conformation. Such extended structures are also stabilized by H-bonds between the main chain atoms of two or more strands in either parallel or anti-parallel orientation. Other frequently populated Φ-Ψ preferences give rise to a variety of other (not necessarily translationally symmetrical) local structural elements, such as turns, hairpins and other, less frequent types of helices. The type and position of such ordered, local structural elements in a protein chain constitute the second level of structure, the so-called secondary structure. The third level of protein structure is the full, 3 dimensional conformation of the whole protein chain. This is typically given by enumerating the coordinates of all the (heavy) atoms of the protein in an arbitrarily chosen orthogonal coordinate system. The main driving forces behind the organization of secondary and tertiary structures are diverse and encompass H-bonds, salt bridges, the covalent bonding of the S atom of cystein amino acids and entropic effects, such as the hydrophobic effect. Tertiary structures can be determined by either X-ray diffraction or NMR measurements. The resulting sets of coordinates for individual proteins and protein

complexes are deposited to and are publicly available in the Protein Data Bank (PDB[1]) database. The fourth level of protein structure describes the spatial orientation of proteins during their interactions. The main features of quaternary protein structures is described in later sections of the introduction (sections 1.2 and 1.5).

1.1.2. Physical description of protein folding

The process during which a polypeptide chain adopts its native tertiary structure is called folding. The typical timescale of protein folding is in the millisecond-second range[2] (depending mainly on proline and cysteine content), and as the temperature inside the cells of living organisms can be considered constant at this timescale, the correct choice of thermodynamic potential for the description of the protein folding problem is the Gibbs free energy (G). The Gibbs free energy of a protein chain can be broken down into two terms:

$$G_{protein} = H_{protein} - TS_{protein} \quad (1)$$

where $H_{protein}$ is the enthalpy and $S_{protein}$ is the entropy of the protein and T is the temperature. In the simplest, two state model of the protein folding process, the protein can exist in either the denatured/unfolded state and the folded state, corresponding to the conformation in which $G_{protein}$ is minimal. The equilibrium between the two states is determined by the following equation:

$$\begin{aligned} -RT\ln K &= \Delta G_{protein} = \\ &= \Delta H_{protein} + \Delta H_{solvent} + \Delta H_{protein-solvent} - T\Delta S_{protein} - T\Delta S_{solvent} \end{aligned} \quad (2)$$

where R is the gas constant, K is the equilibrium constant between the unfolded and the folded state (the joint entropic term of the protein and the solvent has been omitted from the equation as it is negligible compared to other terms). $\Delta G_{protein}$ determines the overall stability of the protein and the contribution of each term varies heavily between individual cases. Even this simple model is applicable to the basic description of the

folding of many proteins and can explain some hallmarks of known folded structures. For example, due to the last term that describes the entropy of the solvent, polar and charged groups of the protein are directed to the surface of the resulting structure, while hydrophobic sidechains form the hydrophobic core shielded from the polar solvent[3]. This hydrophobic effect is typically the strongest driving force in protein folding[4]. However, the simple two state model does not give information about the kinetics of the folding as it does not describe the intermediary conformations the protein goes through during folding. H-bonding (a part of $\Delta H_{protein}$) stabilizes emerging secondary structures and the hydrophobic effect (described by $\Delta S_{solvent}$) drives the hydrophobic collapse of the protein. The temporal order of the two effects during folding has been disputed and there are examples for both scenarios: hydrophobic collapse followed by the formation of secondary structures and vice versa[5].

Although two state folding models are a viable starting point in the description of protein folding, the detailed description of folding requires a more elaborate model where transient states have to be considered. Furthermore, in reality, the final, folded state shows fluctuations as well (captured in the B factors during X-ray crystallography) and the introduction of distinct states is necessary to describe alternative low energy states between which the folded protein alternates. As shown in later chapters, these states can have a profound effect on the function and binding of proteins.

The two state model can be expanded with the addition of folding intermediate states and some of these states can be measured experimentally by techniques with high temporal resolution. However, the full description of the folding kinetics would require the consideration of all possible states of the protein and the possible transitions between them.

1.1.3. The energy landscape view of proteins

The use of energy landscapes in the description of protein folding is an alternative approach that is aimed at describing all possible states of a protein[6, 7]. However, due to the large number of described conformations, this framework is rather qualitative, albeit it proved to be extremely useful in explaining some of the basic properties of protein folding[8, 9]. The energy landscape of a protein is the energy of each possible conformation as a function of the degrees of freedom, such as the dihedral angles along the polypeptide backbone (see Figure 1). The vertical axis represents the internal free energy. The internal free energy contains the enthalpic term, therefore it includes the contributions from hydrogen bonds, ion-pairs and torsion angle energies. Moreover, it also includes hydrophobic and solvation free energies by averaging over the conformational space of water molecules. However, it does not contain the conformational entropy term. Each conformation of the protein is represented by a point on the multidimensional surface, specified by a multidimensional set of coordinates in the conformational space. Conformations that are similar geometrically are close to each other on the surface. However, the energy of similar conformations can still differ significantly, and as a result there are many hills and valleys on this surface. The wider the valleys are, the more conformations are similar to the single conformation at a local energy minimum. Since the true multidimensionality of the surface and the vast number of conformations cannot be easily represented on a figure, usually a highly simplified schematic cartoon is used to illustrate the basic properties of proteins[7]. An example energy landscape of a typical folded protein is shown in Figure 1.

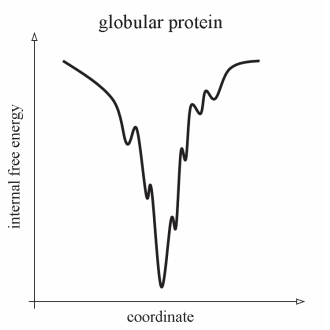

Figure 1: a typical globular protein energy landscape in two dimensions
The internal free energy is sketched against some coordinate representative of the conformation. The coordinate is arbitrary, however each different conformation should have a unique set of coordinates.

It was suggested that the energy landscape of a well-folded globular protein is funnel shaped[5-7]. Although a vast number of conformations are sampled, most of the conformations have high (unfavorable) energy. There are much fewer conformations that have low energy and these are similar to the native state. The bottom of the funnel represents the unique native structure that is stabilized by a large number of intramolecular interactions and by the burying of hydrophobic side chains. This image can also be used to illustrate how globular proteins find their native structure despite the huge conformational space. In 1969 Levinthal noted that it would take an astronomical time for a protein to search through its full conformational space by means of random walk - an apparent contradiction between the large number of possible conformations and the fast folding rates[10]. However, in the folding funnel picture it becomes evident that folding is not a random search as the transition from one conformation to the other is directed by the free energy gradient. According to this, there are multiple parallel pathways that are channeled towards the unique native structure[5]. This process can be visualized as a ball rolling down in a funnel. The funnel shaped energy landscape ensures that the native state is the global energy minimum and it is kinetically accessible.

1.2. Interactions of folded proteins

1.2.1. Phenomenological approach

Proteins are gregarious. Proteins seldom exert their function without interacting with one ore more protein partners and the social network of proteins is built from these specific interactions. These macromolecular interactions form large protein interaction networks in all living organisms[11]. Apart from individual protein structures, the structure of numerous complexes formed between globular, ordered proteins have been solved by X-ray crystallography and NMR and have been deposited into the PDB as well. From the known coordinates of the complex, fundamental properties of the protein-protein interfaces can be calculated. Several systematic studies have analyzed the complexes in terms of their hydrophobicity, accessible surface area, shape complementarity and residue preferences[12; 13]. The comparison of these properties between interior, surface and interface components in oligomeric proteins can reveal some of the basic molecular principles of these interactions[12; 14].

The interface is usually defined as a set of accessible surface residues that become buried during the complex formation. Similar to the core of globular proteins, multiple van der Waals interactions, salt bridges, H-bonds and hydrophobic interactions can be formed across the interface. For strong interactions, a large interface is usually needed. Although the interface properties of these complexes may vary depending on the size and geometry of the partners, the distribution of the interface sizes show a well pronounced maximum at around 1000 $Å^2$ with more than 75% of the interface sizes being in the 500-1500 $Å^2$ range[15; 16]. Larger interfaces are usually formed between permanent complexes. In contrast to the relatively narrow range of the sizes of interfaces, the length distribution of these proteins showed a much larger variance, falling into the range of 50-500 residues and there is no trivial linear dependence between the size of the interface and the protein length. Another consequence of the well defined structure is the segmented nature of the interfaces. During folding, residues that are distant in the amino acid sequence are

brought together in the three-dimensional structure. As a result, residues forming the interface usually belong to several non-contiguous segments and cannot be mapped to a single run of residues in the sequence. Typically the globular interfaces are made up of 2 to 7 segments with some complexes having even higher segmentation values over 10[12; 16].

The proper complementarity of the interfaces is an important signature of specific protein-protein interactions. As a result of multiple atomic level interactions, the interface can be as highly packed as the protein core. On average the residues from the two partners interact via 4-8 atomic contacts[16]. Generally, the interface has an intermediate hydrophobicity between those of the hydrophobic interior and the mostly polar exterior[12; 13]. However, the association of globular proteins is driven not only through hydrophobic patches on the surface, but polar interactions between subunits can also make significant contributions. At closer inspection, various types of complexes showed significant variations in the relative contributions of the different interactions. It was shown that homodimer interfaces were more hydrophobic compared to heterodimers, and that the interfaces of homodimers, permanent hetero-complexes and enzyme-inhibitor complexes were more complementary than antigen-antibody complexes[12].

The functional importance of interface residues is reflected in their evolutionary conservation. Conservational analyses showed that the interface residues are significantly more conserved among homologous sequences that the rest of the surface residues[17]. As promising as this feature may be from the perspective of protein binding site predictions, conservation alone is not sufficient to recognize protein binding sites[13]. Therefore, the conservation values deduced from multiple sequence alignments are usually combined with several physico-chemical parameters in order to highlight residues involved in protein-protein interactions[18]. In addition to these criteria, the most prominent feature of residues in protein binding regions is that they have to be accessible by the interacting partner. However, accessibility can only be predicted from sequence alone with a modest success rate[19]. Furthermore, the segmented nature of globular binding regions means that many, sequentially distant regions should be recognized at the same time. As a

consequence, successful binding site prediction algorithms generally require the solved structure of the interacting proteins. Methods that predict binding sites directly from protein sequence alone are much less common[20] and usually perform with lower accuracy[13].

1.2.2. Thermodynamics approach

The association of protein complexes can be approached based on basic physical principles similarly to protein folding. From a thermodynamics point of view, the interactions between proteins are governed by changes in the Gibbs free energy. The total change in the Gibbs free energy between the initial and final states, $\Delta_r G°$, determines the equilibrium constant K and thus the balance of proteins in the free and the bound form:

$$\Delta_r G° = -RT \ln K \quad (3)$$

Similarly to the description of folding, ΔG° can be divided into enthalpic and entropic terms to give more insight into the binding process:

$$\Delta_r G° = \Delta_r H° - T\Delta_r S° \quad (4)$$

The balance of the enthalpic (H) and entropic (S) terms determines the nature of the interaction, governing the affinity of the binding. The enthalpic term is dependent on the type and the complementarity of the interacting residues. The entropic terms are intimately linked with the flexibility of the partners. The diversity of protein-protein interactions can be traced back to the many different ways these factors can be combined in order to form highly specific functional protein complexes.

1.2.3. Models of molecular recognition

The details of the molecular interactions are determined by the properties of the interacting partners in the initial free state and the final bound state. Accordingly, for the description of protein complexes, the basic properties of the interacting molecules in their

free state also have to be taken into account. Depending on the nature of the unbound state relative to the bound state, various models have been proposed. Although some of these models were developed originally to describe enzyme-substrate interactions, they can be readily applied in a more general way to model generic protein-protein interactions as well. The classical view of molecular recognition was based on the *lock-and-key* model[21], that emphasized the chemical and geometrical complementarity of the interfaces without invoking any changes in the free and the bound protein conformations. This scenario applies, for example, to the complex formed between trypsin and basic pancreatic trypsin inhibitor (BPTI). The experimental data show that both partners have a well-defined structure in the unbound form that is nearly identical to the conformation adopted in the bound form. However, many proteins exhibited a slightly different preferred conformation in the two states. To account for this, the *induced fit* mechanism was suggested[22]. Induced fit arises due to the imperfect complementarity of the interface of the partners. Upon binding, the structure of one or more of the partners is changed by the interaction, and the conformation in the bound form differs from that of the free form. Despite these differences, both the lock-and-key and induced fit models basically assume a single stable conformation under given experimental conditions[23; 24].

An alternative explanation for the conformational differences in the free and bound form is offered by the concept of *conformational selection*[9; 25]. According to this model, one or both of the partners have multiple low energy conformations in the unbound state. The Gibbs free energy differences between these states determine the balance of the population in these conformations. However, during binding the interaction with the partner shifts this equilibrium. The lowest energy conformation adopted in the bound complex is different from the dominant structure of the free state, and it corresponds to one of the higher energy alternative conformations. The importance of the conformational selection model lies in that it can take into account the conformational heterogeneity of proteins. As the resolution of experimental techniques improves, more and more weakly populated, higher energy conformations can be detected, and the importance of these

conformations during the binding process is becoming more apparent[24]. The new examples provide further support to the concept of conformational selection.

1.2.4. The energy landscape view of protein-protein interactions

As the thermodynamic principles of protein folding and binding show a strong similarity, the concept of energy landscapes, introduced in section 1.1.3 can be readily applied to protein-protein interactions as well[24; 26; 27]. A major advantage of the energy landscape view is that conformational heterogeneity naturally follows from it. The energy landscape of the complex is created from the combination of the conformational space of the interacting molecules. However, the interaction with the partner molecule can induce drastic changes in the shape of the energy landscape corresponding to the individual protein. Comparing the shapes of the energy landscape of the free and bound state, the conceptual differences of the various binding mechanisms can be illustrated[27] (see Figure 2). In the *lock and key* model (Figure 2A), both partners have one, well defined minimum in their respective energy functions that corresponds to the native conformation. The combined energy function of the complex also has only one minimum that defines the same conformation for both partners as in their respective unbound native conformations. The *induced fit* model (Figure 2B) starts from the same assumptions for the unbound states of the partners. However, the complex formed by the partners in their respective native states does not correspond to an energy minimum. From this state, the complex reaches the energy minimum by slight alterations in the conformation of one or both partners. The basis of the *conformational selection* model (Figure 2C) is that at least one of the partners has two or more well-pronounced minima that are separated from each other by an energy gap. Upon interaction, the conformation corresponding to one of these minima is selected by the partner.

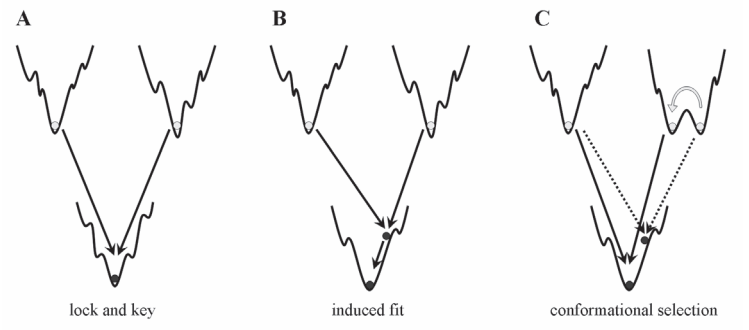

Figure 2: modes of molecular recognition in folded protein complexes

The three classical models for interactions between globular proteins: A) lock and key model, B) induced fit and C) conformational selection. The energy of the system is sketched against a single coordinate of the conformational space. The initial and final states of proteins are represented by light and dark dots, respectively. Arrows mark the pathways of binding and dotted arrows show binding pathways with unfavorable energies.

1.3. Intrinsically disordered proteins

1.3.1. Re-assessing the structure-function paradigm

In the first approximately 40 years of structural biology, the central model underlying all biochemical studies was that a well-formed structure is a prerequisite for a protein to carry out its function. Following the advice of Crick: 'If you want to understand the function, study the structure', this notion motivated a large number of structure-function studies and lead to the structure determination of more than 50 000 proteins. Although some proteins and protein segments were known that either did not lend themselves to structure determination or had sequence features that were seemingly incompatible with a folded structure (eg. highly charged, repetitive sequence regions), these were considered as hallmarks of imperfect experimental conditions or some exotic rarities of nature.

With the explosion of available genome sequences, during the 1990's the known number of these 'rarities' and 'experimental errors' grew steadily up to the point where they could no longer be put down on a side note. This forced molecular biologists to reassess the structure-function paradigm[28]. The world of proteins were extended to include proteins that do not require a stable, three dimensional structure even under physiological conditions in order to fulfill their biological role[29-31]. These intrinsically unstructured/disordered proteins (IUPs/IDPs) lack a well defined tertiary structure and exhibit a multitude of conformations that dynamically change over time and population. The importance of protein disorder is underlined by the abundance of partially or fully disordered proteins encoded in higher eukaryotic genomes[32]. Using bioinformatics predictors it was estimated that 30–50% of eukaryotic proteins contain at least one long disordered segment. The fact that protein disorder is not a tolerated necessity but provides an evolutionary advantage is reflected by studies showing the steady increase of the fraction of disordered proteins in proteomes as organism complexity increases[33; 34]. Furthermore, disordered proteins are involved in many important regulatory biological functions[30] including transcription, translation and cell signaling, complementing the functional repertoire of globular proteins[35].

Recent characterization of IDPs based on their functions shows that disorder can help these proteins to fulfill their functions in various ways[36; 37]. In accord with the wide variety of functions associated with it, protein disorder too comes in a large number of varieties. In some cases disordered regions are short and can be found at the terminal regions of globular domains, such as the disordered N-terminal region of eIF4E. Similarly, globular domains can also harbor flexible loops that appear as missing regions in solved structures. Flexible linkers that connect globular domains, such as Zinc fingers represent another type of localized disorder. In another scenario, especially in complex organisms, protein disorder often encompasses larger, domain sized regions. These regions can exhibit different degrees of flexibility ranging from the near-random conformation of the ACTR domain of the p160 protein through the presence of local transient secondary structural elements – such as in the N-terminal region of p27 – to

compact molten globule regions with considerable amount of secondary structure but without stable tertiary structure, such as the nuclear coactivator binding domain of CBP.

1.3.2. Coupled folding and binding of IDPs

The intrinsic flexibility of disordered proteins is intimately linked to their functions. In the case of entropic chains, the biological function is directly mediated by disorder (e.g. MAP2 projection domain, titin's PEVK domain, NF-M and NF-H between neurofilaments, nucleoporin complex). However, most disordered proteins function by binding specifically to other proteins, DNA or RNA. The lack of structure in the unbound form has profound effect of both the binding process and the resulting complex. In all cases the flexibility of the disordered partner decreases due to the binding, most cases to a level where the resulting complex lends itself to traditional structure determination. In these cases folding is said to be coupled to binding and this coupling modulates the energetic process of binding compared to globular proteins[38; 39]. As discussed earlier in section 1.2.2, the interaction of proteins can be described thermodynamically with the use of the change in the Gibbs free energy (see section 1.2.2, formula (4)). The resulting protein complex corresponds to the state with the minimal Gibbs free energy. However – as opposed to the interaction of globular proteins – in complexes involving IDPs, the loss of entropy during the folding of the disordered partner has to be taken into account and the entropic term (S) can play a much larger role. The loss of degrees of freedom during the coupled folding and binding in the disordered partner is dependent on the flexibility of the IDP in the bound and unbound form. As discussed in the previous section, the starting flexibility can vary in a wide range from near-random proteins to molten globules. On the other hand, it has been shown that IDPs can retain a varying degree of this flexibility in their bound form as well[40]. As a result, ΔS can also cover a wide range and can effectively tune the binding strength over a wide range. This results in a weaker binding compared to that of globular proteins. By uncoupling specificity from binding strength, IDPs are more prone to form specific, yet transient interactions[30; 36], which are

indispensable to regulatory and signaling processes[37]. The increased rate of association and dissociation of disordered proteins increase their temporal binding capacity. Furthermore, as discussed in section 1.5, disordered proteins are able to incorporate a higher fraction of their surface in the binding interface, which increases their interaction capacity in a spatial sense as well[41]. Consequently, disordered proteins in general can mediate a large number of interactions thus serving as hubs of protein-protein interaction networks[42].

1.3.3. Involvement in diseases

Given the functional importance of disordered protein regions, their malfunction is expected to have serious biological consequences. IDPs indeed are often associated with various diseases, including neurodegenerative diseases, amyloidosis, diabetes, cardiovascular diseases and cancer[43-46]. Despite the fact that proteins involved in these diseases are shown to have a higher disorder content, the exact role of protein disorder in these cases are not fully understood. As a consequence, disordered proteins involved in diseases is an intensely studied research area.

Probably the most results published to date concern the involvement of IDPs in cancer[47]. Many notable proteins were studied individually, exemplified by BRCA1, p27, p21 and CBP, that are involved in various forms of cancer. One of best characterized disordered proteins, p53, is known to be directly inactivated in more than 50% of cancers. At a more general level, the higher proportion of disordered proteins among cancer associated proteins was also observed[47]. According to the analysis of the SwissProt database, 79% of human cancer associated proteins have been classified as IDPs, compared to 47% of all eukaryotic proteins. The correlation between protein disorder and cancer was further underscored in the case of two common forms of generic alterations, chromosomal rearrangements and copy number variations[48; 49]. In addition to cancer, disordered proteins were also suggested to be common in diabetes and cardiovascular

diseases. Several disordered proteins – such as Aβ, τ, α synuclein, and prion proteins – are involved in neurodegenerative diseases and are also prone to amyloid formation. Altogether, these results lead to the conclusion that protein disorder comes with a 'biological cost' that is reflected in an increased risk of cancer and other diseases[30; 43]. This calls for the understanding of the role of protein disorder in various diseases.

Apart from basic research interests, the connection between protein disorder and involvement in diseases has implications in therapeutics as well. The pharmaceutical industry is currently struggling to find promising new drug targets, despite substantial increases in research funding. Drug discovery rates seem to have reached a plateau or perhaps are even declining, suggesting the need for new strategies. Until recently, the feasibility of targeting proteins without a well-defined structure was unclear for the purpose of drug development[50]. There is now, however, a newly sparked interest in IDPs as potential drug targets[51]. This is supported by finding specific inhibitors to block the interaction between p53 and MDM2, or between c-Myc and Max. Recognizing the relevance of these proteins stimulated more systematic efforts aiming at their structural characterization and determination of their mechanisms of action.

1.4. Predicting protein disorder

The detailed structural and functional characterization of disordered proteins is quite a challenging task. On one hand, as disordered proteins are generally involved in regulatory functions, their expression levels are lower on average, making them more difficult to isolate. On the other hand, disordered regions are more prone to degradation by proteolytic enzymes than well folded proteins. Furthermore, the existing experimental procedures are highly biased for ordered proteins, and most techniques provide only indirect information about disorder. Consequently, the current list of experimentally verified disordered proteins is rather limited. Currently the largest organized catalogue of experimentally verified disordered proteins and protein segments is the DisProt[52]

database in which over 1,400 disordered regions inside over 650 proteins are collected. In the light of the fact that about half of the human proteins are thought to contain at least one longer disordered segment, the amount of data in the DisProt is scarce at best. This discrepancy faithfully reflects the difficulties of the experimental identification of disordered proteins. Because of these difficulties, bioinformatics tools that target the prediction of protein disorder from the sequence play a very important role in the identification and characterization of IDPs.

1.4.1. Basic sequence properties of IDPs

The first analyses of sequences of disordered proteins revealed significant differences in the amino acid composition of ordered and disordered proteins. Basically, globular proteins have a relatively balanced amino acid composition in terms of hydrophobic and hydrophylic amino acids. Compared to this, the composition of disordered proteins is biased. These proteins are generally depleted in bulky hydrophobic and aromatic amino acids, which would normally form the hydrophobic core of folded globular proteins. On the other hand, they are enriched in polar and charged amino acids. At closer inspection, however, various datasets of disordered protein sequences exhibited further variations in their sequential bias. Differences could be observed depending on the experimental method used to identify disordered regions[53] (e.g. CD, NMR, or X-ray crystallography), depending on the length of disordered regions[54], and the location in the sequence (N and C-terminal, middle regions)[55], although these differences are smaller compared to the differences observed between ordered and disordered proteins.

