Univerzita Karlova
Přírodovědecká fakulta

Studijní program:
Molekulární biologie a biochemie organismů

Zuzana Nováková

Stupeň neuspořádanosti struktur proteinů u prokaryot a eukaryot
Degree of protein structure disorder in prokaryotic and eukaryotic organisms

Bakalářská práce

Vedoucí práce:
prof. RNDr. Jiří Vondrášek, CSc.

Praha, 2024

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

Tímto bych chtěla poděkovat zejména svému vedoucímu prof. RNDr. Jiřímu Vondráškovi CSc., jako i celému týmu skupiny bioinformatiky na ÚOCHB. V neposlední řadě bych také ráda poděkovala své rodině a příteli za podporu při psaní této práce.

Table of Contents

Abstract

The structure-function paradigm of protein biology has been fundamentally changed in the last three decades by the discovery of intrinsically disordered proteins (IDPs) and regions (IDRs). These proteins have been identified as critical components in various cellular processes, including signaling, protein-protein interactions, and regulation.

While it is apparent that IDPs/IDRs are vital in the function of living organisms, the study of their structure has posed a great challenge. Despite recent advancements in NMR spectroscopy and deep learning algorithms for protein structure prediction, IDPs/IDRs remain a relatively unnkown territory, with significant gaps in knowledge about their behavior and function in living systems.

Although IDPs are present in all life forms, their abundance reveals a correlation between organismal complexity and degree of protein disorder. Prokaryotic organisms exhibit a much lower prevalence of IDPs than eukaryotic. Notably, a substantial degree of disorder is observed in unicellular parasitic protists, implying, that IDPs are fundamental in pathogenesis and the progression of diseases like malaria and toxoplasmosis.

In humans, malfunctions in IDPs are linked to many conditions, including neurodegenerative diseases such as Parkinsons's, Alzheimer's as well as various types of cancer. Understanding these proteins could significantly impact the development of therapeutic strategies for these conditions.

This thesis underscores the necessity of specialized computational tools for accurate prediction of IDPs/IDRs to fully understand their function and significance in living organisms.

Key words: Intrinsically disordered proteins, Protein structure prediction, structural biology, Deep Learning

Abstrakt

Paradigma vztahu mezi strukturou a funkcí proteinů prošlo v posledních třiceti letech revolucí s objevem vnitřně neuspořádaných proteinů (IDP) a regionů (IDR). Tyto proteiny se ukázaly být klíčovými pro řadu buněčných procesů, včetně signalizace, interakcí mezi proteiny a buněčné regulace.

Ačkoliv význam IDP/IDR pro funkci živých organismů je nesporný, jejich strukturní analýza představuje významnou výzvu. I přes pokroky v NMR spektroskopii a v algoritmech hlubokého učení pro predikci struktur proteinů zůstávají IDP/IDR stále relativně neznámou oblastí, se značnými mezerami ve znalostech o jejich chování a funkci v živých systémech.

Vnitřně neuspořádané proteiny (IDP) se vyskytují ve všech živých organismech, ale jejich hojnost ukazuje na korelaci mezi složitostí organismů a stupněm neuspořádanosti proteinů. Prokaryotické organismy vykazují mnohem nižší výskyt IDPs než eukaryotické. Zvláště významný stupeň neuspořádanosti je pozorován u jednobuněčných parazitických protistů, což naznačuje, že IDP mají zásadní význam v patogenezi a průběhu nemocí jako je malárie a toxoplazmóza.

U lidí jsou dysfunkce IDP spojeny s mnoha onemocněními, včetně neurodegenerativních chorob, jako je Parkinsonova a Alzheimerova nemoc, jako různé typy rakoviny. Porozumění těmto proteinům by mohlo významně ovlivnit vývoj léčebných strategií pro tyto nemoci.

Tato práce zdůrazňuje nutnost vývoje specializovaných programů pro predikci neuspořádných proteinů, aby bylo možné plně porozumět jejich významu a funkci v organismech.

Klíčová slova: Vnitřně neuspořádané proteiny, Predikce struktury proteinů, Strukturní biologie, Hluboké učení

List of abbreviations

IDP	Intrinsically disordered protein
IDR	Intrinsically disordered region
AA	Amino acid
PTM	Post translation modification
TF	Transcription factor
LDR	Long disordered region
NMR	Nuclear magnetic resonance
MD	Molecular dynamics
STAT1	Signal transducer and activator of transcription

1 Introduction to Protein Structure Disorder

Proteins play a vital role in any living organism, whether as enzymes, catalysing various biochemical reactions, as signalling molecules or as the building blocks of cells. Traditionally, the protein's function has been associated with a three-dimensional structure, often described by a hierarchy of their secondary, tertiary, and quaternary arrangements. This type of ordered protein structure has been believed to play a crucial role in the proteins' ability to function in a multitude of cellular processes (Alberts, 2015).