The amino acid compositional bias of disordered proteins suggests the relevance of hydrophobicity scales for the discrimination of ordered and disordered segments. Among various amino acid scales, properties related to flexibility and coordination number had the highest discriminatory power[56]. Several disordered prediction methods are based on a simple amino acid propensity scale[57], such as the mean packing density of residues

calculated from atomic contacts or the difference of the amino acid propensities to be in coil and regular secondary structure elements. It was also suggested that the combination of low average hydrophobicity and net charge can identify disordered proteins. A specific amino acid scale optimized to discriminate ordered and disordered regions was also constructed[58].

The appeal of single amino acid propensities is that they are easy to calculate and to interpret, however, they are limited to a single effect. This can be insufficient to account for the complex phenomenon of protein disorder. Such properties, however, are also useful to reduce the dimensionality of the input data. By focusing on the relevant properties, an increased performance can be achieved during prediction. Several methods exploited amino acid scales in their predictions, including PONDR VLXT and VSL2 or DisPSSMP.

1.4.2. Machine learning approaches

The prediction of protein disorder can be viewed as a classic binary classification problem (ie. the whole protein or each residue of a protein chain has to be categorized either as ordered or as disordered) and can be addressed by standard machine learning techniques. The underlying assumption is that sequence features calculated from a local sequence window can be directly mapped into the property of order or disorder. Most methods assign disordered and ordered status at the amino acid residue level. Several disorder predictions are based on already existing methods developed for other areas of protein structure prediction, implemented using the specific datasets of disordered proteins. The novelty of many disordered prediction methods based on machine learning approaches lies in the representation of input information, rather than in the algorithms themselves. A comprehensive review of published methods appeared in the literature recently[59].

The two main applied approaches to machine learning methods is neural networks and support vector machines (SVMs). The first method developed for the prediction of disordered proteins is PONDR VL-XT which is based on feed-forward neural networks, one the most common methods in the field of bioinformatics. Elaboration of the used algorithm and the inclusion of position specific scoring matrices, secondary structure predictions and other information gave rise to a newer generation of methods, including DisPSSM and RONN. The first disorder prediction algorithm using SVMs was DISOPRED2. This method was followed by others such as the POODLE family (including the POODLE-I method) and PONDR VSL2 and also algorithms employing a recursive architecture such as DISpro and OnD-CRF. With the increase of available prediction methods, the meta approach is becoming more common. Servers, like MD or metaPrDOS, work by integrating the results of several disorder prediction methods. Although these developments can lead to improved prediction accuracies, there exist other viable alternative approaches.

1.4.3. Physical modeling – the IUPred algorithm

As opposed to the application of various 'black box-like' machine learning algorithms, the prediction of protein disorder can be approached with the direct implementation of physical principles governing the process of protein folding. A prime example of such approaches is the IUPred algorithm[60]. This method captures the essential cause of protein non-folding: if a residue in a protein is not able to form enough favorable intrachain contacts, it will not adopt a stable position in the 3D structure of the chain. If such residues are clustered along a segment of a protein or the whole protein, then this segment or the entire protein will be disordered.

The implementation of the above principle in IUPred is done taking an energetics point of view. For globular proteins, the contribution of interresidue interactions to total energy is often approximated by low-resolution force fields, or statistical potentials, energy-like quantities derived from globular proteins based on the observed amino acid

pairing frequencies[61]. In deriving the actual potentials, different principles have been applied. The resulting empirical energy functions are well suited to assess the quality of structural models and have been used for fold recognition or threading but also in docking, ab initio folding, or predicting protein stability. Their success in a wide range of applications suggests the existence of a common set of interactions, simultaneously favored in all native – as opposed to alternate – structures.

The pairwise energy E of a protein in its native state is the sum of the energies of all its pairwise residue-residue interactions. E is the function of its conformation as well as its amino acid sequence, as these define the list of residue-residue interactions that have a contribution to the total energy. This total energy can be calculated by taking all contacts in the protein, and weighting them by the corresponding interaction energies. The interaction energy between any two types or amino acids can be inferred by calculating the frequency of interactions between these two types in a dataset of known protein structures. These frequencies are transformed into interaction energies using the Boltzmann hypothesis[62] and are described by the 20 by 20 interaction energy matrix of amino acid pairs, **M**. Hence, the pairwise energy content calculated based on the structure can be written as:

$$E_{calculated} = \sum_{i,j} M_{ij} C_{ij} \quad (5)$$

where M_{ij} is the interaction energy between amino acid types i and j, and C_{ij} is the number of interactions between residues of types i and j in the given conformation.

This energy calculation, however, assumes the knowledge of the 3D structure of the protein and as such, is not directly applicable to proteins whose structure can not be determined. To come around this problem, a novel estimation scheme was established and implemented in IUPred to enable the estimation of the E interaction energy without the structure, using the protein sequence alone. The rationale behind this approach is that the energy contribution of a residue depends not only on its amino acid type, but also on its potential partners in the sequence. It is assumed that if the sequence contains more

amino acid residues that can form favorable contacts with the given residue, its expected energy contribution is more favorable. The simplest approximating formula for the specific estimated pairwise energy can be expressed with a quadratic formula as:

$$E_{estimated} = L\sum_{i,j} P_{ij} f_i f_j \quad (6)$$

where L is the length of the protein, f_i is the normalized frequency of residues of type i and **P** is the energy estimator matrix. The elements of **P** are optimized on a set of globular proteins using the least squares method in order to minimize the difference between $E_{calculated}$ and $E_{estimated}$. Equation (6) gives an estimate for the energy of the whole protein, however can be naturally modified to calculate the pairwise energy of single residues as well. For this, it has to be considered that in multi-domain proteins the residues belonging to different domains do not interact. For this end, the amino acid frequencies are only calculated in the sequential neighborhood of the residue in question. The width of this sequence window is marked by w_0 and is set to 100 residues to each side, therefore limiting the amino acid composition calculations to 200 residues, that roughly corresponds to the average domain size. To estimate the interaction energy of residue k (of type j), equation (6)) can be modified:

$$E_j^k = \sum_{i=1}^{20} P_{ij} f_i^k(w_0) \quad (7)$$

where $f_i^k(w_0)$ is the fraction of residues of type i in the w_0 neighborhood of residue k. (Note that lower indices stand for amino acid type, while upper indices stand for position in the chain.) Formula (7) enables the estimation of the intrachain interaction energies of each residue directly from the amino acid sequence. Generally, residues with less favorable predicted energies are more likely to be disordered. Testing on 559 globular and 129 disordered proteins[60] showed that this energy estimation scheme is accurate enough to achieve a high true positive rate (fraction of disordered residues correctly predicted) of 76% while maintaining a sufficiently low false positive rate (fraction of ordered residues incorrectly predicted) of 5%, a standard choice of type II error in prediction methods. The strength of the construction of the method is that its parameters

are derived from a globular protein dataset without the use of specific datasets of disordered proteins. As globular protein datasets are considerably larger than that of disordered proteins, this grants the method substantial stability compared to methods where a large number of parameters are trained on a limited and sometimes ambiguous disordered protein dataset.

The above energy estimation method is implemented in IUPred. The method is accessible via a web server[63] hosted at the Institute of Enzymology (http://iupred.enzim.hu). For the ease of interpretation, the calculated energies are converted into probability values, indicating the probability of each residue being disordered. Figure 3 shows an example output of the IUPred server for the human Wiskott-Aldrich protein (WASp). WASp is a 502 residue long protein that is entirely disordered with the exception of the ordered WH1 domain spanning the 39-148 region. The assigned probabilities are in accordance with the known structural information as the calculated probabilities on the ordered domain lie below 0.5 marking order (low probability of being disordered) and above 0.5 for the rest of the protein (high probability of being disordered).

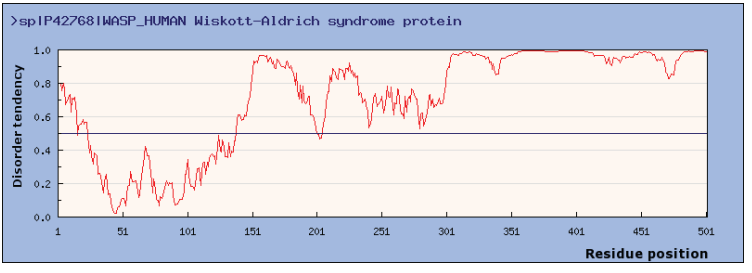

Figure 3: The IUPred server

Screenshot of the IUPred server output for the human Wiskott-Aldrich protein. The horizontal axis represents the protein chain and the vertical axis represents the probability of each residue to be disordered. Residues with values above 0.5 are predicted to be disordered and values below 0.5 indicate an ordered structure.

1.5. Molecular principles of the interaction of disordered proteins

As discussed in section 1.3.2, there is increasing evidence that disordered proteins participate in many vital biological processes and their function often involves protein-protein interactions. While these segments are disordered in isolation, many of them become ordered during binding to their specific partner. As a result of the coupled folding and binding process, the structure of these complexes can be studied with traditional structure determination methods. Although the PDB database contains significantly fewer such cases, even these examples demonstrate the definitive differences of the complexes involving disordered proteins compared to the complexes of ordered globular proteins. Although the structure of the complexes of disordered proteins also shows a rigid conformation, many of their distinct properties give away their inherent flexibility[16].

In most cases, disordered segments adopt a largely extended and open conformation in the complex. The absolute interface size of disordered segments is in the same range observed in the case of globular proteins, with only a few exceptions presenting very large interfaces (over 3000 $Å^2$). However, the length of the regions undergoing disorder-to-order transition is generally significantly shorter. These regions are usually below 100 residues; in many cases the disordered binding regions are less than 30 residues long. Therefore, the relative contribution of the residues to the interface is much higher. Furthermore, the interface area relative to the surface area of bound IDPs is much higher than in the case of globular proteins, meaning that these proteins utilize a much larger fraction of their accessible surfaces compared to globular proteins. An important property of disordered binding regions is that they are usually well localized in the sequence - in about 70% of the cases the interacting residues can be mapped to a single continuous region of residues. These localized interacting regions allow IDPs to have an increased modularity as different binding regions can be incorporated into the same protein without excessively increasing protein length. These binding regions can be close to each other or

can form mutually exclusive overlapping sites. The compact arrangement of multiple binding regions is possibly one of the reasons for the abundance of IDPs among protein-protein interaction network hubs.

The distinct binding mode of IDPs is also reflected in the physico-chemical nature of their interfaces. The interface of disordered proteins is more hydrophobic, and the preferred interaction contacts are also significantly different compared to the more familiar globular proteins. As opposed to the large number of polar-polar interactions at globular interfaces, IDPs tend to favor hydrophobic-hydrophobic contacts with the partner protein. The increased importance of hydrophobic interactions during binding is a hallmark of the complexes involving IDPs. As a result of the binding, the short disordered segments can also adopt both regular (e.g. α-helix) and irregular local conformations. Similar to globular proteins, the interface properties are relatively well conserved. Although disordered regions tend to have lower conservation scores, the scores calculated for the regions becoming ordered during binding and especially for the interacting residues are significantly higher than for the rest of the sequence.

1.6. Linear motifs

The study of protein-protein interactions formed by disordered proteins is based on structural considerations as shown in sections 1.3.2 and 1.5. However, the study of interactions between protein domains and short, linear protein regions – a description which fits most interactions between folded and disordered proteins – has a distinctively separate approach as well. In this case, the interaction is not described focusing on the short partner, but the large one, which is usually a protein domain. It was found for many domains such as SH2/SH3, 14-3-3, WW and MAPK that their interacting partners – albeit in many cases not being homologues – share a limited number (typically between 2-10) of common residues in the short interaction region[64; 65]. These amino acids are interspersed with flexible positions that can accommodate a variety of amino acids

without disrupting the binding[66]. Figure 4 shows the example of nuclear receptors that are able to bind a large variety of protein partners. Although most partner proteins are not homologues, they all share three key leucine residues at their interacting sites. During the interaction, the region that binds to the receptor forms an α-helix and the three leucines form a hydrophobic patch on the surface of the helix. This patch in turn recognizes the appropriate complementary hydrophobic region of the interface of the receptor and anchors the helix to the binding groove. The consensus sequence of the binding region is xLxxLLx, where x can stand for any amino acid, except for proline, as it would disrupt the helix formation. This motif is called LIG_NRBOX and ligands of many nuclear receptors are able to recognize their receptor partners via these sequence patterns. The theory of linear motifs, used to describe such interactions, is based on the assumption that these common residues (constituting the motif) mediate the binding largely independent of the other regions of the protein they are embedded in, functioning autonomously. However, in many cases the role of the context was shown to be larger than originally expected[67].

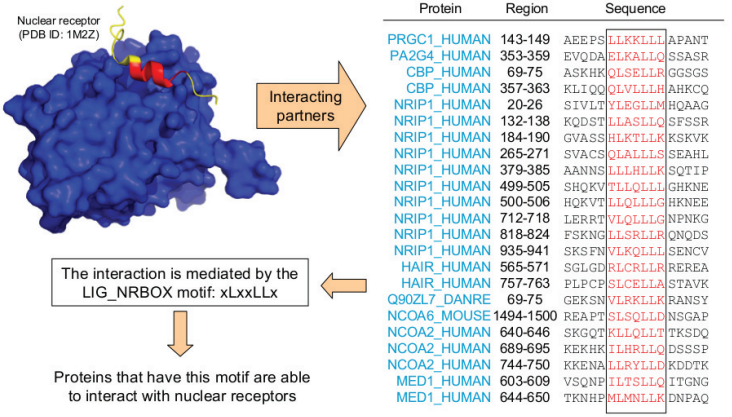

Figure 4: Example of the distillation of a linear motif

The figure shows the known interaction partners of nuclear receptors that all bind using the same binding mode. The upper left structure shows a solved complex structure between a small region of the human NCOA2 nuclear receptor coactivator (shown in red and yellow) and a glucocorticoid receptor (shown in blue). Although the actual sequences around the binding region do not share a high level of similarity, all contain three key leucine residues. These three amino acids interspersed and flanked by flexible positions constitute the consensus LIG_NRBOX motif (shown in red in the structure and the partner sequences).

The majority of protein-protein interaction mediating linear motifs were described in eukaryotes. Known motifs are usually represented by either a sequence logo or a regular expression and are collected in various databases[65; 68]. The most comprehensive and extensive available database of these motifs is the Eukaryotic Linear Motif (ELM) database[68]. Motifs are categorized into four groups: cleavage sites (CLV) mark the target regions of proteases; ligand binding sites (LIG) are generic protein-protein interaction sites that mediate the binding to a diverse set of domains, such as WW, 14-3-3 and SH2/SH3 domains; targeting signals (TRG) include known localization signals such as NLS and NES; modification sites (MOD) describe the regions of proteins undergoing various post-translational modification, such as phosphorylation, sumolation and

amidation. These motifs are known to be found in eukaryotic proteins, however some of these motifs can be expected to be present in other kingdoms of life as well. Furthermore, instances of the retinoblastoma protein-, the SH3- and the 14-3-3 interacting motifs, among others, were identified in various viruses as well.

Linear motifs can be readily used to search for binding partners of a given domain in unknown sequences through basic pattern matches. As an example, the first step in searching for nuclear receptor binding proteins would be to select proteins from a full proteome that harbor the above shown xLxxLLx motif. The strength of this method besides its simplicity is that it automatically gives information about the interacting partner: proteins matching the xLxxLLx motif are supposed to interact with nuclear receptors. This in turn can shed light on the localization and function of these proteins. However, these patterns usually consist of only a few fixed residues, and therefore most motifs are weakly defined, meaning that matches can arise purely by chance with a relatively high probability[69]. As a result, naïve motif searches are hindered by the massive amount of false positive hits: as leucines are hydrophobic, three leucines in close proximity can appear in sequence regions corresponding to the core region of globular proteins. This is partially the result of the incomplete description sequence patterns offer. Inside a living cell, the functionality of the motifs is modulated by structural, spatial and temporal control. Proteins harboring residues matching classical MAPK recognition motifs can be extracellular, hence never encountering MAPKs in reality. Furthermore, the proper structural context of a motif (such as being accessible, flexible and capable of forming the secondary structure necessary to fit into the binding cleft of the target domain) is crucial for its biological relevance and motif definitions do not include any such information.

Currently there are two major objectives around which studies involving linear motifs are centered. The first task is to distill new motifs[70-74]. This can be approached experimentally by identifying candidate interactions between proteins and then determining the residues from the short interacting partner that are essential to the

binding. After the determination of these essential interacting patterns from a sufficient number of partner proteins of a common domain, researchers aim to distill a consensus sequence that fits all the observed individual patterns. Virtually all known and accepted linear motifs have this kind of experimental background. However, there are many proposed bioinformatics methods that aim to reach the same goal based on protein sequences only. Generally the sequences of proteins interacting with a common partner are collected and various methods are used to identify significantly enriched short patterns in these sequences. Various discovery tactics are employed backed up by statistical models to give reliable results. Motifs distilled in such manner can be subjected to experimental validation and in some cases were shown experimentally to yield biologically meaningful motifs.

The second basic task in the field of motifs concerns the application of known motif patterns. Basic pattern matching approaches have a very small predictive power due to the massive amount of false positive hits. Therefore, additional information is introduced into the motif searches that aim to discriminate between true and false motif instances[75-79]. Such information can be based on annotations (eg. in searches using a motif that mediates interactions with nuclear proteins, extracellular proteins can be removed from the candidate list), but also can be based on predictions. Used predictions are usually aimed at the accessibility of the protein region containing the candidate motif hit. Commonly used predictions include domain and accessibility searches (motif hits found in the core regions of domains are not likely to be functional) and disordered predictions, as many motifs were shown to reside in disordered regions.

2. Scientific Aims

The main purpose of my Ph.D. work was to deepen the understanding of the molecular recognition processes of disordered proteins. I approached this aim by developing and using bioinformatics tools and protocols focused on the interactions of disordered proteins, and by applying them to gain biological insights.

Although many protein disorder prediction methods existed, as of 2007 – the start of my Ph.D. studies – there was no publicly available prediction method specifically aimed at identifying binding regions in disordered proteins. Accordingly, my first aim was to develop ANCHOR, a method capable of predicting disordered binding regions based solely on protein sequences, and making it available to the broad scientific community.

Following the completion of ANCHOR, I aimed at applying it in various bioinformatics studies that, on one hand, could serve with meaningful biological conclusions and on the other hand, had practical implications. I focused my studies on the following aspects of structural and systems biology:

- the appearance and presence of disordered binding regions throughout evolution
- the role of protein disorder and disordered binding regions in diseases caused by pathogens, using *Mycobacterium tuberculosis* as a model organism
- the association between protein disorder, interactions, function and involvement in cancer
- the connection of the theory of disordered binding regions with linear interaction motifs
- the possibility of using the theoretical description of protein disorder as a basis of modeling and predicting the various types of protein disorder on a common ground

3. Data and Methods

As usual in bioinformatics studies, my work relies heavily on computational methods and databases. Apart from using standard bioinformatics tools, such as BLAST, I focused on developing custom programs and protocols to specifically target the problem at hand. Furthermore, one of my main projects was the development of a novel prediction algorithm. These programs were written in either C (for computation intensive applications) and Perl (for the development of protocols, assembly of datasets and implementation of statistics methods). For each sub-project, I aimed to test and validate my results using positive and negative datasets. These datasets were assembled from various available data sources, including Pfam, PDB, UniProt, Disprot, the UCSC Genome Browser and the COSMIC database. During the evaluation of my methods and the validation of my results, I quantified their reliability using standard and customized statistical methods. As databases and methods were tailored for each sub-project, it would be difficult to discuss them out-of-context. Accordingly, the following chapter follows the structure of the results section (chapter 4) and each database and method is presented in the order they are referred to.

3.1. Databases

The data acquisition and the assembly of custom databases was done with Perl scripts. In each case I only quote the method used for the assembly of databases and omit the enumeration of their separate protein/domain/structure entries. However, adhering to the concept of reproducibility, the complete lists can be found in the supplementary materials of the referred papers.

Development of ANCHOR and the ANCHOR server (section 4.1)

Short disordered binding sites

Complexes from the PDB[1] (http://www.rcsb.org/) were collected by scanning the chains in the PDB entries against the Disprot database[52] (http://www.disprot.org/). A complex was accepted if it consisted of a chain with length between 10 and 30

residues that was found in the Disprot database as part of an annotated disordered segment and at least one interacting partner that was at least 40 residues long. Furthermore, complexes containing transmembrane proteins, RNA or DNA, chimeras, disulfide bonds between the disordered and ordered chains or a large number of unknown residues (marked with an X) were excluded. A few experimentally verified disordered complexes missing from Disprot were added to this set. A sequence similarity filter of 50% has also been applied to remove closely related proteins or protein segments. This procedure yielded a set of 46 complexes.

Long disordered binding regions

Complexes containing long disordered chains were collected in the same fashion as short ones but with different criteria for the length of the interacting partners. Here the length of the disordered chains was required to be at least 30 residues and they had to have an interacting partner of 70 residues or more. The resulting set contains 28 complexes.

Globular proteins

Globular proteins were collected from PDB entries that had only one chain of at least 30 residues. Also transmembrane proteins and complexes with RNA/DNA were filtered out. This dataset contains 553 proteins.

Disordered proteins

For the analysis of disordered proteins and protein segments the 3.7 version of Disprot database was used, considering only annotated disordered segments of 10 residues or longer.

Biological application of ANCHOR (section 4.2)

Complete proteome dataset

The dataset contains the protein sequences from 736 complete proteomes (53 archaea, 639 bacteria and 44 eukaryota) that were currently available from the SwissProt database (ftp://ftp.expasy.org/) marked as ‘complete proteomes’.

Studies concerning *Mycobacterium tuberculosis* (section 4.3)

Pfam domains

For the domain assignment the protein domains contained in the Pfam database[80] were used (http://pfam.sanger.ac.uk/). Both the manually curated Pfam-A and the automatically generated Pfam-B parts were used.

Sequence Dataset of Complete Proteomes (SDCP)

For the proteome-scale comparative studies, a dataset containing 1,904,578 protein sequences from 467 known complete proteomes was assembled (20 eukaryotic and 447 bacterial proteomes containing 392,401 and 1,512,177 proteins respectively). These proteomes were taken from the UniProt ftp server (ftp://ftp.uniprot.org/).

Large scale analysis of disorder, function and involvement in cancer (section 4.4)

COSMIC

Data were collected from the COSMIC database[81] (http://www.sanger.ac.uk/genetics/CGP/cosmic/). This is currently the most comprehensive catalogue of somatic mutations in cancer. Data are gathered from two sources, publications in the scientific literature, (v52 contains 11,437 curated articles) and the full output of the genome-wide screens from the Cancer Genome Project (CGP) at the Sanger Institute, UK. This dataset also incorporated the outcome of cancer genome projects. A small subset of the COSMIC database was also part of the cancer census dataset that were casually linked to oncogenesis. These genes constituted the COSMIC_census dataset.

Polymorphisms

Polymorphisms were collected using the UCSC Genome Browser[82] (http://genome.ucsc.edu/). Single genes were mapped to the genomic location corresponding to the UCSC Santa Cruz hg19/GRCh37 build. Those sequences, that could not be mapped, were changed or retracted, were discarded from further

analyses. The polymorphism data were obtained by mapping the SNPs of dbSNP (release 132) to the genomic coordinates. This release contained over 13 million SNPs. It also incorporated the results of the 1000 Genomes pilot projects that collected variations via whole genome shotgun sequencing from two families with high coverage and 179 individuals with low coverage. I used the Common SNPs corresponding to uniquely mapped variants that appear in at least 1% of the population. The commonness of these variations suggests that these are likely to be neutral polymorphisms with no clinical relevance. To ensure the quality of the polymorphisms data, I only used validated SNPs.

Human proteome

The proteins of the human proteome were downloaded from the "complete proteome" page of the UniProt database. Only reviewed entries were kept, resulting in a dataset of 20,232 proteins.

Functional annotations

Functional classifications were based on GeneOntology[83] (GO, http://geneontology.org/) terms assigned to human proteins in UniProt. I retrieved all GO terms for all proteins in the human proteome and mapped them to high level GO terms described in the Generic GOslim subset of GO. This subset contains 127 terms covering all three parts of GO annotations: biological processes (50 terms), cellular components (36 terms) and molecular functions (41 terms). All proteins from COSMIC, where possible, were mapped to UniProt sequences and were assigned the relevant GOslim terms.

Interactions

Protein-protein interactions were taken from the current release of the IntAct[84] database (www.ebi.ac.uk/intact/).

Linear motifs (section 4.5)

Linear motif patterns and instances

Linear motif patterns and true motif instances were downloaded from the Eukaryotic Linear Motif database[68] (http://elm.eu.org/). Only LIG motif instances were considered, and were filtered for 'true positive' logical annotation. Furthermore, instances were filtered for similarity using BLAST. Instance sequences producing significant similarity over regions including the same motif were clustered and only one representative from each group was kept.

Sequences from the three domains of life

Sequences from eukaryotes, bacteria and archaea were downloaded from the 'taxonomic divisions' section of the UniProt ftp server (ftp://ftp.uniprot.org/). The eukaryotic, bacterial and archaeal datasets included 171,208, 326,910 and 18,674 protein sequences, respectively.

Pfam domains

The number of PCNA, PDZ and cyclin domain occurrences in the sequences from the three domains of life were collected from the Pfam database (http://pfam.sanger.ac.uk/).

Human proteome

The proteins of the human proteome were downloaded from the "complete proteome" page of the UniProt database. Only reviewed entries were kept; the current release included 20,256 protein sequences.

3.2. Methods

In the following description of the used methods, I follow the order of the presentation of results in chapter 4. Unless otherwise noted, the used methods were implemented in the Perl programming language under Linux operating system. Only custom built methods and protocols are described in detail, standard statistical methods are only referred.

General methods:

Generation of figures:

For the representation of numerical data, the Gnuplot package was used. Figures showing protein structures were made using Pymol. Figures of the IUPred and ANCHOR web servers are screenshots of the actual server outputs. Complex figures were assembled in Powerpoint.

Development of ANCHOR and the ANCHOR server (section 4.1)

Parameter optimization:

The optimal parameters were determined by a three fold cross-validation, by dividing both the negative and positive datasets (Globular proteins and Short disordered binding sites, respectively) into three parts with approximately the same chain length and secondary structure distribution. Only the five parameters specific to ANCHOR, w_1, w_2, p_1, p_2 and p_3 were optimized by a grid search procedure. Specifically, w_1 was varied in the range of 20 to 100 in steps of 10 (giving 9 possible values), w_2 was varied in the range of 5 to 35 in steps of 2 (giving 16 possible values), and p_1, p_2 and p_3 was selected from 1000 sets of randomly generated values (p_1 and p_2 were randomly selected from the interval [-1;1] and p_3 was selected from the interval [0;1] in a way that the sum of their squares is always equal to 1). This yielded 1000 different (p_1, p_2, p_3) combinations. These, combined with all possible values of w_1 and w_2 gave 144,000 different parameter sets in total. These were considered in order to

select the optimal one, containing the five optimal parameters for each round of the cross-validation.

To quantify the performance of the predictor given a set of parameters I calculated the True Positive Rate (TPR) at False Positive Rates (FPR) fixed at 5% calculated on globular proteins as the negative set. Parallelly, the fraction of amino acids that are predicted as binding sites in general disordered regions of Disprot database that are correctly recognized as disordered by IUPred (*F* value) was also calculated (for a more detailed discussion of the F value, see section 4.1.2 of the results chapter).

The best parameter set was chosen manually, by reducing the parameter set in a step-wise manner based on the following steps:

1, Calculate TPR (at fixed FPR=5%) and F for each of the 144,000 candidate sets of parameters

2, Discard all for which F>50%

3, Discard all for which TPR<60%

4, From the remainder choose the 20 for which the difference between TPR and F is the largest

5, Choose the one for which TPR is maximal (the TPR-F difference among these 20 sets vary only within a range of less then 0.02 so that is not a good measure to choose the best one)

The negative and positive sets were divided into three parts, resulting in three different optimal parameter sets. The final predictor algorithm is constructed by averaging these three outputs. As the training sets only contained binding regions of at least 10 amino acids and I aim to identify at least 5 residues of each region, all predicted binding sites were removed that did not exceed 5 consecutive residues. A schematic figure of the training procedure is given in Figure 5 of section 4.1.2 of the results chapter.

Secondary structure evaluation:

Secondary structures were assigned with the DSSP[85] algorithm using the structures of the complexes downloaded from the PDB.

ANCHOR web server:

The core program of ANCHOR is written C, and both the online version and the downloadable version includes a Perl wrapper. This Perl program is called by the web server written in PHP. The graphical output is generated by the JpGraph software (http://www.aditus.nu/jpgraph/). The default option for graphical/text output is automatically determined by the browser type, but it can be changed by user. Additionally, list of sequences can also be submitted to generate simple text output on a larger scale.

Studies concerning *Mycobacterium tuberculosis* (section 4.3)

PSI-BLAST similarity searches and similarity profiles

For the similarity searches between MTB proteins and the proteins in the SDCP (see Data section), PSI-BLAST was used. First, a PSI-BLAST profile was calculated for each of the 3,948 proteins in the MTB proteome using the UniRef90 database, with three iterations. Next, these profiles were used to find hits from the proteins in SDCP. A hit was considered significant (the MTB and the other protein was considered locally similar) and was used further on, if the e-value was below 10^{-4}. Based on the alignments, all locally similar sequences from the SDCP were collected for each protein in the MTB proteome. Next, for each MTB protein a similarity profile was built that contains the number of similar sequences for each of the 467 organism in the SDCP.

Cluster analysis

The input for the clustering algorithm is based on the similarity profiles generated for each MTB sequence. In the cluster analysis Euclidean distance was used together with Ward's method. The result of clustering was largely insensitive to various parameters of the clustering, including the type of the clustering method, various

types of normalizations and parameters of PSI-BLAST. The clustering was implemented in the R program package.

Large scale analysis of disorder, function and involvement in cancer (section 4.4)

Comparison of the cancer databases and the human proteome

The average ratio of disordered residues, ratio of proteins containing >30 residue long disordered regions and average length were calculated in the COSMIC and COSMIC_census datasets. These averages were compared to the average values calculated in the human proteome. Standard errors of the mean were calculated by selecting 10,000 random samples from the human proteome of the same size as the respective dataset. In each of the 10,000 random selections, the means were calculated. From these means the standard error of the mean was established and used to test the difference between the random samples and the database average. For the assessment of significance the confidence intervals of α=0.01 (corresponding to 2.576 standard errors) were considered.

Over- and under-representation of polymorphisms and cancer-associated mutations

For each protein in the COSMIC and COSMIC_census datasets, the sequences were downloaded from the Uniprot database or the UCSC Genome Browser. Using the sequence, IUPred was used to assess which residues were part of disordered regions. These results were also calculated with two other disorder prediction methods, DISOPRED and VSL2. ANCHOR was used to predict regions involved in disordered binding regions. For each protein, the number of polymorphisms and cancer-associated mutations within these regions were calculated. These numbers were compared to the expected number of mutations based on the assumption, that mutations occur in a random way. This expected distribution was calculated in the following way: to calculate the expected number of mutations for ordered and disordered regions, the number of observed mutations was divided according to the ratio of ordered and disordered residues in the given sequence. This model takes into account that the number of mutations can change from one protein to another. The number of expected and observed mutations was summed up separately for ordered

and disordered segments. Using these numbers, the statistical significance of the differences in the two distributions was assessed by χ^2 test.

In the case of cancer-associated mutations, an additional model was used to calculate the expected number of mutations. This took into account the uneven distribution of polymorphisms between ordered and disordered regions. The model was based on a normalization factor calculated from the ratio of the observed number of SNPs relative to their expected number. The normalization factor was calculated for disordered and ordered residues, in each dataset. The expected number of mutations was recalculated by weighting them according to the normalization factor for disordered and ordered residues within each dataset. Using these references, the statistical significance could be calculated similarly to the previous case. Unfortunately, current data does not enable to calculate this factor for proteins individually. However, when datasets were divided into subgroups, for example based on the number of mutations, the results did not change.

Distributions of functional categories

The distribution of each GO term was analyzed using the COSMIC_census dataset. To determine significantly over- or under-represented terms, the distribution of these terms in the human proteome was used as a reference. A random subset was selected from the human proteome dataset and was parsed for occurrence numbers of each term. This was repeated 100 times and then the average occurrence of each term was calculated. These occurrence numbers were compared to the occurrence numbers in the COSMIC_census dataset using left and right sided Fisher tests to assign significance values to the under- and over-representation of terms.

Features

The calculated length, ratio of disordered residues and disordered binding residues, interaction numbers and the number of COSMIC_census mutations for COSMIC_census proteins and the randomly selected reference human proteins were categorized into 5 bins to provide a coarse-grained description. The sixth feature

describing the functional involvement of the proteins was represented by 'functional profiles'. These profiles were calculated based on the significantly over- and under-represented GO terms shown in Table 4 of section 4.4.4. For each protein, a 13 element binary vector was assigned that showed which of the 13 considered GO terms the protein was annotated with.

Mutual information and Jaccard distance

The association between different features calculated on proteins was measured by calculating the mutual information ($I(X,Y)$) between all X and Y pairs of features using the standard formula:

$$I(X,Y) = \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x,y)}{p'(x)p''(y)} \right)$$

where $p'(x)$ and $p''(y)$ are the probability distributions of the features X and Y respectively and $p(x,y)$ is their joint probability distribution. As the maximal information of different features can vary (and hence their maximal mutual information can also vary), to be able to compare the association of different parameter pairs directly, the mutual informations were scaled:

$$D(X,Y) = 1 - \frac{I(X,Y)}{H(X,y)}$$

where $H(X,Y)$ is the joint entropy of X and Y:

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 p(x,y)$$

The resulting $D(X,Y)$ Jaccard distance is a universal metric with $D(X,Y)=1$ if X and Y are completely independent and $D(X,Y)=0$ if X and Y are identical.

The multidimensional scaling of the obtained distances was calculated using the R package.

Linear motifs (section 4.5)

ANCHOR

I used the default version of ANCHOR (http://anchor.enzim.hu), but lowered the cutoff value to 0.4 for disordered binding regions. However, I kept both included filters, meaning that all predicted binding regions shorter than 6 residues and predicted binding regions with extremely low disorder scores were removed. I considered an ELM instance found if there was an overlap between the instance and a binding region predicted by ANCHOR.

Random overlap of ANCHOR regions

The expected overlap between ANCHOR regions and randomly selected protein segments was determined in a stepwise fashion. First, 10,000 regions of length l were selected randomly from the sequences of the UniRef50 non-redundant database. These sequences were input to ANCHOR and the fraction of randomly selected segments overlapping with ANCHOR predicted regions were calculated. This procedure was repeated 10 times and the average overlap % was calculated. This was done with varying the l length between 3 and 20. From this the probability p of a randomly selected segment of length l overlapping with ANCHOR regions was fitted: $p(l) = 0.10984 + l * 0.004494$. The significance of the overlap between real motif regions and ANCHOR was calculated using the binomial distribution using $p(l)$ as the background probability, substituting the average length of the known instances of each motif.

GeneOntology (GO) annotations

GO annotations of the inspected LIG_NRBOX motif was taken from the ELM website. These include annotations from all three main categories of GO (biological process, cellular component and molecular function). From the biological processes the "Regulation of transcription" (GO:0006355) was kept, as the other annotated term ("Positive regulation of transcription", GO:0045893) is a direct child term of GO:0006355. From the molecular function annotations the "Transcription Co-

activation" (GO:0003713) was also omitted due to being a child term of "Transcription Cofactor" (GO:0003712). The "Transcription Factor Binding" (GO:0008134) term was replaced with its ancestor term "Protein binding" (GO:0005515).

GO annotations of human proteins were taken from the Gene Ontology Annotation section of the EBI homepage (http://www.ebi.ac.uk/GOA/proteomes.html). These annotations were mapped to the higher level annotations given in the Generic GOslim subset of GO. However, to remove bias in the analysis, Generic GO terms were slightly modified. All root level terms were removed (biological_process, cellular_component and molecular_function) in order to remove the excessive but uninformative term hits. For similar reasons very broad cellular component terms ("cell", "intracellular" and "organelle") were also excluded. The biological process term "Regulation of biological process" (GO:0050789) was removed as it is not used in the EBI human proteome annotations. Instead, its child term "Regulation of transcription" was added. Furthermore, the molecular function term "Transcription Cofactor" was also added as none of its child or ancestor terms are included in the Generic GOslim.

4. Results and Discussion

4.1. Developing ANCHOR, a method for predicting disordered binding regions

In my work I set out to develop ANCHOR, an algorithm that uses only the protein sequence as an input and can recognize protein segments that are disordered in isolation but can undergo a disorder-to-order transition adopting a well-defined structure upon binding to a globular protein partner[33]. Due to the inherent flexibility, these regions are difficult to study experimentally, making specific prediction methods even more valuable[86]. While there are several methods available for prediction of disordered regions, recognizing disordered binding sites was regarded as a more challenging problem due to the limited number of well-characterized examples. Even today, the number of solved structures of complexes between two proteins that were shown experimentally to be ordered and disordered is in the tens, as opposed to the thousands of solved complexes between ordered proteins. Accordingly, only a handful of dedicated disordered binding site predictors have been developed and as of April, 2012 ANCHOR remains the only general, publicly accessible such method.

The essential feature of disordered binding regions is that they behave in a characteristically different manner in isolation than bound to their partner protein. In their free state, they behave as disordered proteins, existing as a highly flexible structural ensemble. In their bound state they usually adopt a rigid conformation, similar to regions within globular structures. This capability to behave in drastically different ways in different environments is targeted by my approach. Biophysical considerations (see section 1.5) suggest that in most cases there is a strong signal in the amino acid sequence highlighting regions involved in coupled folding and binding and these regions are linear in sequence[16]. As a result, a relatively short sequence segment containing residues with a pronounced tendency to make interactions, leads to a characteristic sequence signal which enables the prediction of these regions from the sequence alone.

4.1.1. The construction of the algorithm

The basis of ANCHOR is a simplified model of the binding of disordered binding regions that is based on biophysical considerations. The three main features of such regions are that they reside in a larger disordered region, they cannot form enough favorable intrachain interactions to fold on their own and they have the capability to energetically gain by interacting with a globular partner protein:

1. The first criterion ensures that a given residue belongs to a long disordered region, and filters out flexible loops of globular domains.
2. The second criterion corresponds to the isolated state and it ensures that a residue is not able to form enough favorable contacts with its own local sequential neighbors to fold, otherwise it would be prone to adopt a well defined structure on its own.
3. The third criterion tests the feasibility that a given residue can form enough favorable interactions with globular proteins upon binding. This basically ensures that there is an energy gain by interacting with globular regions.

In the development of ANCHOR I quantified these three properties using a coarse grained energy-estimation model. The three resulting measures were then combined into a single predictor via optimized weights.

In more detail, the prediction of these three properties relies on the energy estimation framework implemented in IUPred, a general disorder prediction method (see section 1.4.3 and reference[60] for details). The core element of IUPred is the energy predictor matrix **P**. This 20*20 matrix contains the estimated interaction energies between all possible amino acid pairs. **P** can be used to estimate the total interaction energy of a protein formed by the intrachain interactions of its residues without the knowledge of the structure of the protein. The elements of **P** (P_{ij}) were trained on globular proteins with known structures only, without relying on any kind of disordered dataset. These

parameters were determined to minimize the difference between the estimated energies and the energies calculated from the known structures on the dataset of globular proteins. Using the energy predictor matrix IUPred predicts the *E* interaction energy for each residue based on the following formula in default:

$$E_i^k = \sum_{j=1}^{20} P_{ij} f_j^k(w_0) \quad (8)$$

where *i* denotes the type of the *k*-th amino acid, P_{ij} is the element of the energy predictor matrix that estimates the pairwise energy of residue of type *i* in the presence of residue type *j*, $f_j^k(w_0)$ is the fraction of residue type *j* in the sequential environment within w_0 residues from residue *k*. The size of neighborhood considered (w_0) equals 100 residues in both directions. For the final prediction output, the energies calculated for individual residues are smoothed over a window size of 10 (also in both directions from the *k*-th residue so in fact 21 residues are considered in total) and the resulting smoothed energies are transformed into probability values, denoted as s_k.

The disordered binding site prediction is based on three different scores that are calculated with a slight modification of the original energy estimation scheme. The parameters of P_{ij} were taken directly from IUPred. The following three scores are assigned to each residue in a protein according to the above described criteria (1-3):

1. To measure the tendency of the neighborhood of an amino acid for being disordered I use the IUPred algorithm and assign an S_k score to the *k*-th residue of the chain by averaging the IUPred scores in the w_1 neighborhood of the residue in question:

$$S_k = \frac{1}{N} \sum_{k \neq j=b_{lower}}^{b_{upper}} s_j \quad (9)$$

where s_j is the IUPred score of the *j*-th residue of the chain, N is the number of amino acids in the averaging and b_{lower} and b_{upper} are the lower and upper boundaries of the neighborhood of the *k*-th residue, that is $b_{lower} = \max(k\text{-}w_1;1)$ and $b_{upper} = \min(k+w_1;l)$, where *l* is the chain length.

2. I estimate the pairwise interaction energy the given residue may gain by forming intrachain contacts. This is done the exact same way as in IUPred using formula (8), only here the size of the considered neighborhood (w_2) is left as a parameter and is set during the training of the predictor:

$$E_i^{\text{int},k} = \sum_{j=1}^{20} P_{ij} f_j^k(w_2) \quad (10)$$

As can be seen later from the results of the optimization, w_2 is smaller than w_0 used in IUPred. The smaller window size corresponds to more local behavior.

3. The pairwise energy that the residue may gain by interacting with a globular protein is approximated using the average amino acid composition of globular proteins:

$$E_i^{glob} = \sum_{j=1}^{20} P_{ij} \overline{f}_{glob,j} \quad (11)$$

where $\overline{f}_{glob,j}$ is the fraction of residue type j in the averaged reference amino acid composition of globular proteins. By subtracting this energy from $E_i^{\text{int},k}$ one can estimate the energy that the residue may gain by interacting with a hypothetical globular protein compared to forming intrachain contacts: $E_i^{gain,k} = E_i^{\text{int},k} - E_i^{glob}$.

The final prediction score of the residue is given by the linear combination of the above three terms:

$$I_k = p_1 S_k + p_2 E_i^{\text{int},k} + p_3 E_i^{gain,k} \quad (12)$$

where the p_1, p_2 and p_3 coefficients are determined during the training of the predictor together with the optimal values of w_1 and w_2 window sizes. I_k is then converted into a p value that expresses the probability of that residue being in a disordered binding site. For a binary classification residues with scores above 0.5 are predicted to be in a disordered binding site. Since the second and third terms of (12) may vary heavily between neighboring residues, the final score is smoothed in a window of 4 residues.

4.1.2. Parameter optimization

In order to determine the optimal values for the three weights (p_1, p_2 and p_3) and the two window sizes (w_1 and w_2) positive and negative datasets were used. The positive dataset is composed of experimentally verified cases of complexes of globular and disordered proteins. Complexes were collected from the literature[16; 52; 87-91]. Only such cases were accepted where the partners were proven experimentally to be ordered or disordered and the complex has a solved structure with a relatively good resolution (*Short disordered binding sites* dataset, see section of 3.1 Data and Methods). The performance of ANCHOR with a given parameter set on this dataset is quantified by calculating the fraction of positive examples predicted to be binding regions. This measure is termed 'true positive rate' (TPR).

Apart from the positive dataset, two distinct negative datasets are also needed. First, the algorithm should not predict disordered binding regions inside globular proteins. To measure this, I assembled a dataset of ordered monomeric proteins (*Globular proteins* dataset, see Data and Methods). The goodness of a parameter set is given by the fraction of residues inside these proteins that are predicted to be in binding regions. This measure is termed 'false positive rate' (FPR). Second, the algorithm should be able to discriminate between regions of disordered proteins that either bind to a globular protein or not. However, no reliable database can be assembled for this purpose, as there cannot be any conclusive evidence for a disordered protein region that it does not bind to *any* globular protein. In order to circumvent this problem, during the evaluation of different parameter sets, I calculated the fraction of experimentally verified disordered protein segments from the Disprot database[52] (*Disordered proteins* dataset, see Data and Methods) which ANCHOR predicts to be binding regions. This fraction is termed F. The role of this value is to discriminate between general disorder prediction and binding site prediction. It would be possible to achieve a high TPR and a low FPR by predicting every disordered residue as part of a binding region. However, this would yield an F value of 1. In order to train the algorithm to specifically recognize binding regions, the optimal parameter set is which maximizes TPR while minimizing FPR and F.

I carried out the parameter optimization on the above three datasets by three-fold cross validation (see Data and Methods and Figure 5 for a schematic representation and outline of this procedure). The small dataset of known disordered proteins bound to ordered proteins represent a serious bottleneck during optimization. Therefore, it is a clear advantage of my approach that it greatly reduces the dependence on the existing dataset of disordered complexes, and leaves us with only 5 parameters to be optimized on this small dataset.

The behavior of various optimized scores is shown for an example, the N terminal domain (residues 1-100) of human p53 tumor suppressor protein that plays an important regulatory role[92]. Its N terminal region is completely disordered and is known to be able to bind to (at least) three different globular proteins as shown in Figure 6. The segment between residues 17-27 binds to MDM2, the other two binding sites overlap with residues 33-56 binding to RPA 70N and residues 45-58 binding to the B subunit of RNA polymerase II. The three calculated quantities for this domain are also shown in Figure 6. It is worth noting that the MDM2 binding site in the N-terminal region of p53 appears to be on the border of being disordered. Although the disordered prediction is part of ANCHOR, the output of this prediction (E_{int}) is linearly combined with two other quantities meaning that predicted disorder is not strictly a prerequisite of a successful disordered binding site prediction.

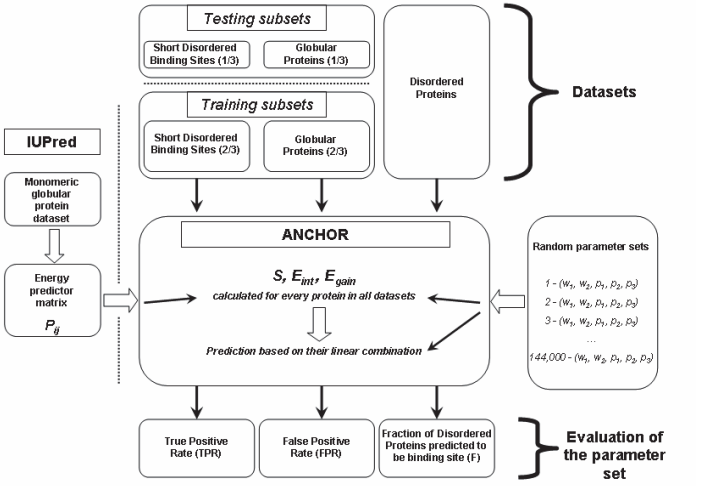

Figure 5: Outline of the training of ANCHOR

In the first step, our Short Disordered Binding Sites dataset and Globular Proteins dataset (positive and negative datasets) are split up and only 2/3 is used in the subsequential steps. Then a parameter set (w_1, w_2, p_1, p_2, p_3) is selected from the 144,000 random ones. This parameter set is used to calculate S, E_{int} and E_{gain} for every position in every sequence in the three input datasets using the fixed energy predictor matrix ***P***. Based on these calculations the evaluating measures are calculated: TPR is calculated on Short Disordered Binding Sites, FPR is calculated on Globular Proteins and F is calculated on Disordered Proteins. Based on these measures, the best parameter set out of 144,000 is chosen (see Data and Methods). This parameter set is then evaluated on the remaining one third of the datasets. These results are reported in Table 1.

This procedure is repeated for all three subsets of Short Disordered Binding Sites and Globular Proteins. The output of the three optimized predictors are combined into one final predictor by averaging their output.