This view of proteins as organized structures has been challenged in recent years by the discovery of intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs). These proteins and regions are characterised by a lack of fixed secondary or tertiary structure. This structural fluidity allows them to fulfil functions that require a degree of versatility that structured proteins usually do not possess (Dunker et al., 2002). IDPs/IDRs are prevalent across all forms of life and while their abundance is varying across domains, they are certainly not anomalies.

1.1 Overview of Protein Structures: Structured vs. Unstructured

Proteins are the most structurally and functionally complex molecules found in organisms. They consist of 20+2 canonical amino acids (AAs) connected through a peptide bond, resulting in the creation of the peptide backbone. Each AA has different physical and chemical properties, such as polarity and pK_a. The sophisticated properties of proteins are a result of the location of every AA in the chain as well as the context of other amino acids (Alberts, 2015).

Ordered proteins maintain their stable, three-dimensional structure trough diverse interactions, such as Wan der Waals attractions, electrostatic attractions, and hydrophobic interactions of nonpolar side chains. These enable the process of protein folding, which happens spontaneously, as it is an energetically favourable process. The main driving force of folding are hydrophobic interactions of aromatic and aliphatic side chains (Dobson, 2003).

The structure of a protein can be described by the AA sequence, also known as primary structure, which folds into a stable 3D structure (secondary structure) or stays unstructured. The overall three-dimensional shape of a protein molecule is known as tertiary structure. Most proteins also interact with each other, very often forming quaternary structure consisting of several polypeptide chains. This is the complex and functional protein

architecture usually found in cells. Each structural level is pivotal as it allows for specific interactions with other biomolecules. All these properties of structured proteins allow them to catalyse chemical reactions, be able to respond to cellular signals and be the building blocks of many cellular structures (Alberts, 2015).

In contrast with this well-ordered hierarchy of framework, IDPs lack fixed three-dimensional structure. These proteins do not fold into a specific conformation, rather, they remain flexible and capable of adopting a multitude of conformations. The intrinsic disorder is not completely random and is regulated by many weak interactions, in a similar way to ordered protein structures (Wright & Dyson, 1999). IDPs show a wide spectrum of conformational stages, ranging from completely unstructured to partially structured parts, often appearing and disappearing quickly as seen in Figure 1. This flexibility enables them to modulate their structure in response to their environment and change their conformational state as needed (Dunker et al., 2008).

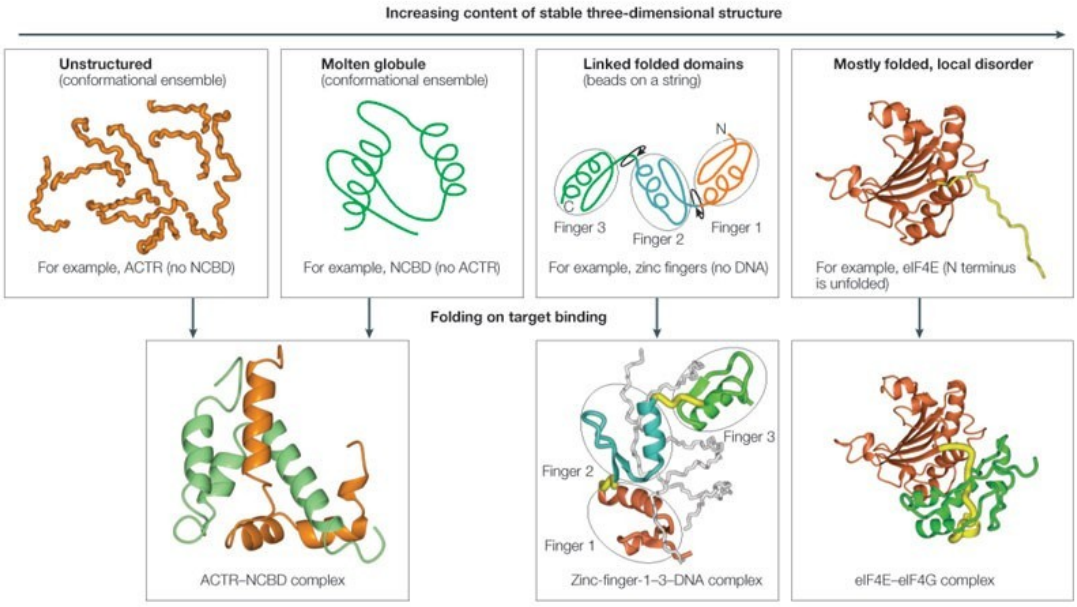

Figure 1-Different conformational stages a protein can adapt, from completely unstructured, partially structured, to completely structured. (Dyson & Wright, 2005)

1.2 The process of protein folding

The process of protein folding is too sophisticated to be left to chance, as presented in the Levinthal Paradox (Levinthal, 1969). Rather, protein folding could be conceptualized as navigating an energy landscape funnel as seen in Figure 2. Within this funnel, ensembles of partially folded intermediates undergo a series of energetically favorable transitions, progressively acquiring a more defined structure. This funnel-like landscape guides the protein towards its native conformation in a non-random, efficient manner (Tovchigrechko & Vakser, 2001). For a big part, these transitions are driven by hydrophobic interactions of non-polar side chains as there is only a limited number of conformations that can create the hydrophobic core (Kalinowska et al., 2017), with maximum efficiency (Cheung et al., 2002)

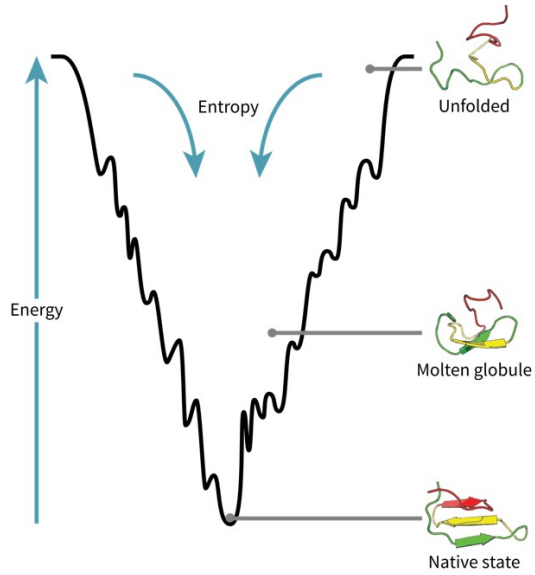

Figure 2-energy landscape funnel (image taken from scistyle.com)

1.3 The Functional Versatility of Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs)

Intrinsically disordered proteins and regions, thanks to their ability to adapt very quickly to their environment, hold a key position in many processes that require immediate response. Their ability to engage in reversible binding with high specificity and low affinity makes them ideal conductors in cellular signalling (Bondos et al., 2022). IDPs and IDRs often contain post-translational modification sites (PTMs) serving as switches for many additional functions. Phosphorylation for instance, allows IDRs to undergo fast conformational changes that alter their binding abilities (Jin & Gräter, 2021).

These changes can be achieved through Molecular Recognition Features, short sequences of disorder that can undergo quick disorder-to-order transition upon binding to their ligand. For instance, the transcription factor (TF) p53 contains a structured DNA binding domain, whereas its N-terminal and C-terminal domains remain unstructured. These unstructured domains constitute only 29% of the protein's total sequence but are responsible for mediating 71% of its protein-protein interactions. Additionally, the vast majority of PTMs, which are critical for the protein's regulatory functions, occur within these unstructured regions (Dunker et al., 2008). IDPs/IDRs play a crucial role in the function of many TFs. Their acidic character is a main driving force of binding events via electrostatic interactions with basic DNA binding sites. Intramolecular interactions of IDRs can also serve as regulatory mechanisms for DNA selectivity resulting in better cellular cycle regulation as shown on the human proto-oncogene MYC, where IDRs serve in several mechanisms resulting in a highly controlled DNA binding (Schütz et al., 2024).

Furthermore, their ability to serve in the process of liquid-liquid phase separation is a crucial intracellular process that allows for compartmentalization of organelles or biochemical processes without the need for membranes. These membrane-less organelles aren’t entirely separated from their surrounding which enables them to interact with the rest of the cytoplasmatic space quickly and effectively. IDPs/IDRs are very abundant in these organelles and are believed to be one of the main inducers of their formation (Uversky et al., 2015).

However, malfunctions in IDPs are linked to the pathogenesis of many diseases. Aggregation of alpha-synuclein (Fink, 2006)and amyloid-beta (Kirkitadze et al., 2001) stand behind the development of Parkinson’s and Alzheimer’s diseases, two most common neurodegenerative diseases. Due to the fact that IDPs play a vital role in cellular signalling, misfolded IDPs play a role in development of many types of cancer (Mark et al., 2005, Hollstein et al., 1991).

2 Properties and Classification of Amino Acids in IDPs

This chapter focuses on different physicochemical properties of amino acids (AAs) and how they influence the formation of structure in proteins. IDPs have different AA composition compared to their structured counterparts, which gives them special physicochemical characteristics.

Disordered proteins can appear in a vast array of conformational states meaning their degree of disorder can vary hugely. Correctly classifying the degree and type of disorder in proteins is vital for their understanding.

2.1 Amino Acid Properties Leading to Structure Disorder

AAs can be divided in groups based on the character of their side chain, which can be charged, uncharged, hydrophobic, or hydrophilic. Hydrophobic AAs are integral in the process of folding as discussed above. While sequence is not the only metric to predict disorder in proteins, there are visible preferences and disordered regions show a bias towards specific AAs. Structured protein regions predominantly contain Cysteine (C), Tryptophan (W), Tyrosine (Y), Isoleucine (I), Phenylalanine (F), Valine (V), Leucine (L), Histidine (H), Threonine (T), and Asparagine (N). AAs such as Aspartic acid (D), Methionine (M), Lysine (K), Arginine (R), Serine (S), Glutamine (Q), Proline (P), and Glutamic acid (E) tend to promote disorder, whereas Alanine (A) and Glycine (G) are considered neutral. This preference is then manifested by low mean hydrophobicity and high net charge of the IDP/IDR. Consequently, they lack a hydrophobic core, thus making it not energetically favorable to fold into a typical globular shape (Uversky, 2002). This allows for a much greater surface-to-volume ratio, enabling these proteins to have a larger interface for interactions. This is especially required in cellular environments with high molecular crowding as it enhances the binding capacity of these proteins (Gunasekaran et al., 2003).