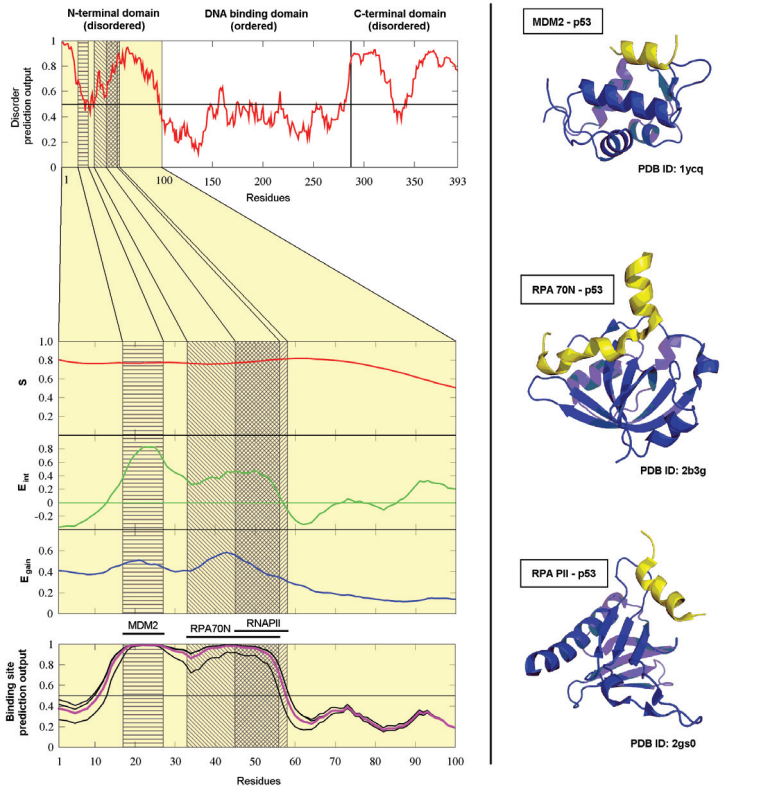

Figure 6: The construction of the ANCHOR prediction method demonstrated on the N-terminal domain of human p53

Left: IUPred prediction score for the full length human p53 (top) and *S*, E_{int} and E_{gain} calculated for the disordered N terminal domain of human p53 (middle). Grey boxes show the three binding sites with the overlap of the RPA70N and RNAPII binding sites shown in dark grey. The outputs of the three individually optimized predictors are shown in black and their average, the final prediction score is shown in purple (bottom).

Right: PDB structures of the binding sites in the N-terminal region of p53 (yellow) complexed with the respective partners (blue): MDM2 (top, PDB ID: 1ycq[93]), RPA 70N (middle, PDB ID: 2b3g[94]) and RNA PII (bottom, PDB ID: 2gs0[95]).

4.1.3. Testing of ANCHOR

I tested ANCHOR by dividing both the negative and positive datasets (*Globular proteins* and *Short disordered binding sites*) into three subsets, training the predictor on two of these and evaluating it on the remaining third one. This was done in all three possible combinations yielding three optimal parameter sets. The parameters calculated on the training sets are shown in Table 1 together with the respective TPR's, FPR's and F values. The optimal parameters were chosen to maximize the amount of correctly predicted disordered binding sites (TPR) while minimizing predicted binding sites in globular proteins (FPR) and also restricting predicted binding sites within disordered regions in general (F). I chose the widely used 5% as a maximal acceptable value for FPR. The fact that the three parameter sets do not differ significantly implies that the buildup of ANCHOR is robust.

Table 1: Parameter and prediction accuracy values obtained during the optimization of ANCHOR

	w_1	w_2	p_1	p_2	p_3	*F* (%)	*TPR* (%)	*FPR* (%)
Training set 1	25	60	0.4630	0.3847	0.7985	46.0	69.8	5.0
Training set 2	27	60	0.6075	0.4149	0.6773	47.4	67.7	5.0
Training set 3	29	90	0.6990	0.4585	0.5488	43.4	64.8	5.0

Optimal parameters of the predictor determined during training. w_1, w_2, p_1, p_2 and p_3 are the optimized parameters, *F* is the fraction of the residues in the disordered regions in the Disprot database that are predicted to be in binding sites, *TRP* and *FPR* are the True- and False Positive Rates, respectively.

The output of the predictor with all three parameter sets and the combined final predictor (the average of these three) are shown for the example of the N terminal region of p53 in Figure 6 of the previous section.

The results obtained on the three independent testing subsets as well as their average are given in Table 2. Since the cutoffs are given by the training process such that I

achieve exactly 5% False Positive Rate (FPR) on the respective training sets (ie. the part of the original Globular proteins dataset that was used in the training of the respective subpredictor), the FPR's are also quoted (they can differ slightly from 5%). Besides the overall TPR calculated on a residue basis (marked TPR_{AA}), I also calculated the percentage of binding sites identified, termed TPR_{SEG}. A binding site was considered to be found if at least five of its amino acids are correctly classified. The results show that ANCHOR performs at 62% TPR_{AA} with a slightly higher TPR_{SEG} of 68% on average, while maintaining a 5% FPR. ANCHOR is also specific to disordered binding sites as opposed to disorder to general. If all disordered proteins had approximately equal capability of binding then the fraction of correctly identified disordered binding sites (TPR) could not be significantly different from the fraction of disordered regions predicted to be binding sites (F value). As this is not the case (TPR=62% vs. F=42%) we can conclude that common features of known disordered binding sites that distinguish them from general disordered protein regions are successfully recognized.

Table 2: Prediction efficiency of ANCHOR evaluated on the testing datasets

	TPR_{AA} (%)	TPR_{SEG} (%)	*FPR* (%)
Testing set 1	61.1	62.5	5.7
Testing set 2	69.5	80.0	4.4
Testing set 3	54.7	62.5	5.1
Average	61.8	68.3	5.1

Results of the testing of ANCHOR on the three testing datasets. TPR_{AA} denotes the ratio of correctly identified amino acids belonging to binding sites. TPR_{SEG} denotes the ratio of binding sites found by the algorithm.

Another standard way of describing prediction algorithms is by Receiver Operating Characteristic (ROC) curves, that is the TPR versus the FPR of the algorithm. This relationship is mapped by scanning the interval between 0 and 1 with the score cutoff. The three ROC curves of the predictor with the three different parameter sets evaluated on the respective testing sets are shown in Figure 7. A single number measure to characterize the performance is the area under the curve (AUC) with random predictors

scoring AUC=0.5 and perfect predictors scoring AUC=1. The AUC values of the predictors trained and tested on the respective subsets are 0.8675, 0.8781 and 0.8993.

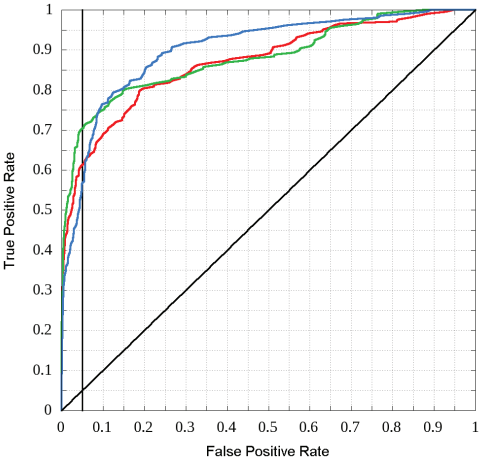

Figure 7: ROC curves obtained during the testing of ANCHOR

ROC curves of the predictor with parameter sets optimized on each of the three training subsets and evaluated on the respective testing subsets are shown with red, green and blue lines. The line with unity slope corresponding to random prediction is also shown. The vertical line corresponds to FPR=0.05, where the final predictor (the average of these three) is used.

Since the interacting regions of a disordered and an ordered protein are inherently different I expect that the predictor will only recognize binding sites in disordered proteins that interact with globular proteins but are not part of globular proteins themselves. In order to verify this hypothesis I tested the combined final predictor on a dataset of complexes containing only ordered chains[16]. The prediction was done on the short interacting chain of the complexes. This gave a false positive rate of only 3.7% that is even lower than the value obtained on the testing set, although this might be only a consequence of the relatively small size of the ordered complex set (72 complexes). Overall, I could ensure that my predictor makes very few mistakes on both globular

proteins and complexes of globular proteins, while it can still recognize the majority of disordered binding regions. This implies that my algorithm is specific to disordered binding sites as opposed to globular proteins, the interface between globular proteins or disordered proteins in general.

4.1.4. Secondary structures and the efficiency of ANCHOR

I assessed the relationship between the efficiency of the prediction and the secondary structure adopted by the residues of disordered binding regions upon binding. For this purpose, I used three types of secondary structural element classes: helix (H, including α- and 3_{10} helices), extended (E) and coil (C, including everything else) as defined by the DSSP algorithm[85]. The number of amino acids in different conformations that can be found in the PDB structures of the positive training set (short disordered complexes), in the interacting residues of these structures and the interacting residues that are correctly identified by the predictor are shown in Figure 8. The secondary structure content in disordered binding regions is heavily biased towards coil conformation. It can also be seen on Figure 8 that the predictor seems to work slightly better for H and E conformations. However, assessing the difference of the distributions of secondary structures in interacting residues and in the subset identified correctly by ANCHOR shows that this difference is not statistically significant at a 5% level ($\chi^2 = 5.32$, $p = 0.070$). Furthermore, a similar result holds true if binding sites are categorized based on their dominant secondary structure type - that is there is no significant correlation between the secondary structure type the binding regions adopt upon binding and the efficiency of the predictor. Overall, this means that there is no significant difference in the efficiency of the prediction on different secondary structural elements.

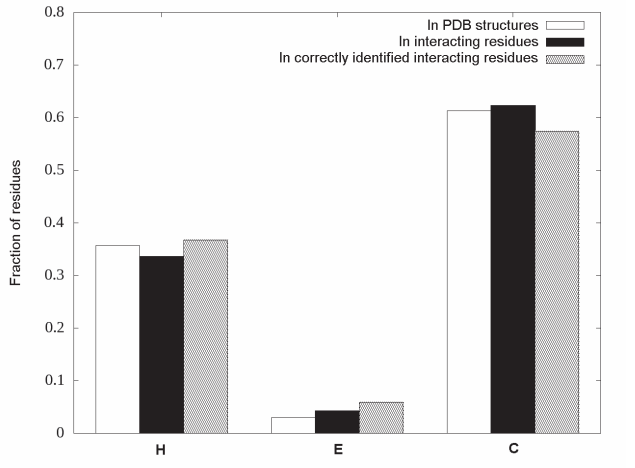

Figure 8: Secondary structure distributions in the short disordered binding site dataset

Fraction of amino acids in different secondary structures in the disordered chains of the complexes. The three groups denote the fractions calculated on all the residues in the PDB structures, only the interacting ones and the ones correctly identified by the predictor.

4.1.5. Testing on long, segmented binding regions

Since ANCHOR was trained on the short disordered dataset it is informative to see how it performs on long disordered binding sites (*Long disordered binding regions* dataset, see section 3.1 of Data and Methods). There is experimental evidence that at least some long disordered chains are not uniform concerning binding strength but contain short stretches of strongly interacting residues separated by segments that interact with the partner only weakly if at all[96]. In these cases, it is expected that the predictor will be unable to identify the weakly interacting parts since – though these parts may also form interchain contacts – they would not be able to bind to the partner in the absence of their sequential neighbors. The distribution of predicted binding regions for the short and long disordered chains in Figure 9A shows a strong preference for predicting multiple interacting regions for longer chains. This inevitably yields lower residue based TPR but

the segment based TPR is not expected to drop. Testing ANCHOR on the long disordered data confirms this assumption with a decreased residue based TPR of 47.7% (as opposed to 65.8% obtained on running the final predictor on the whole set of short disordered complexes) but with a basically unchanged segment based TPR of 78.6% (compared to the 76.1% calculated on short disordered complexes). These data suggest that ANCHOR either finds short disordered binding sites as a whole or completely misses it. However, this may not be true for long binding regions. Figure 9B shows the distribution of the fraction of amino acids successfully identified during prediction in the two types of binding sites. The effect can clearly be seen as about 59% of short binding regions are either fully recovered or are completely missed (the sum of the rightmost and leftmost columns) whereas this ratio is only about 29% for long binding sites.

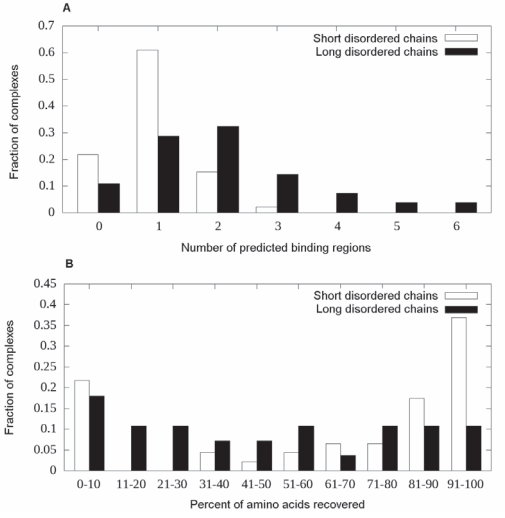

Figure 9: Prediction accuracies and segmentation for the short and long disordered binding sites
A) The distribution of the number of binding segments predicted in short (white bars) and long (black bars) binding sites. It shows the segmented nature of longer binding sites. B) The distribution of the fraction of correctly recovered interacting residues in both the short (white bars) and long (black bars) disordered binding sites.

I illustrate this type of behavior on the disordered human p27. This protein is involved in controlling eukaryotic cell division through interactions with cyclin-dependent kinases. Its kinase inhibitory domain binds both subunits of the CDK2-cyclin A complex in an extended conformation (PDB ID: 1jsu[88]). It is known from kinetic measurements that the binding of p27 is hierarchical through its three domains: first, the D1 domain (residues 25-36) binds to cyclinA which anchors the neighboring LH domain (residues 38-60) that exhibits transient helical structure in monomer state as well. After the binding of D1 this transient structure is stabilized and positions the rest of the chain (D2 domain, residues 62-90) in the correct position to bind to CDK2.

Figure 10 shows the prediction output for p27. Four interacting regions are identified with the first one (27-37) clearly corresponding to D1. The gap between the first two regions (38-58) coincides with the weakly interacting LH domain. The last three regions (59-67, 74-77 and 79-90) cover the strongly interacting D2. Figure 10 also shows the number of atomic contacts/residue for p27 (averaged in a window of size 3). This contact number profile exhibits well pronounced peaks that line up with the regions that are predicted by my algorithm. The figure also shows the four predicted regions mapped to the crystal structure of the complex.

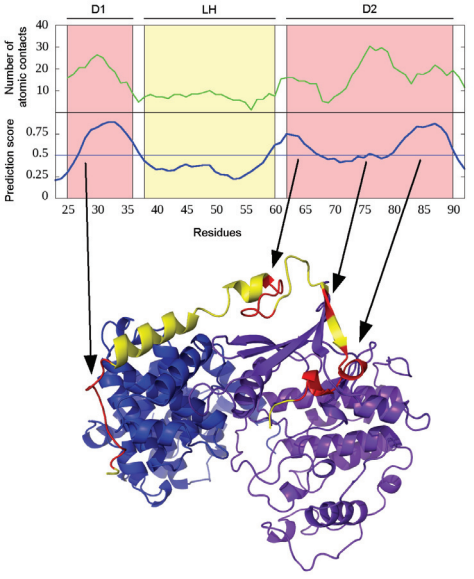

Figure 10: ANCHOR prediction for human p27

Top: Number of atomic contacts (green) and prediction output (blue) and for the N-terminal binding region of human p27. "D1"and "D2" denote the two strongly interacting domains (red boxes) and "LH" denotes the weakly interacting linker domain between them (yellow box).

Bottom: Crystal structure of human p27 (red and yellow) complexed with CDK2 (magenta) and Cyclin A (blue) (PDB ID: 1jsu[88]). Red parts denote regions that are predicted to bind by the predictor. These regions correspond to the experimentally verified strongly binding regions of p27. The figure was generated by PyMOL.

4.1.6. Discussion

With the development of ANCHOR I aimed to recognize disordered binding regions from the amino acid sequence. So far, the limited number of well characterized examples hindered the development of general prediction methods[97].

My approach relies on a basic physical model of disordered binding sites and it is based on modeling the interaction capacity in the free disordered state and in the bound ordered state. Previously, it was shown that ordered proteins can be discriminated from disordered proteins based on estimated pairwise energy content[60] and this approach was implemented in IUPred, a general disorder prediction method. This method takes into account that disorder/order tendency can be modulated by the sequential neighborhood simply at the level of amino acid composition, without attempting to model the specific interactions. Taking it one step further, I used the same energy estimation calculations to identify disordered binding regions in proteins. My model assumes that the specific properties of disordered binding sites are dictated by the combination of preferences to bind to an ordered protein on the one hand, and the ability to remain in a disordered state in isolation, on the other. Based on this simple model, ANCHOR achieved approximately 67% accuracy at predicting 5% false positive rate.

During binding, the formation of intermolecular contacts is accompanied by the formation or the stabilization of secondary structure elements. It was found that the adopted secondary structure can be predicted from the amino acid sequence with similar accuracy as in the case of globular proteins[98], suggesting that the adopted secondary structure can be imprinted into the sequence of the binding motif. However, the secondary structure observed in the complex can also be dictated by the template structure. An extreme example of this is the C-terminal region of p53, observed in all three secondary structure classes[99]. Hence, it is clear that not all of these conformations can be the result of inherent preferences. Interestingly, ANCHOR does not seem to be sensitive to the adopted secondary structure conformation and it works with the same accuracy for all secondary structure conformations. This independence of secondary structure elements underlines the generality of ANCHOR. These results also suggest that disordered binding sites can be recognized without taking into account of the adopted secondary structure in the majority of cases. Nevertheless, the details of conformational preferences can be still crucial in selecting the specific binding partner, or determining the kinetic and thermodynamic properties of the associations.

Beside ANCHOR, a previously published method called α-MoRF predictor also exploited a general disorder prediction method to recognize short binding elements[97; 100]. Although the direct comparison between the two methods was not possible, because the α-MoRF predictor is not publicly available, some basic differences between the two methods should be noted. First, the α-MoRF predictor directly relies on the prediction output of PONDR VXLT, which essentially predicts binding regions as ordered structural elements, and a subsequent neural network is applied to filter out valid disordered binding sites. Although very high accuracies were reported for the performance of the neural network based filtering, the complete method is limited by the efficiency of finding the local drops in predicted disorder tendencies (dips) based on PONDR VLXT. Therefore it should be taken into account that this program is a first generation prediction method that was trained on only 15 proteins. In the case of IUPred, dips corresponding to certain binding sites were also observed, although to a smaller extent[97]. This observation, however, is not directly exploited in ANCHOR. Instead, the core parameters of the energy prediction of IUPred are used to create three separate scores characterizing three important attributes of disordered binding regions. The second main difference is that ANCHOR is not restricted to a single secondary structure class like the α-MoRF predictor that was trained to recognize only α–helical segments. The example of the C-terminal region of p53, where four short overlapping regions were shown to bind in different conformations representing all three secondary structure classes, indicates that such restriction can be a serious disadvantage for recognizing some extremely adaptable disordered binding motifs.

In my work I assumed, that short binding regions undergoing disorder-to-order transition can be viewed as elementary binding units that are necessary for the molecular recognition. Therefore, such examples were used for the optimization of ANCHOR. In accordance with their elementary unit picture, ANCHOR recognized them generally as a single continuous binding site. Regions undergoing disorder-to-order transition, however, are not limited to such short segments as there are several examples of longer disordered segment becoming ordered upon complex formation. Such segments can be as long as 100 residues. However, these longer regions can contain segments which bind only

weakly or might not become ordered at all[101; 102]. This segmentation of longer binding regions can occur for structural reasons. The segmentation can prevent the accumulation of the critical amount of residues that would lead to the formation a collapsed structure or non-specific aggregates. The possible functional advantages of the segmented nature of a binding site were demonstrated for the well characterized example of p27. The segmented nature of binding is reflected in the prediction output, with predicted binding sites corresponding to the strongly interacting regions. In the dataset of longer disordered binding segments, I found this segmentation to be quite general. In these cases, the predicted sites generally give only partial coverage of the PDB structure, and multiple binding sites are predicted in the majority of cases. This suggests that ANCHOR is likely to find those sites that interact more strongly, anchoring the disordered segments to their partner protein.

The success of ANCHOR has both technical and theoretical implications. Apart from the applications that will be discussed in later chapters, from a theoretical point of view, the relatively high accuracy of the method indicates that the underlying simplified biophysical model is capable of describing the majority of disordered binding regions. The basis of the description is that these regions can be characterized by highly disordered sequential neighborhood, unfavorable intrachain energies and more favorable interaction energies with a globular partner. The resulting model is accurate and general enough to recognize the majority of disordered binding sites independent of their secondary structure or amino acid composition. As such binding sites are essential functional elements of disordered proteins, their prediction directly provides information about functionally important residues in these proteins. In this way, ANCHOR broadens the repertoire of prediction methods for functional sites in proteins aiming to decrease the large number of unannotated sequences. Generally, the complete understanding of protein-protein interactions involving disordered binding sites requires the knowledge of their partners as well as possible post-translational modifications that can influence their binding. While predictions can be made even without taking the partner molecule into account, certain cases might require incorporating the specific feature of the partner.

Nevertheless, ANCHOR can provide the starting point for such scientific explorations, by finding potential regions involved in such binding.

4.1.7. Availability and the ANCHOR server

Following the publication of ANCHOR, to better target the wider scientific community, ANCHOR was put online in the form of a web-server[103]. The server is freely accessible and offers the option to download the ANCHOR program for local use as well. This does not require registration and is also free of charge for academic users. The server is complete with the short description of the method itself, help pages and examples to aid the users in the efficient use of ANCHOR.

ANCHOR is hosted on the servers of the Institute of Enzymology and is accessible at http://anchor.enzim.hu. The minimum input of the web server is a single amino acid sequence. Sequences can also be specified by their corresponding UniProt IDs or ACs. A list of motifs can also be submitted, specified as regular expressions with or without their names. A few examples, including known eukaryotic linear motifs are given in the help to guide the user with the format. The motif search, however, is not restricted to known linear motifs, any kind of regular expression can be specified.

The basic output of the server is the probability score, indicating the likelihood of the residue to be part of a disordered binding region along each position in the sequence. The returned plot shows the prediction profile calculated by ANCHOR and also incorporates the disorder profile calculated using IUPred. Predicted disordered binding regions and matched motifs are also indicated underneath the profile as horizontal bars. The graphical output is followed by a simple text output, summarizing the predicted and filtered binding regions, the location of the found motifs and the returned prediction profile. An example for the graphical output is presented on Figure 11.

I wrote the core program of ANCHOR in C, while motif searches are carried out by a Perl wrapper. This program is called by the web server written in PHP. The graphical

output is generated by the JpGraph software. The default option for graphical/text output is automatically determined by the browser type, but it can be changed by user. Additionally, list of sequences can also be submitted to generate simple text output on a larger scale.

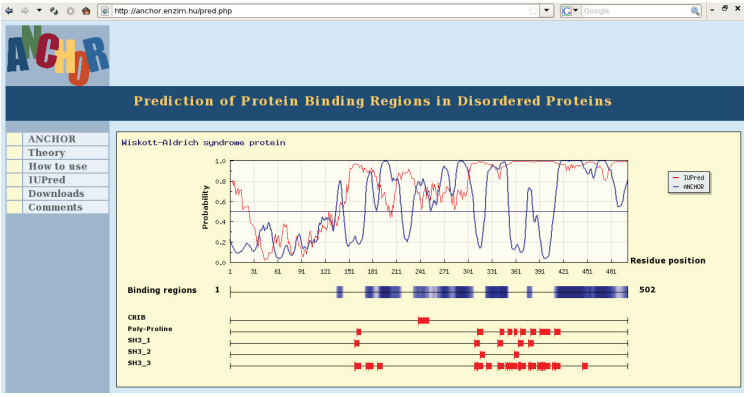

Figure 11: An example of the ANCHOR server graphical output

The human Wiskott-Aldrich Syndrome protein (WASp) was used as an input with various motif searches. The N-terminal of the protein contains an ordered domain, otherwise it is largely disordered. Red line shows the disorder tendency and blue line shows the ANCHOR prediction. Predicted binding regions are characterized by scores above 0.5 and a condensed output shows predicted binding regions under the prediction profiles with blue boxes. In WASp multiple disordered binding regions were predicted, and several of these can be confirmed experimentally. The results of the motif searches shown with red bars, show regions containing various SH3 binding sites as specified in the ELM database. Additionally, proline rich regions and the CRIB motif implicated in binding to Cdc42 can also be located.