Although the AA composition of a protein plays a role in its degree of disorder, the distribution of different types of amino acids is crucial as well. IDPs/IDRs seem to be organized into statistically significant “modules” of amino acids, that show similar properties, such as charge or polarity. These “modules” repeat along the sequence in an organized matter. Therefore, it is possible to say that the distribution of amino acids along the sequence of IDPs/IDRs is not random and a repeating pattern is observed (McConnell & Parker, 2023).

2.2 Specific Physicochemical Properties of IDPs and IDRs

IDPs have unique properties due to the lack of three-dimensional structure and their specific AA composition. This leads to different responses to environmental changes such as temperature or pH in comparison with globular proteins.

When exposed to high temperature, globular proteins usually denature and lose their tertiary structure, which leads to losing their ability to function correctly (Wu & Wu, 1925). However, in IDPs high temperature promotes the formation of fully reversible partial secondary structure. This effect is accounted to the fact, that high heat strengthens the force of hydrophobic interactions, allowing IDPs to undergo partial folding (Uversky, 2009).

Likewise, pH can affect the structure of IDPs significantly. Rapid shifts in pH either towards more acidic or basic lead to partial folding of IDPs. In normal conditions with neutral pH, IDPs have a high net charge, leading to electrostatic repulsion, keeping them unfolded. Rapid change of pH decreases net charge of the protein leading to lower electrostatic repulsion. This allows for hydrophobic interactions to become more dominant and induce partial folding of the protein (Uversky, 2009).

These specific properties are vital for the function of IDPs as these mild changes of environment allow them to react quickly to different types of signals.

2.3 Classifying the Degree of Disorder in Proteins

Protein structure disorder is a wide spectrum of conformational states with a varying level of secondary structure present. One way to classify the degree of disorder is through seeing proteins as a composition of foldons - smaller units of protein that fold independently and often work as smaller functional units in the protein (Panchenko et al., 1996). This perspective leads to the conclusion that every foldon within a protein can possess different degree of structure at a given time.

The structure of foldons can be divided into four categories. Independently foldable foldons adopt and maintain a stable three-dimensional structure which is usually closely tied to their function. Inducible foldons can rapidly change their structure upon binding to a ligand and are often present in signalling pathways. Semifoldons are regions that are always partially folded, representing an intermediate between order and disorder. Nonfoldons are regions that remain completely unfolded in physiological conditions, thus show the highest degree of

structural disorder. They serve as sites for PTMs as interaction platforms for different binding partners (Uversky, 2015).

This classification shows that disorder in proteins is not binary but a broad spectrum of different conformational states that can change quickly in time.

3 Methodologies for Identifying IDPs/IDRs

The identification of IDPs has been proven to be problematic in many ways. Traditional experimental methods, such as X-ray crystallography, show bad accuracy with IDPs (Dunker et al., 2001), so different approaches must be implemented (for example NMR spectroscopy). This results in the lack of experimental data, which, combined with specific characteristics of IDPs are both contributing to the fact, that predictive tools are still not able to bridge this gap in our knowledge of IDPs entirely (Kurgan, 2022). This chapter focuses on different approaches for identification of protein disorder and structure of IDPs while outlining the fact, that readily available tools are still lacking the robustness seen in prediction of structured proteins (Pearce & Zhang, 2021).

3.1 Experimental Approaches to Define IDPs/IDRs

IDPs and their structure are hard to observe via X-ray crystallography. Due to their fluid character and movement, the resulting image shows bad resolution, but X-ray crystallography is the primary source of structural information about proteins (currently there are 188 120 protein structures in PDB, 162 435 of them were acquired by X-ray crystallography (Bank, Retrieved 18 March 2024)). IDPs can be detected via X-ray crystallography only as the lack of electron density, which, however, can have different reasons.

A method that gives more precise information about the structure of IDPs is nuclear magnetic resonance (NMR) spectroscopy. IDPs exhibit a larger amount of signal overlap in NMR than structured proteins. This problem can be overcome by using isotopic labelling, such as nuclei of ^{15}N, ^{1}HN and ^{13}CO (Yao et al., 1997) and multidimensional NMR (Bermel et al., 2006). In recent years this method has been tailored to the specific challenges associated with IDPs, such as combining it with small-angle-X-ray scattering (SAXS) (Bernadó & Blackledge, 2009) or focusing on specific residues (Felli et al., 2021).

With these advancements, NMR spectroscopy can be a reliable source of information about IDPs, though there is a need for high concentration of protein which can be hard to acquire, and it is not a method suitable for large proteins.

3.2 Traditional Predictive Methods

Methods that predict the presence of IDRs can be classified in several categories based on the way they determine whether IDRs are present. One of the most widely used methods is the combination of supervised learning on experimentally acquired data and the usage of different factors implying disorder. A pioneering program of this type is PONDR, a program still widely utilized in research nowadays. Using a moving window of several residues it takes in account neighbouring AAs. The model uses different factors that make it easier to describe the level of disorder, such as frequency of each AA in the window, K2-entropy (measuring the complexity of the local sequence, as less complex areas are more likely to be disordered) and flexibility index of every AA (PENG et al., 2011). The model then uses a neural network based on a training set of experimentally acquired data (Peng et al., 2006). Another widely used predictive tool of similar type is DISOPRED that uses training data acquired from high resolution X-ray crystallography to determine disordered regions and predicts disorder based on them (Ward, McGuffin, et al., 2004). There are more predictive methods that work in a similar manner such as DisEMBL (Linding et al., 2003) or NORSnet (Schlessinger et al., 2007).

Another way to determine whether a protein contains IDRs is through physicochemical properties of AAs and their interactions. This approach is utilized by IUPred, program that approximates pairwise energy of sequence and, as IDPs have a higher pairwise energy, determines whether the sequence is structured or not using a scoring table (Dosztányi et al., 2005, Erdős et al., 2021).

3.3 Comparing Accuracy of Predictive Methods

A multitude of large-scale analyses on the accuracy of IDP predictors have been conducted. A problematic aspect of this testing is the relatively small amount of experimentally acquired data from NMR spectroscopy to compare with predictions. There has also been a visible disparity between the quality of prediction for different proteins and regions in proteins, with each predictor performing the best on different parts of given dataset. In a comparative analysis, the predictor with the highest accuracy in general was the most accurate in just 30 % of all predictions (Katuwawala et al., 2020). What's more, in the same article, it was concluded, that proteins with the highest degree of disorder (from experimental data) are in general harder to predict (Katuwawala et al., 2020).

DISOPRED3 and SPOT-Disorder seem to be the most accurate widely used predictors, but even the overall worst performing predictors were the most accurate for at least one protein. There is still around 18 % of proteins tested that none of the available tools could predict accurately as showed in Figure 3B (Katuwawala et al., 2020). The remarkable conformational plasticity of alpha-synuclein: Blessing or curse? *Trends in Molecular Medicine*

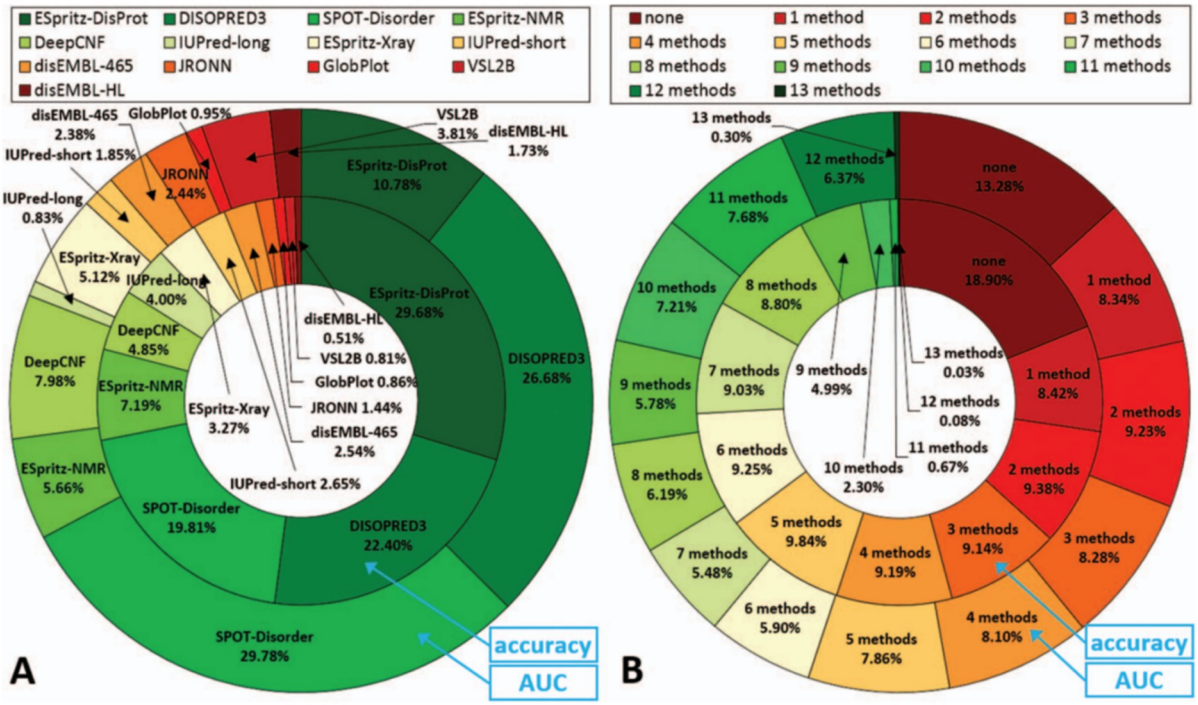

Figure 3- Comparison of 13 disorder predictors. Panel A shows the fraction of proteins for which the given predictor gives best results compared to others.