4.2. Biological application of ANCHOR on whole proteomes

Apart from the study of individual proteins, ANCHOR opened up new ways to analyze biological data on a larger scale as well, making it possible to gain insights about disordered binding regions at an evolutionary level. Following the completion of ANCHOR, I studied the appearance of protein disorder and disordered binding regions throughout evolution[33] by employing a large scale scan using ANCHOR on a set of 736 complete proteomes (53 archaea, 639 bacteria and 44 eukaryota, see Data and Methods) that were currently available from the SwissProt database (ftp://ftp.expasy.org/) as of 2009. In agreement with previous analyses[32; 87] there is a clear trend of increasing amount of protein disorder as the complexity of the organism increases (see Figure 12). However, Figure 12 also shows that the fraction of disordered amino acids predicted to be in disordered binding sites increases even compared to fraction of disordered residues, as the complexity of organisms grows. Generally, archaea have the least amount of both disorder and binding sites. On the other hand, eukaryota have generally the largest ratio of disordered and binding amino acids with bacteria being between these two groups on average. However there are a few exceptions to these general trends, marked separately on Figure 12.

Considering archaea, mesophiles generally contain a larger amount of disorder and a larger fraction of disordered binding sites than most extremophiles (thermophiles, cryophiles and acidiphiles). However the group of halophile archaea (archaea that favor high saline concentration) is a distinct exception with fraction of disordered amino acids ranging from 0.2 to 0.25 as opposed to other extremophiles' values not exceeding 0.07. This group includes all the halophile archaea in my study, namely *Natronomonas pharaonis*, *Haloarcula marismortui*, *Haloquadratum walsbyi* and two types of *Halobacterium salinarum*. *Cenarchaeum symbiosum*, the only example of obligate endosymbiont among archaea also has an unusually large amount of disordered protein segments in its proteome (0.12). While *Cenarchaeum symbiosum* is closely related to

thermophile archaeas, it is adopted to the much lower living temperature of its host. This adaptation could explain the relatively large amount protein disorder and disordered binding sites. In general, these clear differences in the predicted disorder between various archaea organisms points to different strategies to adapt to various extreme environmental conditions resulting in biased amino acid compositions. However, it cannot be ruled out that under such extreme conditions, as high salt concentration or high temperature, the amount of disorder can be over- or under-predicted depending how these conditions affect the presence of protein disorder.

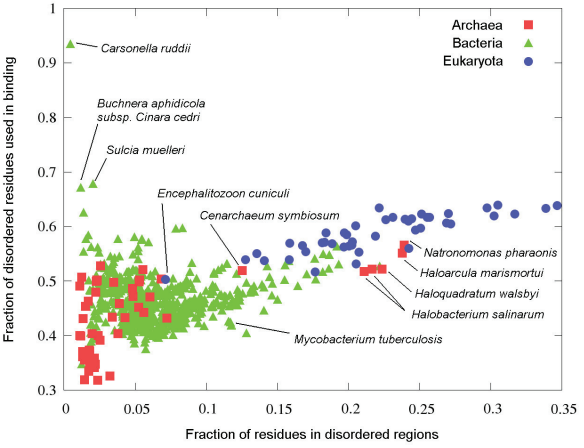

Figure 12: Fraction of disordered and disordered binding site residues in complete proteomes
The number of amino acids in disordered binding sites divided by the number of amino acids in disordered regions plotted as a function of the number of amino acids in disordered regions divided by the total number of residues in the proteome of the organism for the 736 complete proteomes deposited in the SwissProt database, colored according to the three kingdoms of life. The outlying points are marked with the name of the corresponding organism.

Among bacterial proteomes, there are a few examples of organisms that seem to utilize a surprisingly large fraction of their disordered amino acids in binding. The three most extreme cases (*Carsonella ruddii*, *Sulcia muelleri* and *Buchnera aphidicola subsp. Cinara cedri*) are marked separately on Figure 12. These are the three smallest complete

bacterial proteomes, none of them reaching the size of the smallest archaea proteome. These organisms present extreme cases of streamlined genomes as a result of endosymbiosis. As these proteomes are very small, the predicted amount of disorder and disordered binding sites are within the false positive range, and should be treated more cautiously. Additionally, some other bacteria are hallmarked by an unusually high ratio of protein disorder. One such case is *Mycobacterium tuberculosis*, the main causative agent of TB that – in terms of ratio of disordered residues – ranks among the top 10% of bacteria. A separate analysis of the proteome of *Mycobacterium tuberculosis* is presented in chapter 4.3.

Eukaryotes tend to appear more consistent in using both larger amount of disordered residues and larger fraction of disordered residues for binding compared to the other two kingdoms (Figure 12). The only notable outlier both in terms of extremely low amount disordered proteins and disordered binding sites is *Encephalitozoon cuniculi*. This organism is the only microsporidian parasite in the dataset and has an extremely small proteome. This lack of complexity and dependence on a eukaryotic host to function might explain the lack of disordered proteins.

I also analyzed the length distributions of the predicted disordered regions and binding sites in the three kingdoms of life. These results are shown in Figure 13A and Figure 13B, respectively. As complexity increases, longer disordered segments are preferred, and the difference between eukaryota and lower complexity organisms becomes even more apparent for longer regions (over 30 residues). A similar trend can be observed in the length distribution of disordered binding sites. While in archaea and bacteria predicted binding regions are generally below 30 residues, longer binding sites in eukaryota organisms are much more common. There are at least three different effects that can contribute to this phenomenon. First, as the number of binding sites rise there is also an increasing possibility of these binding sites becoming very close to each other or even overlapping with each other. This scenario was demonstrated in the case of the N-terminal domain of p53 (see Section 4.1.1, Figure 6). Second, extremely large disordered binding regions may be needed for special functions. Some members of the mucin protein

family provide an example for this. Human MUC1 contains a large repeat region (20-120 repeats, one repeat being 20 amino acids long) that enables it to aggregate and to perform its function. As each repeat is correctly identified as a disordered binding site, the whole repeat region is predicted as one large binding region. This mechanism can create binding sites up to the length of several hundreds of residues in extreme cases. Third, it cannot be excluded that longer binding sites are not always segmented by weakly interacting regions thus forming long, continuous binding regions. Nevertheless, the majority of predicted binding sites is shorter than 30 residues, although such restriction on the length of disordered binding sites was not enforced.

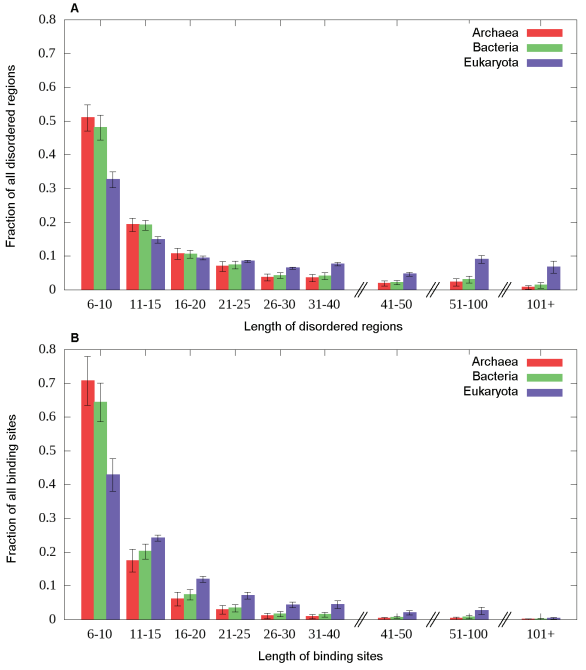

Figure 13: Length distribution of disordered and disordered binding sites in complete proteomes

The length distribution of A) the disordered protein segments determined by IUPred and B) predicted disordered binding sites determined by ANCHOR for the 736 complete proteomes available, grouped according to the three kingdoms of life.

4.3. The effect of protein modularity in pathogen virulence: a case study of Mycobacterium tuberculosis

Out of the several identified organisms harboring an unusually high ratio of residues in disordered and disordered binding regions (discussed in section 4.2), I analyzed the proteome of *Mycobacterium tuberculosis* (MTB) in detail[104]. MTB is the main causative agent of TB, a disease that demands 2 million human lives worldwide annually[105]. As a result of the lengthy co-evolution of *Homo sapiens* and MTB, the bacterium became a dramatically successful pathogen species that presents considerable challenge for modern medicine[106]. The continuous and ever increasing appearance of multi-drug resistant mycobacteria necessitates the identification of novel drug targets and drugs with new mechanisms of action[107]. However, further insights are needed to establish automated protocols for target selection based on the available complete genome sequences.

To uncover the factors resulting in the success of MTB, I employed a proteome-wide analysis. As a first step, as already presented in chapter 4.2, I calculated the amount of protein disorder using IUPred and the amount of disordered binding regions using ANCHOR. At the residue level, 11.8 % and 5.7 % of residues were predicted to belong to a disordered segment or a disordered binding region, respectively. Although these values were relatively small, they represented significantly higher values compared to many other bacteria (for reference data see Section 4.2, Figure 12). The fraction of disordered proteins and disordered binding regions were even comparable to that of simpler eukaryotes.

4.3.1. Similarity based clustering of MTB proteins

The uncovering of proteins involved in species-specific processes is usually focused on identifying proteins that are unique to the organism and have no homologs in other organisms[108; 109]. Despite its rationale, this approach has strong limitations as proteins are

highly modular and species-specific processes can be brought about not only by unique proteins but by unique combination of otherwise ubiquitous domains and protein regions as well. This can be shown specifically for MTB by analyzing the domains present in MTB proteins, using the Pfam database (see Data and Methods). Figure 14 shows the organism specificity of domains present in the MTB proteome. Altogether only 5 of the total 2099 different domains of MTB are species-specific, being present in only MTB or the highly similar *Mycobacterium bovis* and cannot be found in any other organism. The majority of domains however, are ubiquitous among bacteria and eukaryotes with 812 of them present in the human proteome as well. This evident lack of MTB specific protein building blocks calls for a different approach at pinpointing proteins responsible for the unique properties of MTB.

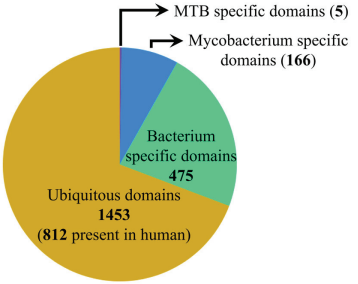

Figure 14: Occurrences of domains of *M. tuberculosis* in other organisms

The distribution of the 2099 Pfam domains present in the proteome of MTB in Eukaryotes and Bacteria. Slices of the pie chart correspond to different levels of specificity with purple showing domains that can be found exclusively in MTB, blue and green showing domains found in mycobacteria or in bacteria in general, respectively and orange showing ubiquitous domains that can be found in organisms from MTB to eukaryotes. Numbers of domains are given for each slice, with number in parenthesis for ubiquitous domains showing the number of domains present in human proteins.

As an alternative approach, I carried out a large-scale sequence similarity search for all proteins in MTB by comparing them to the proteomes of a wide range of other organisms (see Data and Methods). By virtue of this analysis, the number of similar

proteins in other bacterial or eukaryotic proteomes was determined for each protein present in MTB. To capture the modularity of MTB proteins, I applied local similarity searches (see Data and Methods). Generally, the number of sequences similar to an MTB protein sequence across various species can show quite large variations. The various scenarios include organism-specific proteins, nearly ubiquitous proteins for which the number of homologs is relatively constant from bacteria to eukaryotes, and many other cases for which significant enrichment/depletion of certain protein families can be seen at certain points in evolution.

In order to identify some of the basic trends, I carried out a cluster analysis of the similarity profiles of MTB proteins (see Data and Methods). Similarity profiles were constructed using the number of similar sequences in the proteome of other organisms of each MTB protein. This is in contrast with the binary profiles commonly used in phylogenetic profiling[110; 111]. Using the results of the clustering, I constructed a hierarchical tree. This tree could be dissected into major branches, grouping the MTB proteins into distinct clusters.

This analysis identified two groups of proteins that showed highly unusual evolutionary profiles. One of these represents a group of proteins that are present in MTB in a large number, but completely missing from bacteria other than mycobacteria and are generally not present in eukaryotes either. All of these proteins belong to a mycobacteria specific class of PE/PPE proteins[112]. The hallmark of the proteins of the other group is that they have an exceptionally high number of similar sequences in eukaryotes. This cluster is comprised by the pkn protein family that is defined by the presence of a eukaryotic-like kinase domain that enables these proteins to be involved in regulatory processes[113].

The two protein groups of pkn and PE/PPE families stand out in several respects. The homologs of the pkn family are more common in eukaryotes, while members of the PE/PPE family are basically mycobacterium specific. However, both of these groups show a drastic domain enrichment in MTB. Beside their very unusual evolutionary

profiles, they also exhibit high disorder content. Both of these properties could indicate their functional importance. Further insights can be gained by looking at the functional and structural properties of these two families in more detail.

4.3.2. pkn protein family

Members of the pkn family belong to the group of eukaryotic-like Ser/Thr protein kinases (STPKs)[113]. Originally these proteins were thought to be unique to eukaryotes, however, the accumulation of genomic sequences revealed that some prokaryotes also contain members of this group. The bacterial signaling pathways usually rely on two-component systems, basically consisting of a sensor histidine kinase and a response regulator. The eukaryotic-like protein kinase genes, however, represent an independent, additional mode of bacterial regulation. In mycobacteria, genome sequence data indicate that the number of STPK genes is in fact either commeasurable or even considerably higher than those representing the usual bacterial two-component system genes[114]. In the MTB genome, 11 STPK genes can be identified (from pknA to pknL) and with the exception of pknG and pknK, all of these proteins are highly probable to be localized to the membrane. Furthermore, members of the pkn family exhibit a significant amount of disorder and contain a large number of disordered binding regions. Although functional annotation of pkn proteins remain scarce, they are reported to be involved in a wide range of functions, including cell elongation, growth and division, regulation of lipid biosynthesis, membrane transport, nitric acid stress response, regulation of glucose transport and the barrier septum formation, transcriptional regulation, regulation of DNA binding and response to stress and host immune response.

Reflecting the functional diversity of this family, members of the pkn family are different structurally as well. Atomic level information is available for the pknB, pknD and pknG proteins. Apart from the kinase domain, several pkn proteins contain additional domains, such as PASTA or NHL domains. Of special interest is the soluble pknG protein which consists of a rubredoxin and a tetratricopeptide (TPR) domain flanking the

kinase domain. The rubredoxin domain was found to be essential for the function and might be responsible for regulating the activity of pknG depending on the redox state of the environment. Although the exact function of the TPR domain in this case is unknown, pknG was experimentally shown to be essential for avoiding the degradation of MTB cell in macrophages by disrupting the fusion of MTB with lysosomes[115].

4.3.3. PE/PPE protein family

PE and PPE proteins represent the most variable group of proteins in pathogenic mycobacteria[112]. The PE/PPE protein family contains 167 members and can be further divided into the PE, PE-PGRS and the PPE protein sub-groups (with 35, 64 and 68 members, respectively). Almost all proteins contain a domain at the N-terminal region that defines the sub-group (PE domains in the PE and PE-PGRS groups and PPE domains in the PPE group) and many PE/PPE proteins incorporate other domains as well. Accordingly, some PE/PPE proteins are highly modular and can be up to 3300 residues in length, and their structural and functional characterization is definitely of importance.

In vivo essentiality screens showed that several of the PE/PPE proteins are essential for growth[116]. Due to their variability these proteins are regarded as a possible source of variable surface antigens which provide a means to exploit and possibly escape the host immune system during pathogenesis[117]. Although the exact function of none of the PE/PPE proteins or of their complexes has been revealed, available findings delineate a consistent picture which suggests that the PE/PPE proteins are involved in a highly plastic host-pathogen interaction network[112]. Although, despite their importance, these proteins comprise a yet greatly unexplored area as both structural and functional data concerning them are scarce.

My analysis showed that protein disorder is not homogeneously present in all three sub-groups (PE, PPE and PE-PGRS). The majority of the disordered regions can be found in the PE-PGRS proteins. Although most disordered parts do not include any

predicted Pfam domains, some domains significantly overlap with these regions. On the other hand, some domains, such as the α/β hydrolase domain together with various Pfam-B domains of unknown function seem to be entirely ordered and hence might lend themselves to traditional structure determination possibly yielding potential drug targets.

4.3.4. Implications for target selection in drug design

The presented comparative genomic study based on the result of large-scale sequence similarity searches is completely general and could be applied to any kind of organism with an annotated genome. In my work I focused on MTB, the causative agent of tuberculosis. My analyses revealed two protein families in the proteome of MTB that stand out in several aspects. These proteins were also shown to have a functional importance essential for the survival of this pathogen and can be potential targets for drug design[118].

The common properties of both the pkn and PE/PPE families include unusual domain accretions specific to this organism. This is combined with an increase in their disorder content. Both families carry out important functions in the MTB and are involved in the interactions with the host cell. Various members were shown to be essential for the organism[116] and according to a recent analysis using guinea pig model, representatives of these families are significantly enriched in the early and chronic stages of infections[118]. Furthermore, many of them are either located in the surface of the bacteria or are exported into the host cell. The properties of these protein families underscore their biological importance and suggest that they would be ideal candidates for drug design. However, conventional drug design procedures generally overlooked such proteins as targets by largely focusing on metabolic processes. The need for novel drugs for the treatment of MTB forces researchers to explore new directions for target selection. The pkn and PE/PPE families, through their complex architectures offer several options in this regard.

Both pkn and PE/PPE proteins contain long disordered segments. Until recently, the feasibility of targeting proteins without a well-defined structure was unclear for the purpose of drug development. There is now, however, a newly sparked interest in intrinsically disordered proteins as potential drug targets[50; 51]. The low binding free energy of these interactions indicates that they would be relatively easy to block by small molecules[50]. Generally, the analysis of known examples of the druggable regions of disordered proteins indicated that these segments overlapped with the binding regions predicted by ANCHOR[119]. Therefore, ANCHOR and other disordered binding region prediction algorithms that will be hopefully developed in the years to come can be extremely useful to highlight potential druggable sites directly from the amino acid sequence, especially in combination with other methods.

Although some of my findings are specific to MTB, there are several more general implications of this study. The exclusivity of certain proteins to a given pathogen is often one of prime criteria used in various target selection protocols. However, my results indicate that species-specific functions are not necessarily brought about by species-specific proteins. In contrast, many novel functions developed from already existing proteins. In the case of eukaryotes, there are several notable examples, such as the development of olfaction, reproduction, and immunity[120], where the combination of gene duplication, divergence and recombination led to the expansion of protein families and provided jumping points in evolution. The example of MTB shows that such complex evolutionary scenarios play important roles in prokaryotes as well and can be detected by species-specific enrichment of certain protein domains or families. Protein families emerging as a result of such processes often have complex domain architectures. Consequently, these proteins can be approached from multiple directions for the purpose of drug development and taking the various factors into account can help to improve the success rate of target selection protocols and drug development process.

4.4. Large scale analysis of protein disorder, protein function and involvement in cancer

As demonstrated in the previous chapter, protein disorder can play a significant role in the pathogenicity of certain bacteria. Consequently, the presence of protein disorder has been linked to various infections. However, disorder has been linked to other classes of diseases as well that can develop without pathogens, including diabetes, neurodegenerative diseases and cancer[43] (see section 1.3.3). In these cases, the correlation between protein disorder and the development of the disease has been shown[48; 49]. However, correlation does not imply causality and hence the popular claims of protein disorder imposing a 'biological risk' or 'biological cost' are unfounded at best. To address this question, I analyzed the link between protein disorder, disordered binding regions and the involvement in cancer concentrating on human proteins and their cancer-associated mutations[121].

4.4.1. Data collection

In this study, genetic variations were restricted to single amino acid substitutions, therefore proteins that were associated with cancer via chromosomal translocations or copy number variations were not considered. The dataset of missense mutations was compiled from the COSMIC database[81] (COSMIC, see Data and Methods). It included cancer mutation data collected both from the literature and the outcomes of large-scale cancer genome projects. An additional dataset corresponded to a more restrictive subset of proteins in COSMIC that were part of cancer census genes. These proteins could be casually linked to oncogenesis[122] (COSMIC_census). I also assembled a database of neutral mutations (polymorphisms), taken from the UCSC Genome Browser[82] (see Data and Methods). The number of proteins, amino acids and mutations in each dataset are given in Table 3.

Table 3: Datasets used in the analysis of cancer associated mutations

Datasets	Number of			
	proteins	residues	mutations	polymorphisms
COSMIC	8 957	6 898 559	22 708	26 435
COSMIC_census	261	238 130	5 375	673

The number of proteins, residues, mutations and polymorphisms are shown for the cancer-associated mutation databases.

4.4.2. Protein disorder in cancer-associated proteins

I evaluated the disorder content in the datasets to confirm that protein disorder is common in human cancer-associated proteins using the complete human proteome as reference. The disorder content was calculated using IUPred. Figure 15 shows the disorder content and the percentage of proteins with disordered regions over 30 residues, as well as the average length of proteins in the various datasets as compared to the average values of the human proteome (see Data and Methods) obtained with IUPred. Contrary to previous results[47], the overall disorder content of the full COSMIC database was not elevated compared to the reference. However, when restricting the analysis to the census part of COSMIC, the obtained results are in agreement with earlier results, with the percentage of disordered residues being significantly higher (Figure 15). These results did not depend on the choice of the disorder prediction software, as DISOPRED and VSL2, two other fundamentally different methods produced remarkably similar outputs (data not shown).

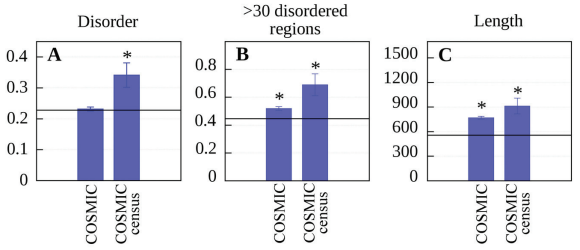

Figure 15: Length distribution and disorder content of cancer associated proteins

Average ratio of disordered residues (A), ratio of proteins containing >30 residue long disordered regions (B) and length (C) in the datasets analyzed. Black horizontal lines represent the average values obtained for the proteins of the human proteome taken from SwissProt. Flags show the confidence interval of α=0.01 calculated from the standard error of the mean of random selected samples from the human proteome (see Data and methods). Significant differences are marked with an asterisk.

There was, however, a significant increase in the proportion of proteins containing long disordered segments among both COSMIC and COSMIC_census proteins compared to the human proteome. The results calculated with IUPred (Figure 15B) were again confirmed by the two other prediction methods (data not shown). In agreement with earlier results, cancer-associated proteins were also significantly longer. The increase in length and in fraction of proteins with long disordered segments points to the increased modularity and complexity of cancer-associated proteins.

4.4.3. Polymorphisms and cancer-associated mutations in ordered, disordered, and disordered binding regions

The rates of evolution are largely governed by the stringency of functional and structural constraints. As ordered and disordered segments in proteins have distinct properties in these regards, these characteristic differences are expected to be reflected in the distribution of genetic variations in these regions. To test this assumption, I analyzed the differences in the distribution of polymorphisms (SNPs) and cancer-associated

mutations within ordered, disordered and disordered binding regions of cancer-associated proteins using IUPred and ANCHOR. Residues were categorized into three groups: residues predicted by ANCHOR were considered to be a part of 'binding regions', the rest of the residues were either grouped to 'disordered regions' or 'ordered regions' based on IUPred predictions.

For each protein in the datasets, I tallied the number of observed polymorphisms and cancer-associated mutations in ordered, disordered and disordered binding segments. These numbers were compared to the expected number of polymorphisms based on the assumption that the mutations are distributed evenly in the sequence (see Data and Methods). The results presented on Figure 16A show the relative difference between the observed and expected number of polymorphisms.