Panel B shows the fraction of proteins for which the given number of predictors gives accurate results. The inner ring shows results when using accuracy and outer ring shows the area under curve (Katuwawala et al., 2020).

As each of these models was trained on specific data set, their accuracy varies across different types of proteins. Based on these findings, it is best to use a consensus-based analysis, as even the best performing predictors don't give accurate results in many cases and a single predictive tool that is universally reliable is yet to be developed (Kurgan, 2022).

4 Advancements in Deep Learning for Prediction of IDP structure

Many computational tools dedicated to IDPs have been introduced during the last two decades, but they were mostly focused on identifying the existence of disordered regions in the sequence and not their actual 3D structure. Since the introduction of tools such as AlphaFold, the possibilities of protein 3D structure prediction have changed drastically and are much more reliable (Jumper et al., 2021). While the 3D architecture of structured proteins has been shown to be predicted with high accuracy, IDPs and IDRs are still harder to identify even with these advanced methods. In this chapter I focus on several methods that have emerged in the last years and that provide some insight into the architecture of disordered proteins.

4.1 Deep Learning based Methods for Predicting IDPs: An Overview

Since the introduction of AlphaFold2 in 2021 protein structure prediction has become almost as precise as experimental methods in many cases, outperforming any other predictive tool by far (Pereira et al., 2021, Jumper et al., 2021). This however is not true for IDPs, for which AlphaFold does not produce the same precision of results as for structured proteins. The predictions are often inaccurate, based on low structural probability with low confidence level (Aderinwale et al., 2022). This can be largely accounted to the fact that AlphaFold wasn't trained on a dataset with high content of disordered proteins (Jumper et al., 2021).

Molecular Dynamics (MD) simulations can be used as a tool in combination with deep learning methods and predict the structure of IDPs thanks to the development of specific force fields tailored for them (Wu et al., 2018). These simulations combined with deep learning methods are utilized by models such as ALBATROSS, that provides structural information from sequence. While this model shows good results when compared to experimentally acquired data, it still underestimates several factors such as hydrophobicity of aliphatic residues or interactions of charged residues. These factors could lead to inaccurate results in some cases (Lotthammer et al., 2024).

Phanto-IDP is also a deep learning model combined with MD simulations. It can provide reliable data for backbone structure and dynamic of IDPs. This model still has some limitations that arise from its high reliance on training data set. This can result in incorrect predictions for different states of proteins such as before and after binding to a ligand if the program was only trained on one of them (Zhu et al., 2023)

4.2 Evaluation of AI Predictive Accuracy

AlphaFold can be efficient in predicting structures of IDPs that adopt a defined conformation upon binding to a ligand. Those conditionally folded proteins are usually predicted with high confidence score and the structure shown resembles the bound and structured state (Alderson et al., 2023). Proteins that do not adopt structure under physiological conditions at all are predicted with low confidence and the predictions are not precise (Aderinwale et al., 2022)

Combining more tools together can produce more accurate results. Using AlphaFold in combination with Rosetta ResidueDisorder, tool specifically developed for identification of disordered regions, provides promising results at 73.7% accuracy compared to NMR data (He et al., 2022).

New methods combining MD simulations with deep learning models seem to have promising results, yet their predictive accuracy is not robust and usually works only for a subset of IDPs (Zhu et al., 2023).

4.3 Concluding Remarks on the Impact of AI in IDP Research

Deep learning models have changed the way we study protein structure significantly (Jumper et al., 2021). While predictive methods such as AlphaFold are immensely accurate for structured proteins, their use for IDPs is still limited. Programs tailored for predicting structures of IDPs are being constantly developed, yet they still lack the accuracy compared to structured protein prediction. The complexities of IDPs pose a great challenge in predicting their 3D structure in a way that accurately reflects their real movement and quick changes in degree of disorder, but new methods are being developed constantly and with the use of deep learning methods it’s it’s conceivable that future breakthroughs may well enable us to predict the structures and behavior of IDPs with high precision.

5 Defining Model Organisms and Occurrence of IDPs/IDRs

While IDPs are present in all forms of life, there is a visible corelation between the portion of disordered regions and increasing organism complexity. Their presence in bacteria and archaea is lower in comparison to eukaryotes as seen in Figure 4 (Schad et al., 2011). This disparity could be accounted to a bigger demand for complex signalling and regulatory pathways in compartmentalised cells of eukaryotes. What's more, the transition in IDP/IDR fraction of proteome is not smooth; while most prokaryotic and archaeal proteomes consist of a maximum of 28 % of IDPs/IDRs, eukaryotic organisms usually have no less than 32 % of disordered regions (Xue et al., 2012).

This relationship between biological complexity (defined by the number of unique specialized cell types) and degree of disorder is not observed in eukaryotes (Schad et al., 2011) . The highest degree of disorder in eukaryotic organisms is surprisingly found in protists, unicellular organisms. This fact sets the whole corelation off. The specific phyla with the highest degree of disorder are parasites responsible for serious diseases both in humans and other animals. These findings could be of great clinical importance disordered proteins play a crucial role in their pathogenesis as discussed in chapter 3.2 (Pancsa & Tompa, 2012).

In this chapter, I focus on three model organisms to demonstrate the difference in abundance of protein disorder and pinpoint different ways IDPs work in organisms with extremely high occurrence of disordered proteins. Moreover, I choose to compare cellular localisation of IDPs as this information can be very beneficial in deepening our understanding of the various roles IDPs take on in the cell.

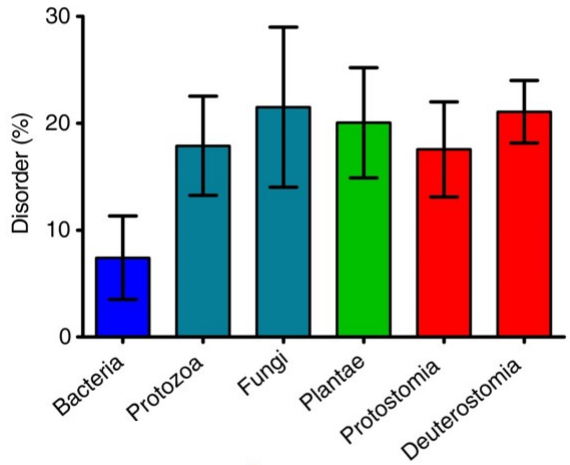

Figure 4-Species divided into six groups and their average degree of protein structure disorder. The graph shows major difference between bacteria and eukaryota. In this experiment only a small subset of each phyla was taken into account, which explains such a large standard deviation in fungi and plants, yet it still demonstrates the large variability between these kingdoms *(Schad et al., 2011)*.

5.1 Comparison of IDPs in selected organisms

To accurately compare the degree of disorder in different organisms, I examine the abundance and localisation of IDPS/IDRS in *Escherichia Coli* as model organism of bacteria, *Homo Sapiens* as an example of a multicellular eukaryote and *Toxoplasma Gondii*, as it is one of the eukaryotes with the highest known degree of protein disorder (Ahrens et al., 2018). This comparison shows the extreme variety in abundance of disorder between prokaryotes and eukaryotes and highlights the specific functions IDPs are crucial for.

E. Coli is the most used model organism of prokaryotes. It has a genome of around 4.5-5.5 Mb and approximately 4500 coding sequences (Engelbrecht et al., 2017). Using two different predictive methods (PONDR and IUpred), it was estimated that long sequences of disorder (measured as sequence of >30 residues, as opposed to shorter disordered sequences that are often found in linkers) make up only 3.6 or 5.5 % of total proteome respectively (Tompa et al., 2006). Highest degree of disorder was found in proteins taking part in cell cycle regulation, and organelle/membrane organisation.

Humans, in contrast, have a much higher degree of protein disorder. Out of more than 100 000 proteins encoded by our DNA, 35 % of them contain sequences of disorder longer than 30 residues and 21 % have disordered sequences of at least 50 residues (Ward, Sodhi, et al., 2004).

Parasitic protists possess the highest known degree of protein disorder. The degree of disorder is dependent on the lifestyle of protists, as obligate parasites with a simple life cycle have a much smaller amount of IDPs compared to nonparasitic species or species that have a complicated life cycle with a multitude of hosts. Largest known degree of protein disorder is found in *Toxoplasma Gondii*, a parasite responsible for toxoplasmosis, with 65 % of its proteins containing at least one long (<30 residues) disordered sequence (Ward, Sodhi, et al., 2004). This fact cannot be accounted to extremely reduced genome as it is in some parasitic species, as *Toxoplasma Gondii* has more than 8 000 proteins encoded in its genome.

Toxoplasma Gondii can infect most warm-blooded animals, meaning it needs to have a plethora of different cellular mechanisms enabling it to adapt to such a wide variety of hosts. A big part of proteins taking part in these cellular processes shows a significant degree of disorder, highlighting the functional and structural versatility of IDPs (Feng et al., 2006).