There are significant differences among the three sets in the distributions of observed SNPs (Figure 16A). While SNPs were clearly overrepresented in disordered segments and underrepresented in ordered regions, disordered binding regions fell between these two categories, but their behavior was still closer to disordered segments. These data are in agreement with the basic assumption that neutral polymorphisms are less likely to occur in positions with stronger structural and functional constraints. In globular proteins, functionally relevant sites are often restricted to a few residues that form the active site, but nearly all residues contribute to the formation of the 3D structure at some level. This represents a large evolutionary constraint for globular proteins. Functionally important residues of IDPs, such as residues directly involved in binding or undergoing post-translational modifications, can experience constraints similar to the active sites of globular proteins. In terms of structural constraints, however, mutations generally are expected to have smaller impact on the structural properties of disordered segments, due to the lack of well-defined structure. Accordingly, disordered proteins exhibit a lower evolutionary conservation, observed at various levels[123; 124].

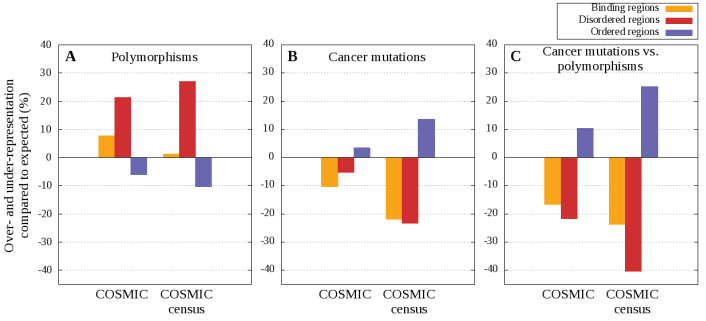

Figure 16: Distribution of polymorphisms and cancer-associated mutations

Over- and under-representation of mutations in disordered binding regions (orange), disordered (red) and ordered regions (blue) calculated with ANCHOR, as compared to background distributions (see Data and Methods). (A) the distribution of polymorphisms as compared to the uniform random distribution; (B) the distribution of cancer-associated mutations as compared to the uniform random distribution and (C) the distribution of cancer-associated mutations as compared to the expected values weighted by the distribution of polymorphisms shown in (A). All differences were significant.

Compared to polymorphisms, cancer-associated mutations followed a reversed trend and were more likely to appear within ordered regions (Figure 16B). Using the distribution of SNPs as an expected distribution for cancer-associated mutations (instead of the random distribution, see Data and Methods), these differences became even more pronounced (Figure 16C).

Together with the results obtained on the distribution of polymorphisms, these results suggest that disordered residues are more tolerant to mutations at two levels. First, disordered regions can allow a larger number of genetic variations without affecting the function. Second, if a mutation occurs, it is more likely to cause cancer if the affected residue is located within an ordered region. The lower sensitivity of disordered regions to genetic variations is likely to originate from the specific structural properties of these regions. The analysis of disordered binding regions showed that functionally relevant sites within disordered regions can slightly deviate from this behavior. Disordered

binding regions could be placed between disordered regions in general and ordered regions, both in terms of the appearance of polymorphisms and cancer associated mutations. These suggest stronger evolutionary constraints within disordered binding regions, in accordance with their functional importance. Nevertheless, within the broader context of binding regions, only a few residues might be directly responsible for the specificity of the binding and these residues could present even higher evolutionary constraints. Altogether, these results clearly contradicted the original hypothesis about the increased risk of cancer associated with protein disorder, at least in terms of single nucleotide mutations.

4.4.4. Functional correlations

I also analyzed cancer-associated proteins in terms of their functional categories and their number of protein-protein interactions. First, I assessed which functional groups were overrepresented within cancer-associated proteins. For this analysis, the GeneOntology[83] functional categories were used (see Data and methods). The occurrence of each of the considered 50 biological processes and 41 molecular functions in the COSMIC_census dataset was compared to the expected occurrence of these functions in the human proteome. The list of biological processes and molecular functions that exhibited statistically significant differences is shown in Table 4. The significantly enriched processes among cancer-associated proteins included signal transduction, involvement in cell-cycle and proliferation, DNA- and protein binding, phosphorylation and regulation of transcription. These proteins on the other hand were significantly depleted in transport processes in general and particularly in ion transport. In other cases, the differences were not significant at the α=0.01 level. In general, my results are in complete agreement with an earlier study, and correlate well with the functional enrichments of disordered proteins.

Table 4: Significant annotations of COSMIC and COSMIC_census proteins

	GO ID	Description	Number of COSMIC census proteins with the given term	Expected number of proteins with the given term	p-value	Over- or under-representation
Biological processes	GO:0007165	signal transduction	51	26	$1.418*10^{-3}$	0.96
	GO:0008283	cell proliferation	17	4	$3.055*10^{-3}$	3.25
	GO:0006811	ion transport	0	8	$3.696*10^{-3}$	-1.00
	GO:0006810	transport	9	24	$5.370*10^{-3}$	-0.63
	GO:0007049	cell cycle	20	7	$8.084*10^{-3}$	1.86
Molecular functions	GO:0005515	protein binding	184	65	$1.305*10^{-26}$	1.83
	GO:0003677	DNA binding	84	27	$4.907*10^{-10}$	2.11
	GO:0000166	nucleotide binding	72	25	$6.844*10^{-8}$	1.88
	GO:0004672	protein kinase activity	36	6	$5.573*10^{-7}$	5.00
	GO:0003700	transcription factor activity	44	12	$3.463*10^{-6}$	2.67
	GO:0016301	kinase activity	37	8	$3.192*10^{-6}$	3.63
	GO:0016740	transferase activity	48	18	$5.276*10^{-5}$	1.67
	GO:0030528	transcription regulator activity	17	5	$7.340*10^{-3}$	2.40

List of GO biological processes and molecular functions that are significantly over- or under-represented in the COSMIC census database as compared to the human proteome (see Data and methods). p-values were obtained using the exact Fisher test.

Cancer-associated proteins represent a specific group of proteins that are enriched in certain functions, contain more disordered regions, generally are longer and are involved in a larger number of interactions (25.5/protein as compared to 5.5/protein in the human proteome). However, all these features also correlate with each other. To untangle these complicated relationships, I studied the association between these distinct features. Specifically, I considered the length of the protein, the ratio of its residues residing in disordered segments or disordered binding regions, the number of cancer-associated mutations taken from the COSMIC_census database and the number of protein-protein interactions as well as the above identified significant functional classes (see Data and methods). The mutual information and the Jaccard distance were calculated between all pairs of features. The obtained distances between the different features are shown in Table 5. These distances were also subject to multidimensional scaling to reduce the dimensionality to two. The resulting scaled location of each feature is presented in Figure 17.

Table 5: Jaccard distances of features

	Length	Disorder %	Binding regions %	COSMIC census mutations	Interactions	Functions
Length	0.0000	0.9871	0.9860	0.9597	0.9776	0.9157
Disorder %		0.0000	0.5170	0.9753	0.9896	0.9208
Binding regions %			0.0000	0.9732	0.9860	0.9162
COSMIC census mutations				0.0000	0.9444	0.8808
Interactions					0.0000	0.8670
Functions						0.0000

Jaccard distances of the 6 features calculated on the COSMIC census database as compared to the human proteome (see Data and methods).

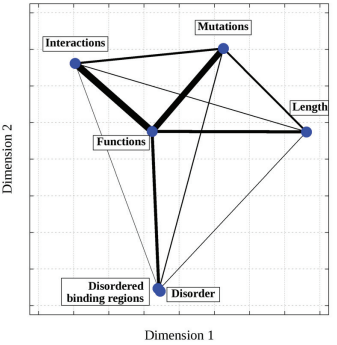

Figure 17: Two dimensional mapping of various features based on the distances calculated on the COSMIC census database relative to the human proteome

Coordinates were obtained using multidimensional scaling (see Data and methods) by projecting the original Jaccard distances into two dimensions. The widths of the connecting lines are inversely proportional to the original Jaccard distances (see Table 5).

It can be seen that the association between the ratio of residues in disordered regions and disordered binding sites is the highest indicating the relatively constant ratio of disordered residues that are involved in binding. Apart from this strong association, the functional features shared the most information with all the other features. This indicated the central role of function that largely determines the disorder content together with the

amount of disordered binding regions, the number of protein-protein interactions, the required length for a given protein and its involvement in cancer.

In conclusion, my results clearly show that protein disorder in itself is not responsible for the increased biological risk in terms of cancer-associated mutations. It seems plausible that the functional involvement of a protein determines both its disorder content and its involvement in cancer, thus presenting a correlation between these two features, without an existing casual link between them. My study was restricted to single amino acid changes, however, other type of genetic alterations can also lead to cancer.

4.4.5. Connection with other types of genetic variations

My general finding is in contrast with the results obtained in the analyses of another major form of genetic aberrations leading to cancer, chromosomal translocations[48]. In this case, a direct link between disorder and cancer was found. This was rationalized based on that ordered proteins are more likely to be misfolded and degraded as a result of translocation, while disordered proteins could survive with an aberrant function. A third form of commonly occurring genetic variations is copy number variation (CNV), which corresponds to the enrichment or depletion of certain genomic regions. CNVs are frequently observed in cancer and other diseases. In a recent study, a strong correlation between dosage sensitive gene products and protein disorder was found, and it was related to the interaction promiscuity of IDPs[49]. In order to resolve these seemingly contradictory results, cancer-associated mutations have to be placed into a network context. The network view was also suggested to be crucial in order to reduce the complexity of the landscape of cancer genomes. The exploration of the role of protein disorder in these cases necessitates many further studies and taking into account the specific functions of these proteins and the way they are regulated. The present work, nevertheless, demonstrated that genetic mutations affect ordered and disordered regions in different ways, in accordance with the distinct structural and functional properties of these segments. In order to understand the background of various diseases, these differences have to be taken into account.

4.5. Disordered binding regions and linear motifs – bridging the gap between two models of molecular recognition

The disordered binding region and the linear motif concepts (introduced in sections 1.5 and 1.6, respectively) describe molecular interactions on different bases: the former focusing on the structure (or the lack and formation of it) and the latter approaching the problem through the sequence. However, the interactions described by the two concepts share a high degree of similarity. In both cases the interaction is confined to a relatively short, linear sequence region in one of the partners. Additionally, many known linear motif instances were shown to reside in disordered protein regions[125]. Accordingly, in many cases the same interaction was categorized as an example of both linear motif mediated binding and of disordered binding regions. Examples include the binding of p53 to MDM2 and the N terminal region of p27 binding to the cyclinB-CDK2 complex. However, despite the growing number of common examples, the complementarity of linear motifs and disordered binding regions has not yet been directly addressed.

In this section I study this connection through two prediction methods, each tailored specially for identifying the respective type of interaction sites. Linear motif searches are carried out by using regular expressions taken from the ELM database[68] and disordered binding regions are identified by ANCHOR[33]. Through the overlap of these two approaches I set out to take the next step in the integration of the two concepts.

4.5.1. Predictive power of linear motifs

One of the main limitations of using linear motifs in the prediction of protein-protein binding regions is the weak definition of the motifs. Basic pattern-matching scans through databases are hindered by the overwhelming number of false positive hits. The exact quantification of the false positive rate of motif pattern matches would require a protocol that is able to determine if a match is false or true. This biological knowledge,

however, is not available for the majority of protein sequences. Several studies used statistical measures of how well a motif is defined based on the sequence pattern itself[68; 69; 71; 73]. Such measures are also incorporated into the ELM server, where these measures can serve as a warning for the user of what order of magnitude of false positives can be expected when using only the pattern to search for true motif instances. I chose a way of demonstrating the weakly defined nature of most motif patterns based on biological considerations.

For this purpose, I used the motifs collected in the ELM database. As these motifs were described mostly in eukaryotes, there should be a strong bias of real occurrences to appear in eukaryotic proteins as opposed to bacterial and archaeal proteins. In contrast to this, scanning bacterial and archaeal protein datasets (see Data and Methods) for ELM motif patterns yields hit numbers comparable to that of searches in eukaryotic proteins (see Figure 18A). These hit numbers include both real instances and false positive (random) hits. Although the ratio of true and random hits is unknown, real hits are expected to show a pronounced enrichment in eukaryotes. On the other hand, random occurrences are expected to appear with approximately the same frequency in all three kingdoms of life. The lack of difference between eukaryotes and prokaryotes in this regard is the most alarming in the case of TRG motifs, as the lack of cell compartments in prokaryotes makes such a widespread usage of target signals controlling subcellular localization very improbable.

Figure 18B shows that the normalized number of matches from the three domains of life are mostly indistinguishable even when assessed for each ligand binding (LIG) motif separately. The horizontal axis is a list of all LIG motifs and the height of the graph for a given motif shows the average number of matches per proteins in the three domains. Some well defined motifs – such as the GYF domain binding motif – have pattern descriptions that only match a handful of protein sequences (18 out of all 171,208 eukaryotic sequences from SwissProt and none of the archaeal or bacterial sequences). These motifs are grouped at the left hand side of the figure. However, there are only a handful well defined motifs, with nearly 76% of the LIG motif patterns matching at least

1 out of 100 proteins in all three domains of life. These motifs cover a wide range of functions such as the interaction with 14-3-3, WW, PDZ, PCNA domains, nuclear receptors and even the interaction with MDM2 via a motif that is experimentally described exclusively in the p53 protein family. Considering the biological meaning of these motifs, it is clear that with a few exceptions, naïve motif searches are dominated by false positives.

Ligand binding motifs mediate interaction with a well defined protein partner domain. The occurrence of three example LIG motifs are shown in Figure 18C. The top part of Figure 18C shows the occurrence of PCNA, PDZ and Cyclin binding motif hits (random+real occurrences). The position of these three motifs are shown in Figure 18B with vertical lines (note that there are three sub-types of PDZ motifs and in Figure 18C the occurrence of all three types are added). The bottom parts of Figure 18C show the occurrence of the corresponding interacting domains in the three domains of life. The occurrence of PCNA, PDZ and cyclin domains is highly unbalanced with PCNA domains being absent in bacteria, PDZ domains being absent in archaea and cyclin domains being exclusive to eukaryotes. The presence of real motifs is linked to the presence of the interacting partner domain, however, the corresponding motif hits do not reflect these specific distributions and all three motif patterns can be found ubiquitously in all three domains of life.

The same over-prediction trend can be shown for targeting signals as well. Scanning the human proteome (see Data and Methods) for TRG motifs, about 92% percent of the proteins match motifs that – in biologically active form – are exclusively found in membrane proteins (TRG_ENDOCYTIC_2, TRG_ER_diArg_1, TRG_ER_diLys_1 and TRG_LysEnd motifs). Furthermore, 41% of human proteins match classical nuclear localization signals and 33% are predicted to be localized to the peroxisome. The irrationally high numbers for these localizations and the large overlap between incompatible localizations (95% of proteins matching NLS's also match membrane localization motifs) show that targeting motifs suffer from the same under-definition as ligand binding motifs.

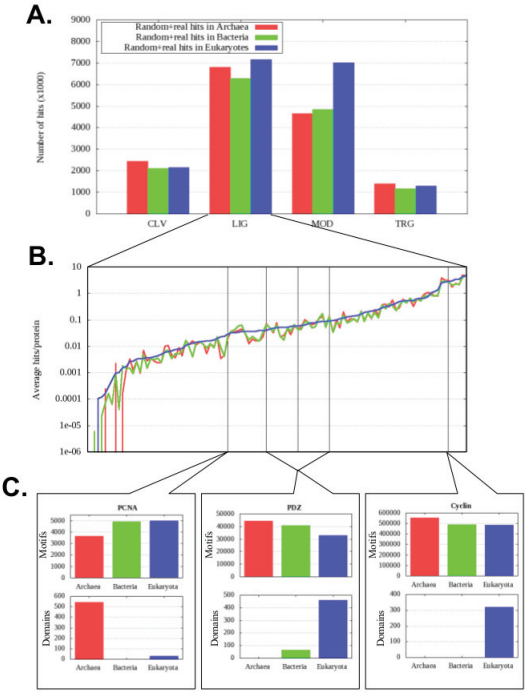

Figure 18: Results of motif scans in the three domains of life

A: the number of found motif hits in the eukaryotic (blue), bacterial (green) and archaeal (red) proteins included in the UniProt database. As the size of the three databases are different, the number of actual hits in the prokaryotic sets were scaled with the ratio of the number of residues in each dataset. B: The average number of motif hits per protein for the three databases covering the three domains of life. Again, hit numbers in prokaryotic sets are corrected for different number of residues compared to the eukaryotic dataset. Coloring is identical to that of part A (red – archaea, green – bacteria, blue – eukaryotes). C: The upper bars show the number of found hits in the three domains of life for PCNA, PDZ and Cyclin binding motifs. Lower bars show the actual number of corresponding partner domains that can serve as interaction partners for these motifs in the same datasets. Prokaryotic hit numbers are corrected for different number of proteins and the coloring scheme follows that of parts A and B.

4.5.2. Combining linear motif and disordered binding region predictions

Overall efficiency and the reduction of false positives

The overlap between predicted disordered binding regions and linear motifs was tested using ANCHOR predictions and annotated ligand binding linear motif (LIG) instances from the ELM database. For this purpose a more permissive version of ANCHOR was chosen, where the prediction threshold was reduced to 0.4 instead of the original 0.5. Motif instances were checked and filtered for similarity to minimize redundancy (see Data and Methods). The majority of annotated LIG motif instances were recognized by ANCHOR as binding regions yielding a recovery rate of 66%. In contrast, the overlap between ANCHOR predictions and unfiltered motif pattern matches in the eukaryotic sequences in UniProt (containing both random and true motif instances) is significantly lower with 17.6% (see Figure 19). In total 7,164,890 LIG motif hits were found in the total of 171,208 sequences. Upon filtering the hits with ANCHOR, only 1,262,532 LIG motif hits remained, yielding a reduction of over 82%.

The strong connection between true linear motif instances and ANCHOR predictions is supported by the fact that the disordered binding regions predicted by ANCHOR overlap with known linear motifs with a significantly higher ratio than expected from random (see Data and Methods). The fraction of linear motifs recognized by ANCHOR is very similar to the true positive rate of ANCHOR as measured on true disordered binding regions (66% versus 68%, respectively). Furthermore, ANCHOR is much more sensitive to true motif instances than for protein segments simply matching a motif pattern. This can be used to enrich the number of true positive motif hits when scanning through unknown sequences by discarding the motif hits that do not overlap with ANCHOR predictions. The results obtained with ANCHOR filtering are more reliable as correct

motifs are enriched while the total number of hits are reduced by nearly an order of magnitude.

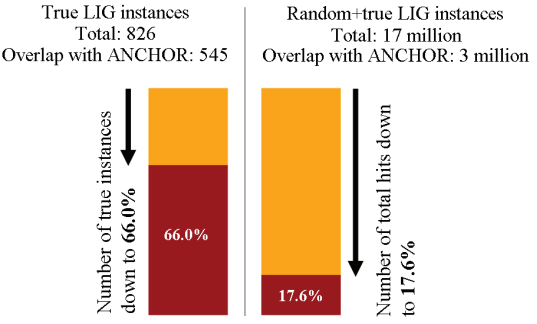

Figure 19: The predictive power of ANCHOR as a filter in motif searches

Left: fraction of known instances recognized by ANCHOR. Right: the reduction in the number of ligand binding motif matches in the eukaryotic sequences of UniProt.

ANCHOR's recovery rate and the reduction of hits, however, is highly uneven between different motifs. At one extreme, all 22 instances of the nuclear receptor box motif (LIG_NRBOX) were recognized and at the other, none of the 5 TPR binding motifs were found. To give a more detailed picture on the efficiency of ANCHOR in motif recognition, recovery rates and the reduction of hits (calculated on the eukaryotic sequences in UniProt) were calculated for each motif separately. Figure 20 shows the total number of instances and the number of these overlapping with ANCHOR predictions for all LIG motifs that had at least three independent annotated instances. For each motif the rate of recovery was compared to the random overlap between ANCHOR predictions and randomly chosen protein segments (see Data and Methods). For motifs marked with asterisk the number of overlap is significantly higher than expected from random.

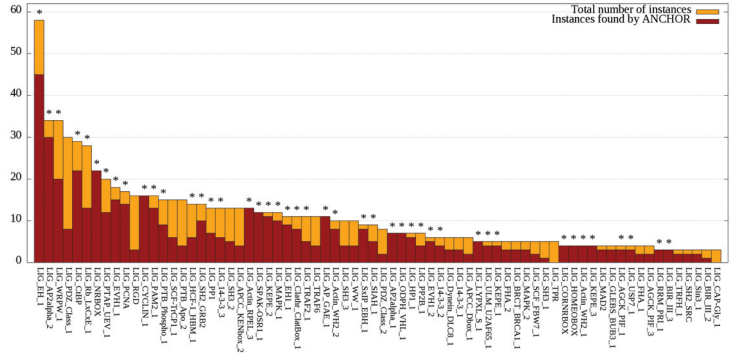

Figure 20: Efficiency of ANCHOR for individual LIG motifs

The figure shows the total number of annotated instances for each of the ligand binding motifs that have at least three independent instances in the ELM database. Dark red bars show the number of instances overlapping ANCHOR predicted binding regions. Stars mark the motifs for which the recovery rate is significantly higher than that expected by chance alone (see Data and Methods).

Considering the reduction of the total number of hits, my analyses show that for well defined motifs giving a moderate number of hits in the eukaryotic UniProt sequences ($<10^4$) the reduction rate is lower with an average of approximately 60%. However, for more ill-defined motifs ($>10^4$ hits), the reduction rate increases and reaches approximately 85%. This shows that ANCHOR can be especially useful for filtering hits of poorly defined motifs, whereas for well-defined motifs the definition already guarantees a more moderate false positive rate.

The combination of ANCHOR and linear motif prediction can yield a combined prediction tactic that is able to make use of the distinct advantages of the two methods. On one hand the use of linear motifs inherently gives information about the interacting partner. Furthermore, it is able to capture the essentiality of certain positions inside a binding region. On the other hand, the incorporation of ANCHOR makes it possible to take into account the influence of the residues surrounding the core residues of the motif. In many cases the effect of this context in motif mediated binding was shown to be

considerable[67]. In addition, ANCHOR can effectively discriminate between the presence of different structural tendencies on and around the binding region. Furthermore, as ANCHOR uses a cutoff value to give predictions, this makes the resulting, combined approach tunable and its specificity and sensitivity can be tailored to suit the need of various applications.

Efficiency by structural context

The use of ANCHOR assumes that true motif instances reside in disordered protein regions. Although this holds for most motif instances[125], some true motif instances are known to reside in accessible surface loops of globular domains. Furthermore, some motifs are generally found at terminal regions of proteins. For example, the PDZ motifs occur exclusively at the C terminus of proteins and are usually preceded by a folded domain. As ANCHOR relies heavily on the disordered state of the protein region to recognize disordered binding motifs, in these cases its efficiency is expected to be lower.

To test this, LIG motif instances were grouped according to the disorder or order of the sequence regions flanking the instance. Based on this, three groups were established. A motif instance is categorized as disordered, if both the N- and C terminal flanking regions are predicted to be disordered by IUPred. Mixed instances are flanked by a disordered region on one side and by an ordered one on the other side. Ordered instances reside in a sequential environment fully predicted to be ordered.

Figure 21 shows the efficiency of ANCHOR on all three groups. This efficiency varies heavily between the groups. Only 19.7% of ordered instances are found, but the recovery rate increases to 60.5% and 86.0% for mixed and disordered instances, respectively. These results are largely independent of the prediction method used for the assignation of disorder status, and remained consistent upon using DISOPRED2 or VSL2 (data not shown).

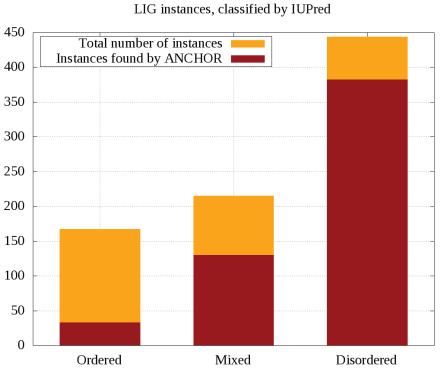

Figure 21: The ratio of motif instances annotated in the ELM database identified by ANCHOR
Instances are classified according to the predicted disorder status of their flanking sequential environment.

4.5.3. Examples

The majority of known linear motifs reside in a disordered protein region to make the interacting segment accessible for the partner molecules (see reference[125] and Figure 21). One such example is show in Figure 22A for the nuclear receptor binding motif NRBOX in the human nuclear receptor coactivator 2 protein (NCOA2). NCOA2 is a 1,464 residue long transcriptional coactivator for steroid receptors and nuclear receptors. Its dysfunction has been linked to acute myeloid leukemias. The protein contains four verified instances of the NRBOX motif through which it can bind to the human NR3C1 glucocorticoid receptor. The motifs reside in the unstructured regions of the NCOA2 protein between residues 641-882. The NRBOX motif consists of three leucine residues in an xLxxLLx configuration (see section 1.6). This hydrophobic sequence signal is readily picked up by ANCHOR and the motif regions are correctly predicted as disordered binding regions. Figure 22A also shows the known structure of one of these motif instances bound to its receptor partner.

Although in fewer numbers, there are many examples of biologically functional motif instances that are found inside structured domains. An example is shown in Figure 22B:

the MAP kinase binding motif of the human DUS6 protein. DUS6 is a 381 residue long protein implicated in various signaling pathways, including apoptosis, growth and cell speciation. It consists of two structured domains, a rhodanese and a tyrosine-protein phosphatase domain, connected by a linker region. The motif region is in a surface accessible part of the rhodenase domain and therefore can be bound by the target kinase. However, as the monomeric structure shows in Figure 22B, the motif region is structured even without the presence of the binding partner. As the identification of linear motif instances with ANCHOR relies heavily on the presence of protein disorder, these motifs cannot be identified with ANCHOR. This motif has an ordered structural context, where the performance of ANCHOR is very low (see Figure 21). The identification of motif instances similar to these calls for the application of domain and accessibility predictions.

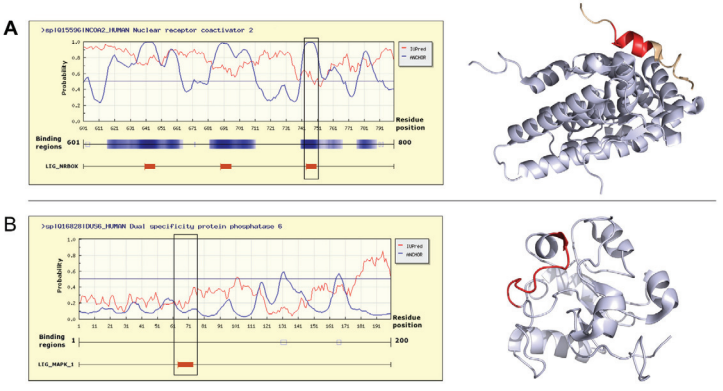

Figure 22: Examples of true motif instances with ANCHOR predictions

A: Three instances of the nuclear receptor binding motif (LIG_NRBOX) in the human nuclear receptor coactivator 2 protein (NCOA2). Left: IUPred (red) and ANCHOR (blue) predictions for the 601-800 region of NCOA2. Red bars mark the motif instances with the black box showing the instance for which the corresponding bound structure is shown. Right: the structure of NCOA2 (salmon) bound to the glucocorticoid receptor (grey) with the motif shown in red (structure 1m2z[126]). B: MAP kinase binding motif (LIG_MAPK_1) in the rhodenase domain of the human DUS6 protein. Left: IUPred (red) and ANCHOR (blue) predictions with the red bar and black box indicating the position of the motif. Right: the structure of DUS6 in monomeric form (structure 1hzm[127]) with the motif shown in red.

4.5.4. Application to whole proteome scans

To test the usability of ANCHOR in a large scale scenario, I scanned the human proteome for the nuclear receptor binding motif LIG_NRBOX and applied the ANCHOR filtering to the resulting motif hits. For NRBOX motifs the efficiency of ANCHOR is 100% on known instances with all 22 known true motifs overlapping predicted binding regions. In total 7,897 of the scanned proteins match the NRBOX motif at least once, accounting for roughly 39% of all human proteins. The number of proteins containing motif matches is reduced to 1,623 (8%) after applying ANCHOR filtering (see Figure 23A).

NRBOX motifs are annotated with Gene Ontology (GO) terms from all three existing categories (biological process, cellular component and molecular function). Proteins with both unfiltered and filtered NRBOX motif matches were grouped according to their GO annotations (see Data and Methods). In the case of all three annotation types (biological processes, cellular components and molecular functions), ANCHOR filtering increased the ratio of proteins matching the annotations of NRBOX motifs 1.4-2.3 fold (see Figure 23B-D). In all three cases, the number of proteins bearing no annotations at all was high and did not change significantly due to the filtering. This shows that the relatively low ratio of proteins with correct annotations even after filtering is a consequence of the generally poor GO annotation of human proteins. Furthermore, proteins can participate in several processes, can perform multiple functions and can have multiple localizations. As a result, the proteins with annotations not matching those of NRBOX proteins are not necessarily false positives. Due to these limitations the ratios of proteins with correct GO terms in themselves are not indicative. However, the significant enrichment of these proteins as a result of ANCHOR filtering shows that the filtering procedure greatly increases the ratio of correct motif hits while reducing the total number of hits by 80%.

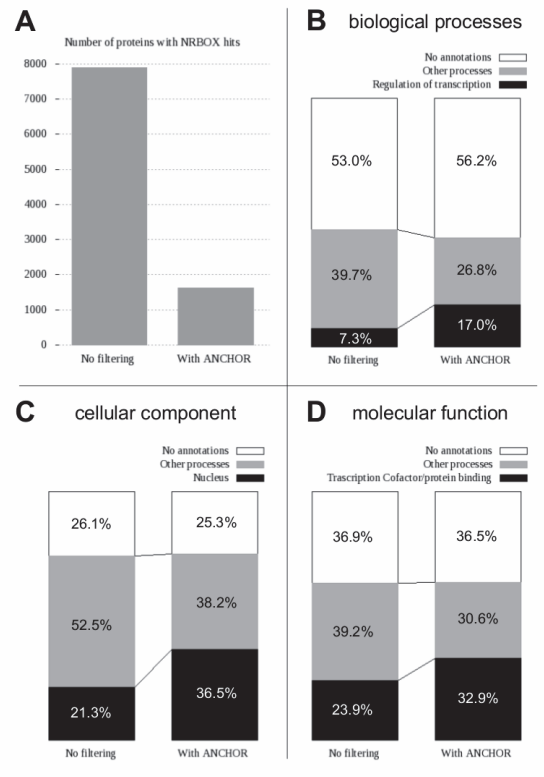

Figure 23: Application to whole proteome scans

Results of applying ANCHOR as a filter for scanning the human proteome for instances of the nuclear receptor interacting motif (LIG_NRBOX). A: number of proteins matching the motif; B-D: fraction of proteins containing NRBOX matches with biological process, cellular component and molecular function GO annotations (B, C and D, respectively) matching the annotations of true NRBOX instances (black boxes), with other annotations (grey boxes), and no annotations (white boxes). The height of bars in B-D represent 100% of all found motifs and thus in each sub-figure the complete left bar stands for 7,897 proteins and the complete bar on the right stands for 1,623. The two different number of hits are scaled to accurately represent enrichments of correctly annotated proteins.

4.5.5. Implications

Altogether, the presented results support the complementarity of the linear motif and disordered binding region concepts. This can serve as a stepping stone for creating new models of molecular recognition that take into account the relevant features of both current approaches, thus the integration of the two concepts can provide a deeper and a more complete picture of the molecular details of protein-protein interactions. However, apart from the theoretical message of these results, the presented results have strong practical implications as well. In general, the combination of the two predictions reflecting the two binding models enables us to get the best of both worlds: predict interactions with relatively low false positive rate, with structural context and with information about the partner. This can aid the prioritization of candidate motifs for experimental works and improve the quality of proteome-wide systems biology analyses. Furthermore, unlike many filters commonly applied in motif hit filtering[77], the efficiency of ANCHOR can be quantified separately for different bound secondary structures, structural context of the binding site and even for individual motif types. Based on this, researchers can decide before commencing a study whether ANCHOR results should be incorporated in their protocol. As interactions mediated by both disordered proteins and linear motifs were mostly described in regulatory proteins, the more precise prediction of binding regions has high importance in assembling protein-protein interaction networks of various organisms. This can aid the understanding of the intricate interplay of proteins communicating through transient interactions. The pinpointing of proteins and the exact protein regions that are involved in these regulatory pathways in turn can be used as a starting point in pharmaceutical studies aimed at drug target identification.

4.6. Towards a unified view of protein structure and interactions – limitations and possibilities

Although many parameters of biological systems are continuous or quasi-continuous, there is a universal trend of categorizing in sciences in general and especially in biology. A prime example is the interaction of molecules. Albeit affinities of interactions are continuous and can range several orders of magnitude, in many cases the results of affinity measurements are condensed into simple binary statements of whether the two molecules interact or not. These simplified statements are at the heart of the studies of signaling pathways, interaction networks of organisms and pharmacological studies. The rationale behind this simplification tendency is that condensing a huge amount of information to a level perceptible to the human brain, researchers can deduce further results much more easily. "Not getting lost in the details" has its clear advantage, but after a certain sophistication of the field, this approach can present new burdens.

Similarly to almost all fields of molecular biology, the introduction of distinct categories has been heavily applied in the field of protein disorder as well. After the realization of the fact that a well defined 3D structure is not a prerequisite of protein function, a new category of "disordered proteins" was established to describe such proteins. This framework is currently widely used when studying a protein structure. Protein disorder is routinely inferred for example from X-ray structure determination: a position is either visible in the structure (ordered) or not (disordered). Results from other experimental measurements (such as CD, NMR, SAXS, etc.) are also represented in a binary form, although all of these measurements provide continuous output values. Consequently, this binary representation is present in databases as well (eg. in the DisProt database) and has also percolated to the bioinformatics tools targeting protein disorder. Accordingly, almost all disorder prediction methods mark each residue in the input protein as either ordered or disordered. Although most prediction methods assign a continuous score to residues as well, this score is not optimized to reflect any biologically relevant feature, but only reflects some internal score. During the testing of the

algorithms, an optimal cutoff is set regarding this score that best separates disordered and ordered residues. Then the assigned score of a residue can be converted to a probability value showing the reliability of the prediction (for a comprehensive review see [59]).

In reality, from a structural point of view, disordered segments are heterogeneous and affect various levels of protein structure[29; 59]. Some of them exist in the form of (near) random-coils that corresponds to a largely random distribution of conformations dominated by extended structures. In reality, however, no protein is ever random coil, and the macroscopic properties compatible with random coils do not exclude the possibility of transient short-range or long-range interactions resulting in transient structural elements. Indeed, transient secondary structure elements were observed in a number of cases. Disordered proteins can also exist as molten-globules and exhibit a compact but disordered state with some secondary structure content. Generally, various types of disorder and the transition between these states can be linked to specific function of the proteins.

As protein flexibility is inherently not discreet, the strict binary categorization of residues of a protein into "ordered" and "disordered" groups is a great oversimplification. Disorder is a complex phenomenon, and there are many examples that go beyond the classical ordered/disordered classes. In these cases, there is no single good answer from the perspective of predictions. The inability of prediction methods to handle various types of protein disorder causes a serious limitation in their efficiency. I illustrate this problem through the example of human calpastatin that contains multiple disordered binding regions. Although calpastatin does not have a stable three dimensional structure on its own, the binding regions exhibit strong structural preferences. This places them at the borderline of order and disorder in various aspects. The comparison of the behavior of several disordered prediction methods can provide insights into their general features and usability.

Calpastatin is a 708 residue long protein that is a specific inhibitor of calpain, a Ca^{2+} activated cystein protease. The calpain-calpastatin interaction is part of multiple larger

networks of interactions involved in the regulation of cell division, cell motility and muscle protein degradation. Calpastatin contains four repeats of the calpain inhibitory domain and thus is able to inhibit four different calpain molecules at the same time. Each inhibitory domain binds to calpain via three separate binding sites (A, B and C). The center binding site B binds to the active site of calpain in an extended conformation, while the other two sites A and C bind as α-helices and increase the specificity of the interaction between the two molecules. Although calpastatin is fully disordered along its entire length, the binding sites exhibit considerable transient structure in isolated form as well[102]. These transient, preformed structural elements correspond to the secondary structure that these segments adopt upon binding to calpain, namely α-helical structure for sites A and C but site B also has highly nonrandom conformational preferences.

Figure 24 shows prediction profiles from 8 different disorder prediction algorithms covering the most commonly used prediction algorithm architectures (see Section 1.4) for the first inhibitory domain of calpastatin (residues 137-277). The output of each method is a continuous score in the [0;1] interval assigned to each residue in the sequence. This score shows the probability of each residue to be disordered (for a more detailed example of a disorder prediction profile given by IUPred, see section 1.4.3 and Figure 3). All methods are trained according to the aforementioned binary approach, where residues are categorized as either ordered or disordered. As a result, all of these methods are optimized for this binary classification and traditionally, prediction outputs are condensed to a binary output as well: if a residue is assigned a score below 0.5, it is considered ordered, and scores above 0.5 indicate disorder.

The presence of the preformed structure of the binding regions is reflected in almost all of the prediction outputs as they generally assign a lower score to the binding sites than to the rest of the protein. Although the dips apparent near the three binding sites are relatively consistent among different methods, they react to these segments in a variety of ways. Some predictors only react to the general structural content of the inhibitory domain as a whole and give a slight dip in the middle of the domain coinciding with binding site B (VSL2B and POODLE-I), while some others give three distinct dips

approximately corresponding to the three separate binding regions (DISOPRED2, IUPred, OnD-CRF, RONN and DISpro). VL-XT also reacts to the presence of residual structure, albeit in a relatively erratic fashion. The average score on linker regions between binding sites is generally larger than on the binding regions themselves, reflecting the fact that these regions retain their disordered nature even in the bound form. On the other hand, the large variation in the prediction scores on the binding regions shows that at these regions, conclusions drawn from a single predictor or a naïve consensus prediction is either meaningless or can be very misleading.

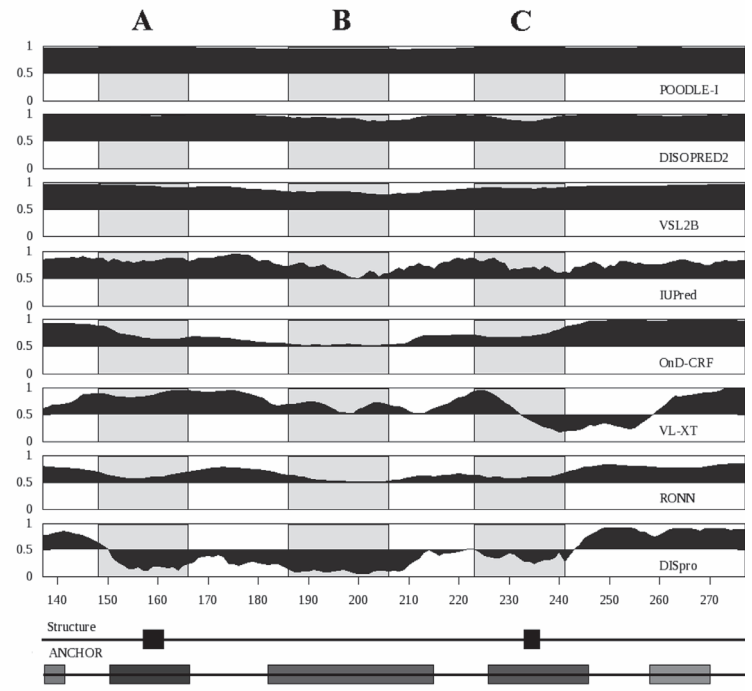

Figure 24: Disorder predictions for the first inhibitory domain of human calpastatin (UniProt AC: P20810).

In the case of OnD-CRF and DISOPRED2 the original prediction scores were rescaled linearly to be directly comparable with other methods. Disordered predictions were sorted top to bottom by decreasing average predicted disorder tendency calculated on the shown sequence part. Grey boxes labeled A, B and C on the prediction outputs mark the three binding regions. Underneath the prediction outputs, the sequence parts that were shown experimentally to adopt α-helical structure when bound to calpain (based on the PDB entry 3df0) are shown ("Structure"). The bottom line shows the disordered binding site prediction by ANCHOR. Shading of the boxes corresponds to the overall confidence of the predicted binding region, with darker shades corresponding to higher confidence.

Similarly to disordered binding regions, other "flavors" of disorder exist as well[128], such as coiled-coil or molten globule regions. These structural elements challenge the classical definition of protein disorder, as for example coiled coils always occur as oligomers (formed by 2-7 monomer proteins). Each protein adopts an α-helical conformation that is unstable on its own, however, the helices are stabilized by the interaction with other helices. The resulting structure is stable and lends itself to structure determination. This structural element is on the verge of order/disorder, as the constituent monomers do not have a stable structure, but their obligate complex does. Such structural elements pose a challenge similar to that of disordered binding regions to prediction algorithms. The common feature in these problematic structural regions is their intermediate flexibility. This suggests that approaches that go beyond the binary classification of proteins as ordered or disordered are necessary. Although it can be tempting to identify the continuous score provided by prediction methods as a measure of flexibility, no such information is used in the construction of these algorithms and the final score is not optimized for this. The lack of consensus, or even similarity between different methods, as illustrated by calpastatin underlines the inability of most current methods to directly capture flexibility. Furthermore, the proper identification of proteins and protein regions with transient/intermediary structural content is not simply a structural problem, but also a biological one, as the specific functional modes of disordered protein regions are directly linked to their intrinsic flexibility. This identification problem can be tackled with the use of specific prediction methods, such as ANCHOR for the identification of disordered binding regions or COILS for the recognition of coiled coil regions.

Although structural heterogeneity restricts the brute-force application of prediction methods for regions containing certain structural elements, the basis of physics based disorder prediction methods can be a starting ground to get more information about the presence and type of transient structure. During the development of IUPred[60] and ANCHOR[33], it became clear that the phenomenon of the lack of structure can be understood and modeled on the basis of the energy of interresidue interactions. Using this concept, not only disordered segments, but regions undergoing disorder-to-order

transition could be recognized as well. This suggests that models incorporating basic biophysical properties of disordered segments hold the key to more detailed predictions of protein disorder. Although these methods currently do not outperform advanced machine learning methods, they are rooted in a strong biophysical model that – as opposed to machine learning approaches – can be improved and fine-tuned.

The development of ANCHOR on the basis of IUPred demonstrates that in order to describe protein disorder beyond a binary classification, the existing models have to be elaborated. The common physical description of protein structure and various 'flavors' of protein disorder based on the energy landscape model can guide the elaboration of our models. Both the folding and the binding of both ordered and disordered proteins can be described on a common ground, as shown in the Introduction (sections 1.1.2-1.1.3 and 1.2.2-1.2.4). Conformational heterogeneity naturally follows from the energy landscape view. The funnel-like energy landscapes of strictly ordered proteins and the plateau-like energy function of highly disordered proteins represent two extreme scenarios, as shown in Figure 25. In reality, every protein can adopt a vast number of different conformations and each of them can occur with non-zero probability. However, conformations with lower energy are more probable, while higher energy conformations are present less frequently. Therefore, every protein is inherently dynamic, although the details of dynamic behavior differ from one protein to another. In the case of globular proteins, the ensemble is dominated by a single narrow range of conformations that have significantly lower energy compared to other conformations. This leads to the presence of a well-defined structure. There could be other low-energy conformations even in the case of ordered proteins, represented by valleys on the figure. These alternative conformations, that can have important functional roles, are becoming more commonly detected as the resolution of experimental techniques improves. In contrast to globular proteins, the energy surface of IDPs has multiple local minima that are energetically near-identical. These proteins virtually exhibit a continuum of allowed conformations. The significant differences in the free state of various proteins can also have a large impact on the way these proteins interact.

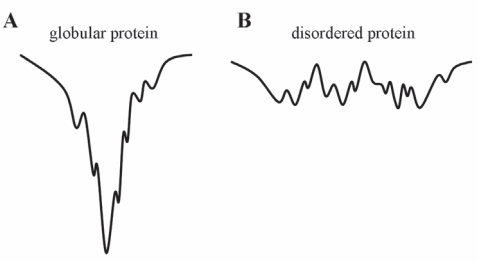

Figure 25: Extreme examples of energy landscapes

Schematic representation of the energy landscape of a globular protein (A) and a disordered protein (B). The energy of the system is sketched against a single coordinate of the conformational space.

The various scenarios of the binding of proteins can be treated analogously to their folding using the energy landscape view. Apart from describing the various "classical" binding modes of globular proteins (introduced in sections 1.2.3 and 1.2.4), energy landscapes offer a way to integrate the binding of disordered proteins as well[129]. The energy landscape of the complex is created from the combination of the conformational space of the interacting molecules. However, the interaction with the partner molecule can induce drastic changes in the shape of the energy landscape corresponding to the individual protein. Disordered segments that adopt a single well-defined structure as a result of the complex formation are expected to have a funnel-like energy landscape with a single well-defined minimum. However, disordered proteins sample a large number of different conformations in their initial state prior to the interaction. Thus, the final conformation is chosen from a conformational ensemble instead of a limited number of conformations. This process can be viewed as a continuous version of the classical conformational selection model. Figure 26 shows a schematic representation of the binding mode of IDPs. This type of binding is not compatible with either the lock-and-key or with the induced fit model. However, it was suggested that as the conformational space narrows down during the formation of the complex, these mechanisms might come into play. Interestingly, during binding the disordered segments do not always become fully ordered but can retain their dynamic nature even as part of the complex, resulting in a 'fuzzy complex'[40]. Such a dynamic complex is formed, for example, between the

intrinsically disordered Sic1 with its partner Cdc4 during regulation of yeast cell cycle progression. These complexes can be described only by a set of alternating conformations. The resulting energy profile is also shown on Figure 26.

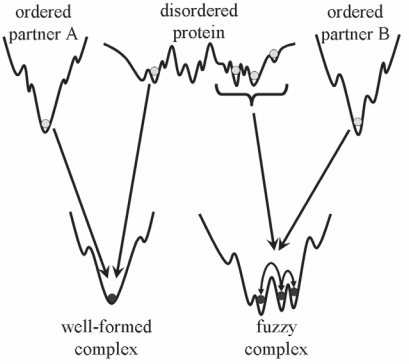

Figure 26: The energy landscape of the interaction between a globular and a disordered protein
The energy of the system is sketched against a single coordinate of the conformational space. The initial and final states of proteins are represented by light and dark dots, respectively. The globular protein has a funnel-like-, while the disordered proteins have a flat and highly rugged landscape. The resulting complex can become completely ordered, represented by a funnel, or can retain some flexibility, resulting in a fuzzy complex.

These examples show that although the various levels of flexibility present in proteins and their complexes allow them to carry out their functions in different ways, the kaleidoscope of protein interactions, however, is built upon the same physical principles. The current energy estimation scheme behind IUPred and ANCHOR (introduced in section 1.4.3) is able to give an estimation of the depth of the minimum of the energy function of a given protein or a protein segment. In order to incorporate the estimation of local flexibility/order of a protein, the width of these valleys and the presence of other, energetically near-identical conformations have to be described as well. This effectively means the estimation of entropic terms of a protein chain that would open the way for describing the presence of local structure of disordered proteins via short range interactions. Understanding these principles and modeling the formation of protein

structure based on them will help to develop better prediction methods, more well-designed experiments, and novel approaches to aid drug development.

5. Conclusions and future directions

The past two decades saw the rapid accumulation of molecular biology data made possible by the development of high-throughput experiments and rapid sequencing techniques. At the same time, our theoretical knowledge and interpretation of experimental results are lagging behind the amount of data at hand. The fact that we have immensely more data than we can make sense of, necessitates the development of bioinformatics methods through which the analysis and processing of available data can be achieved in a reliable and cost efficient way. This need for theoretical advances and their practical applications is even more pronounced in the field of disordered proteins. In the past twenty years after their recognition, it became clear that these proteins and their interactions play a fundamental role in the regulation and signaling of living cells. However, both their theory and the available practical methods aimed at analyzing them are not in proportion with their importance.

Disordered proteins and their binding are governed by the same physical principles as the folding and interactions of ordered/globular proteins. Through this, the thermodynamics description of globular proteins can serve as a starting point in the modeling of disordered proteins. The implementation of physical modeling enabled the development of the successful and novel prediction of protein disorder from the protein sequence alone. As opposed to various machine learning methods that do not have a physical background, IUPred uses statistical potentials to estimate the potential interaction energy a protein chain can form on its own via intra-molecular interactions. Although the applied model is coarse grained without considering atomic details, it correctly describes the main driving force behind protein structure formation and thus is applicable to modeling both ordered and disordered proteins.

Based on the validity and success of the residue-residue interaction energy prediction scheme implemented in IUPred, it was possible to extend this model to the interaction of disordered proteins (see section 4.1). Specifically, in my work I aimed at developing a prediction method that is able to recognize protein regions from the sequence that are

disordered in isolation but can adopt a well defined structure when binding to an ordered protein partner. Such disordered binding regions, compatible with coupled folding and binding, have to fulfill distinct energetic requirements that can be quantified with the energy prediction scheme. The possible interaction energy a residue can gain by interacting with a globular protein partner (inter-molecular interactions) can be modeled in the same fashion as intra-molecular interactions. Using this approach I developed ANCHOR, a method that is able to recognize around 70% of known disordered binding sites correctly from the sequence alone. Although ANCHOR was optimized on short disordered binding sites, it can correctly identify long segmented binding regions as well. Furthermore, the efficiency is largely independent of the amino acid composition or the type of bound structures of the binding sites. This generality on one hand has theoretical implications. The applicability of ANCHOR to different types of binding regions supports the generality of the underlying model. On the other hand, from a practical point of view, ANCHOR can be applied to unknown sequences without restrictions. This method was the first (and up to date remains the only) general, publicly available such method. It is accessible via its own dedicated web server and can be downloaded for local use as well.

Apart form the analysis of single proteins, ANCHOR can be applied – in conjunction with other prediction algorithms – to gain system- or evolutionary level conclusions. Using IUPred for the prediction of protein disorder and ANCHOR for the prediction of disordered binding regions I was able to demonstrate that the presence of both disorder and disordered binding sites increase with the complexity of the studied organism (see section 4.2). In general, eukaryotic proteomes contain a larger fraction of these structural elements than bacteria and archaea, furthermore, in complex organisms the typical length of disordered and disordered binding regions are significantly longer. My results imply that throughout the course of evolution, protein disorder serves as an advantage and new disordered regions are introduced to harbor binding regions. This mechanism can support the emergence of complex signaling pathways and regulatory networks.

The proteome-wide analysis of disorder and disordered binding regions provided interesting example organisms that seem to deviate from the general trends. One such organism is *Mycobacterium tuberculosis* (MTB), the main causative agent of TB. This remarkably successful obligate intracellular organism is predicted to contain an unusually high fraction of disordered proteins. Via thorough sequence analysis employed on the MTB proteome using domain analysis and sequence profiling, I was able to pinpoint two protein families that can play a major role in the successful adaptation of MTB (see section 4.3). The representative proteins of these families are generally long, modular and contain large disordered regions. Although developed on the MTB proteome, the proposed protocol is independent of the organism and can be used generally on any organism of interest, thus aiding drug target searches in identifying promising drug target proteins.

I carried out another large scale sequence analysis focusing on point mutations connected to human cancer (see section 4.4). It was shown by previous studies that proteins involved in cancer exhibit a high disorder content. This induced popular claims that 'disorder entails a biological cost', arguing that disorder makes proteins more vulnerable to mutations. Through analysis of the distribution of cancer-associated mutations across various structural regions of proteins I was able to show that taking the appropriate background distributions into consideration, disordered regions in fact are depleted in cancer-associated mutations. Through functional analysis using various statistical measures, I also demonstrated that the association between protein disorder and the involvement in cancer is indirect and can be explained through the function of proteins.

Parallel to the disordered binding region concept, interaction between short regions of proteins and globular domains has been extensively studied using the concept of linear motifs. In this framework, the description of the molecular recognition is based on sequential properties instead of structural ones. The interaction between certain globular domains and their binding regions has been shown to be mediated by a limited number of residues in the short interacting partner. These residues form the motif which is supposed

to mediate the binding largely independent of the rest of the protein chain. Although these motifs are known to generally reside in disordered protein regions, the connection between the 'disordered binding region' and 'linear motif' models present a largely uncharted territory. I studied this connection by representing each concept with its dedicated prediction method (see section 4.5). I used ANCHOR for the prediction of disordered binding regions and the regular expression representation of ligand binding motifs from the ELM database. Using the annotated examples of known motifs it was possible to show that there is a significant correlation between binding sites predicted by ANCHOR and the occurrence of true motif instances. The resulting combined method presents the best of both worlds: the motifs can take into account the essentiality of certain key residues indispensable for the interaction and provide information about the interacting partner. Furthermore, the incorporation of ANCHOR dramatically reduces the number of false positives – the main limiting factor in the naïve use of linear motifs for prediction. In addition, ANCHOR also introduces a way to take into account the effect of the 'context' residues that flank the core motif residues – an effect that is largely overlooked, albeit recent studies have estimated it to be more significant than previously anticipated. I demonstrated the efficiency of using ANCHOR as a filtering procedure for linear motif searches through the large scale scanning of the human proteome for nuclear receptor binding motifs.

In the commonly used description of disordered proteins, disorder is considered as a binary feature and based on experimental results, proteins or protein regions are categorized as either ordered or disordered (see section 4.6). Accordingly, current disordered prediction methods are used as binary predictors to reproduce the same binary classification using a bioinformatics approach. In reality, however, protein flexibility is a continuous property and as a result protein disorder is heterogeneous ranging from the rigid structure of trypsin to the near random-coil behavior of ACTR. This heterogeneity can be explained in the common thermodynamical description of proteins. Although there exist dedicated prediction methods for the identification of 'structurally ambiguous' regions (such as COILS for the prediction of coiled coil regions), these methods are generally not based on biophysical considerations. In this light, one of the most important

theoretical message of the success of IUPred and ANCHOR is that the common physical description of both disordered and ordered proteins can be modeled in a unified framework. By further refining the underlying model, this approach can be developed to accurately describe the alternative conformations of proteins based on their sequence. By modeling short range interactions and effectively estimating the entropic terms, this approach can serve as a basis of developing more sophisticated prediction algorithms and the deeper understanding of the continuous spectrum of protein disorder. As the function and the mode of interaction of proteins are intimately linked to their flexibility, these results will deepen our understanding of the molecular recognition of disordered proteins.

Summary

In the last decade of the 20th century the results of high-throughput genomic studies drastically changed our view of structures and biological roles of proteins. Until the early 1990's the basic assumption of structural biology was that the structure of a protein is indispensable to its proper function. However, the accumulation of known proteins contradicting this 'structure-function paradigm' forced molecular biologists to reassess this prevailing view. These intrinsically disordered/unstructured proteins (IDPs/IUPs) do not have a stable, three dimensional structure in isolation, even under physiological conditions, yet they are able to perform highly specific and crucial functions in signaling, transcription and various regulatory processes, such as the control of cell division and apoptosis. Given the functional importance of IUPs, many proteins containing disordered regions have been associated with various diseases, such as cancer, diabetes, amyloidosis and neurodegenerative diseases.

Although lacking a stable structure in isolation, IUPs can adopt a well-defined conformation upon interaction with partner molecules. This coupled-folding-and-binding process distinguishes the binding of disordered proteins from that of globular proteins. The energetics of this special binding mode can be modeled via estimating the residue-residue interactions using statistical potentials. These potentials, derived from globular protein structures using the Boltzmann hypothesis are at the heart of IUPred, a protein disorder prediction algorithm. By extending the core model of IUPred, I was able to develop ANCHOR, a method designed to identify 'disordered binding regions' – protein regions that are disordered in isolation but are able to bind to globular partner proteins via coupled-folding-and-binding. The method only needs the sequence of a protein as an input and hence is applicable to all proteins with known sequences. ANCHOR is publicly available to academic users at the http://anchor.enzim.hu web-server.

The development of ANCHOR opened up a novel, fast and cost-efficient way to analyze individual proteins, as well as to conduct large scale bioinformatics studies. I applied ANCHOR to the analysis of whole proteomes to gain insights about disordered binding regions at an evolutionary level. In combination with other bioinformatics tools, I developed a novel protocol for the identification of potential drug target proteins and tested the protocol using *Mycobacterium tuberculosis* as a model organism. I also applied ANCHOR to address the connection between protein disorder and disordered binding sites and cancer-associated mutations. It could be demonstrated that although many disordered proteins can be linked to cancer – contrary to widespread claims – protein disorder in itself does not entail a biological cost, at least in terms of single amino acid mutations. I also demonstrated that protein disorder and involvement in cancer do not share a causative relationship, but are linked by protein function. ANCHOR also enabled the partial integration of two distinct models of molecular recognition: the 'disordered binding region' and the 'linear motif' concepts. The benefit of these results is twofold: from a theoretical point of view they deepen our understanding of the molecular recognition of disordered proteins and from a practical point of view they can serve as readily applicable tools for planning experiments and interpreting results ranging from basic science to pharmaceutical drug design.

References

1. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**, 899-907.
2. Baldwin, R. L. (1994). Protein folding. Matching speed and stability. *Nature* **369**, 183-4.
3. Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501-9.
4. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-55.
5. Dill, K. A. (1999). Polymer principles and protein folding. *Protein Sci* **8**, 1166-80.
6. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167-95.
7. Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol* **4**, 10-9.
8. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci U S A* **89**, 8721-5.
9. Tsai, C. J., Kumar, S., Ma, B. & Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Sci* **8**, 1181-90.
10. Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Fold Des* **2**, S69-75.
11. Blow, N. (2009). Systems biology: Untangling the protein web. *Nature* **460**, 415-8.
12. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20.
13. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform* **10**, 217-32.
14. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**, 2177-98.
15. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* **336**, 943-55.
16. Meszaros, B., Tompa, P., Simon, I. & Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol* **372**, 549-61.
17. Valdar, W. S. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108-24.
18. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* **307**, 447-63.

19. Bonvin, A. M. (2006). Flexible protein-protein docking. *Curr Opin Struct Biol* **16**, 194-200.
20. Ofran, Y. & Rost, B. (2007). ISIS: interaction sites identified from sequence. *Bioinformatics* **23**, e13-6.
21. Fischer, E. (1894). Einfluss der Konfiguration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **27**, 2985-93.
22. Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **44**, 98-104.
23. Teilum, K., Olsen, J. G. & Kragelund, B. B. (2009). Functional aspects of protein flexibility. *Cell Mol Life Sci* **66**, 2231-47.
24. Boehr, D. D., Nussinov, R. & Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **5**, 789-96.
25. Ma, B., Kumar, S., Tsai, C. J. & Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng* **12**, 713-20.
26. Papoian, G. A. & Wolynes, P. G. (2003). The physics and bioinformatics of binding and folding-an energy landscape perspective. *Biopolymers* **68**, 333-49.
27. Csermely, P., Palotai, R. & Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci* **35**, 539-46.
28. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**, 321-31.
29. 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. (2001). Intrinsically disordered protein. *J Mol Graph Model* **19**, 26-59.
30. Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**, 197-208.
31. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527-33.
32. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**, 161-71.
33. Meszaros, B., Simon, I. & Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol* **5**, e1000376.
34. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635-45.
35. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N. & Obradovic, Z. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* **6**, 1882-98.
36. Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**, 3346-54.

37. Galea, C. A., Wang, Y., Sivakolundu, S. G. & Kriwacki, R. W. (2008). Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* **47**, 7598-609.
38. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739-56.
39. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* **12**, 54-60.
40. Tompa, P. & Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* **33**, 2-8.
41. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* **18**, 343-84.
42. Dosztanyi, Z., Chen, J., Dunker, A. K., Simon, I. & Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* **5**, 2985-95.
43. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**, 215-46.
44. Cheng, Y., LeGall, T., Oldfield, C. J., Dunker, A. K. & Uversky, V. N. (2006). Abundance of intrinsic disorder in protein associated with cardiovascular disease. *Biochemistry* **45**, 10448-60.
45. Uversky, V. N. (2009). Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci* **14**, 5188-238.
46. Uversky, V. N., Oldfield, C. J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L. M., Obradovic, Z. & Dunker, A. K. (2009). Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* **10 Suppl 1**, S7.
47. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**, 573-84.
48. Hegyi, H., Buday, L. & Tompa, P. (2009). Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol* **5**, e1000552.
49. Vavouri, T., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**, 198-208.
50. Cheng, Y., LeGall, T., Oldfield, C. J., Mueller, J. P., Van, Y. Y., Romero, P., Cortese, M. S., Uversky, V. N. & Dunker, A. K. (2006). Rational drug design via intrinsically disordered protein. *Trends Biotechnol* **24**, 435-42.
51. Metallo, S. J. (2010). Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**, 481-8.
52. 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. (2005). DisProt: a database of protein disorder. *Bioinformatics* **21**, 137-40.
53. Garner, E., Cannon, P., Romero, P., Obradovic, Z. & Dunker, A. K. (1998). Predicting Disordered Regions from Amino Acid Sequence: Common Themes

Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform* **9**, 201-213.
54. Radivojac, P., Obradovic, Z., Smith, D. K., Zhu, G., Vucetic, S., Brown, C. J., Lawson, J. D. & Dunker, A. K. (2004). Protein flexibility and intrinsic disorder. *Protein Sci* **13**, 71-80.
55. Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform* **10**, 30-40.
56. Xie, Q., Arnold, G. E., Romero, P., Obradovic, Z., Garner, E. & Dunker, A. K. (1998). The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. *Genome Inform Ser Workshop Genome Inform* **9**, 193-200.
57. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415-27.
58. Campen, A., Williams, R. M., Brown, C. J., Meng, J., Uversky, V. N. & Dunker, A. K. (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett* **15**, 956-63.
59. Dosztanyi, Z., Meszaros, B. & Simon, I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform* **11**, 225-43.
60. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827-39.
61. Thomas, P. D. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A* **93**, 11628-33.
62. Shortle, D. (2003). Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci* **12**, 1298-302.
63. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-4.
64. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N. P., Trave, G. & Gibson, T. J. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci* **13**, 6580-603.
65. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**, 265-74.
66. Neduva, V. & Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Lett* **579**, 3342-5.
67. Stein, A. & Aloy, P. (2008). Contextual specificity in peptide-mediated protein interactions. *PLoS One* **3**, e2524.
68. Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jodicke, L., Dammert, M. A., Schroeter, C., Hammer, M., Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., Luck, K., Via, A., Chatr-Aryamontri, A., Haslam, N., Grebnev, G., Edwards, R. J., Steinmetz, M. O., Meiselbach, H., Diella, F. & Gibson, T. J.

(2012). ELM--the database of eukaryotic linear motifs. *Nucleic Acids Res* **40**, D242-51.

69. Davey, N. E., Edwards, R. J. & Shields, D. C. (2010). Estimation and efficient computation of the true probability of recurrence of short linear protein sequence motifs in unrelated proteins. *BMC Bioinformatics* **11**, 14.
70. Rigoutsos, I. & Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**, 55-67.
71. Edwards, R. J., Davey, N. E. & Shields, D. C. (2007). SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* **2**, e967.
72. Marschall, T. & Rahmann, S. (2009). Efficient exact motif discovery. *Bioinformatics* **25**, i356-64.
73. Frith, M. C., Saunders, N. F., Kobe, B. & Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* **4**, e1000071.
74. Neduva, V. & Russell, R. B. (2006). DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res* **34**, W350-5.
75. Davey, N. E., Haslam, N. J., Shields, D. C. & Edwards, R. J. (2011). SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res* **39**, W56-60.
76. Chica, C., Labarga, A., Gould, C. M., Lopez, R. & Gibson, T. J. (2008). A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* **9**, 229.
77. Via, A., Gould, C. M., Gemund, C., Gibson, T. J. & Helmer-Citterich, M. (2009). A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics* **10**, 351.
78. Rajasekaran, S., Balla, S., Gradie, P., Gryk, M. R., Kadaveru, K., Kundeti, V., Maciejewski, M. W., Mi, T., Rubino, N., Vyas, J. & Schiller, M. R. (2009). Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res* **37**, D185-90.
79. Dinkel, H. & Sticht, H. (2007). A computational strategy for the prediction of functional linear peptide motifs in proteins. *Bioinformatics* **23**, 3297-303.
80. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res* **36**, D281-8.
81. Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. & Futreal, P. A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-50.
82. Sanborn, J. Z., Benz, S. C., Craft, B., Szeto, C., Kober, K. M., Meyer, L., Vaske, C. J., Goldman, M., Smith, K. E., Kuhn, R. M., Karolchik, D., Kent, W. J., Stuart, J. M., Haussler, D. & Zhu, J. (2011). The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Res* **39**, D951-9.
83. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E.,

Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

84. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. & Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res* **38**, D525-31.
85. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.
86. Receveur-Brechot, V., Bourhis, J. M., Uversky, V. N., Canard, B. & Longhi, S. (2006). Assessing protein disorder and induced folding. *Proteins* **62**, 24-45.
87. Gunasekaran, K., Tsai, C. J. & Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* **341**, 1327-41.
88. Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J. & Pavletich, N. P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* **382**, 325-31.
89. Sorenson, M. K., Ray, S. S. & Darst, S. A. (2004). Crystal structure of the flagellar sigma/anti-sigma complex sigma(28)/FlgM reveals an intact sigma factor in an inactive conformation. *Mol Cell* **14**, 127-38.
90. Hertzog, M., van Heijenoort, C., Didry, D., Gaudier, M., Coutant, J., Gigant, B., Didelot, G., Preat, T., Knossow, M., Guittet, E. & Carlier, M. F. (2004). The beta-thymosin/WH2 domain; structural basis for the switch from inhibition to promotion of actin assembly. *Cell* **117**, 611-23.
91. Huber, A. H. & Weis, W. I. (2001). The structure of the beta-catenin/E-cadherin complex and the molecular basis of diverse ligand recognition by beta-catenin. *Cell* **105**, 391-402.
92. Chumakov, P. M. (2007). Versatile functions of p53 protein in multicellular organisms. *Biochemistry (Mosc)* **72**, 1399-421.
93. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J. & Pavletich, N. P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science* **274**, 948-53.
94. Bochkareva, E., Kaustov, L., Ayed, A., Yi, G. S., Lu, Y., Pineda-Lucena, A., Liao, J. C., Okorokov, A. L., Milner, J., Arrowsmith, C. H. & Bochkarev, A. (2005). Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein A. *Proc Natl Acad Sci U S A* **102**, 15412-7.
95. Di Lello, P., Jenkins, L. M., Jones, T. N., Nguyen, B. D., Hara, T., Yamaguchi, H., Dikeakos, J. D., Appella, E., Legault, P. & Omichinski, J. G. (2006). Structure of the Tfb1/p53 complex: Insights into the interaction between the p62/Tfb1 subunit of TFIIH and the activation domain of p53. *Mol Cell* **22**, 731-40.
96. Lacy, E. R., Filippov, I., Lewis, W. S., Otieno, S., Xiao, L., Weiss, S., Hengst, L. & Kriwacki, R. W. (2004). p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. *Nat Struct Mol Biol* **11**, 358-64.

97. Cheng, Y., Oldfield, C. J., Meng, J., Romero, P., Uversky, V. N. & Dunker, A. K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* **46**, 13468-77.
98. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. (2004). Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **338**, 1015-26.
99. Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N. & Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**, S1.
100. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N. & Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry* **44**, 12454-70.
101. Galea, C. A., Nourse, A., Wang, Y., Sivakolundu, S. G., Heller, W. T. & Kriwacki, R. W. (2008). Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. *J Mol Biol* **376**, 827-38.
102. Kiss, R., Kovacs, D., Tompa, P. & Perczel, A. (2008). Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain. *Biochemistry* **47**, 6936-45.
103. Dosztanyi, Z., Meszaros, B. & Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745-6.
104. Meszaros, B., Toth, J., Vertessy, B. G., Dosztanyi, Z. & Simon, I. (2011). Proteins with complex architecture as potential targets for drug design: a case study of Mycobacterium tuberculosis. *PLoS Comput Biol* **7**, e1002118.
105. Onozaki, I. & Raviglione, M. (2010). Stopping tuberculosis in the 21st century: goals and strategies. *Respirology* **15**, 32-43.
106. Pieters, J. (2008). Mycobacterium tuberculosis and the macrophage: maintaining a balance. *Cell Host Microbe* **3**, 399-407.
107. Borrell, S. & Gagneux, S. (2009). Infectiousness, reproductive fitness and evolution of drug-resistant Mycobacterium tuberculosis. *Int J Tuberc Lung Dis* **13**, 1456-66.
108. Hasan, S., Daugelat, S., Rao, P. S. & Schreiber, M. (2006). Prioritizing genomic drug targets in pathogens: application to Mycobacterium tuberculosis. *PLoS Comput Biol* **2**, e61.
109. Raman, K., Yeturu, K. & Chandra, N. (2008). targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Syst Biol* **2**, 109.
110. Kim, Y., Koyuturk, M., Topkara, U., Grama, A. & Subramaniam, S. (2006). Inferring functional information from domain co-evolution. *Bioinformatics* **22**, 40-9.
111. Glazko, G. V. & Mushegian, A. R. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol* **5**, R32.
112. Brennan, M. J. & Delogu, G. (2002). The PE multigene family: a 'molecular mantra' for mycobacteria. *Trends Microbiol* **10**, 246-9.

113. Alber, T. (2009). Signaling mechanisms of the Mycobacterium tuberculosis receptor Ser/Thr protein kinases. *Curr Opin Struct Biol* **19**, 650-7.
114. Wehenkel, A., Bellinzoni, M., Grana, M., Duran, R., Villarino, A., Fernandez, P., Andre-Leroux, G., England, P., Takiff, H., Cervenansky, C., Cole, S. T. & Alzari, P. M. (2008). Mycobacterial Ser/Thr protein kinases and phosphatases: physiological roles and therapeutic potential. *Biochim Biophys Acta* **1784**, 193-202.
115. Scherr, N., Muller, P., Perisa, D., Combaluzier, B., Jeno, P. & Pieters, J. (2009). Survival of pathogenic mycobacteria in macrophages is mediated through autophosphorylation of protein kinase G. *J Bacteriol* **191**, 4546-54.
116. Sassetti, C. M. & Rubin, E. J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci U S A* **100**, 12989-94.
117. Gey van Pittius, N. C., Sampson, S. L., Lee, H., Kim, Y., van Helden, P. D. & Warren, R. M. (2006). Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol Biol* **6**, 95.
118. Kruh, N. A., Troudt, J., Izzo, A., Prenni, J. & Dobos, K. M. (2010). Portrait of a pathogen: the Mycobacterium tuberculosis proteome in vivo. *PLoS One* **5**, e13938.
119. Dunker, A. K. & Uversky, V. N. (2010). Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Curr Opin Pharmacol* **10**, 782-8.
120. Babushok, D. V., Ostertag, E. M. & Kazazian, H. H., Jr. (2007). Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci* **64**, 542-54.
121. Pajkos, M., Meszaros, B., Simon, I. & Dosztanyi, Z. (2012). Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst* **8**, 296-307.
122. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83.
123. Xia, Y., Franzosa, E. A. & Gerstein, M. B. (2009). Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Comput Biol* **5**, e1000413.
124. Liu, J., Zhang, Y., Lei, X. & Zhang, Z. (2008). Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol* **9**, R69.
125. Fuxreiter, M., Tompa, P. & Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23**, 950-6.
126. Bledsoe, R. K., Montana, V. G., Stanley, T. B., Delves, C. J., Apolito, C. J., McKee, D. D., Consler, T. G., Parks, D. J., Stewart, E. L., Willson, T. M., Lambert, M. H., Moore, J. T., Pearce, K. H. & Xu, H. E. (2002). Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. *Cell* **110**, 93-105.
127. Farooq, A., Chaturvedi, G., Mujtaba, S., Plotnikova, O., Zeng, L., Dhalluin, C., Ashton, R. & Zhou, M. M. (2001). Solution structure of ERK2 binding domain of MAPK phosphatase MKP-3: structural insights into MKP-3 activation by ERK2. *Mol Cell* **7**, 387-99.

128. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins* **52**, 573-84.
129. Meszaros, B., Simon, I. & Dosztanyi, Z. (2011). The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins. *Phys Biol* **8**, 035003.