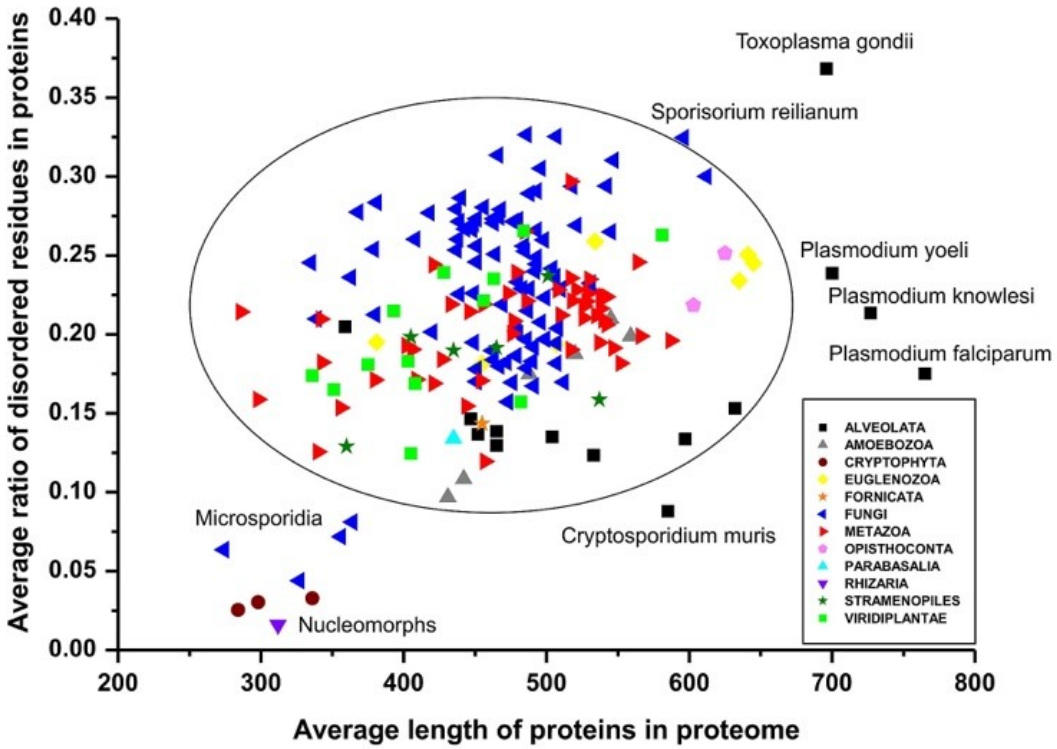

Figure 5-The relationship between average protein length and average degree of disorder in proteomes of different eukaryotic organisms (Ward, Sodhi, et al., 2004).

5.2 Insights from organisms with high IDP profiles

As mentioned in chapter 3.1, the highest degree of disorder in eukaryotic organisms is found in *Apicomplexa* which can be seen in Figure 5. This phylum is entirely parasitic, with some species causing serious human and animal diseases such as malaria, toxoplasmosis, or coccidiosis. These protists utilize IDRs for the process of infection. For instance, TgIST (Toxoplasma inhibitor of STAT1-dependent transcription), an IDP found in *Toxoplasma gondii*, s ilences STAT1 (signal transducer and activator of transcription) signalisation and effectively blocks interferon mediated response to infection (Huang et al., 2022). While the complete structure of this protein is still unknown, it has been found that it is completely unstructured when not in bound state and undergoes a fast disorder-to-order transition upon binding to STAT, which is a common mechanism seen in IDPs. The disordered character allows this protein to react quickly to its environment and mediate this pathway in many different species (Huang et al., 2022). Another reason why this parasite is so widespread is its resilience. This can be in part accounted to Late Embryogenesis Abundant proteins, a group of intrinsically disordered

proteins enabling *Toxoplasma* embryos to endure extreme conditions such as high and low temperature, high salinity and many more, making them extremely resistant to environmental stress (Arranz-Solís et al., 2023).

High degree of disorder can also be found in many different proteins of *Plasmodium*. These proteins play a crucial role in the process of infection as well. One of them is EBL (erythrocyte binding ligand) protein that is required for successful invasion of host red blood cells. Many surface proteins of plasmodium exhibit a high degree of disorder. This attribute is beneficial for the parasite as it helps with epitope masking in host organisms (Naung et al., 2022). For instance, MSP2 (merozoite surface protein 2) of *Plasmodium falciparum* can escape antibody recognition in many cases (Morales et al., 2015). *Plasmodium* in general possesses a high degree of disorder localised in its parasitophorous vacuole, apical complex and exported proteins, all of which are highly involved in the infection process of this organism.

This insight can be of great use when designing vaccines against malaria by targeting specific IDPs such as EBL or MSP2 and it can lead to more effective ways to prevent this parasite from evading the host immune system. Deeper knowledge of these proteins can lead to the development of vaccines that induce a robust and lasting immune response against different life stages of *Plasmodium (Naung et al., 2022)*.

To conclude, it is evident that IDPs/IDRs play a vital role in some parasitic eukaryotes. Their versatility grants them the ability to infect many different species and survive extreme conditions, complicating the host's efforts to fight these pathogens. More insight into the disorder-to-order transitions and different interactions with other proteins is needed, but it is evident that these proteins could be a good possible target for novel therapeutic strategies as seen in recent research (Eacret et al., 2019), while still posing a challenge due to their high degree of polymorphism, which has been one of the main problems in designing effective vaccines (Barry & Arnott, 2014; Takala & Plowe, 2009).

5.3 Cellular Localization of IDPs: Comparative Analysis

Thanks to their specific function, IDPs are distributed rather unevenly across different cellular compartments. This disparity can be best observed in eukaryotic, highly compartmentalised cells.

A large-scale analysis of cellular localization of almost all human proteins was done in 2021, using Gene Ontology annotation of cellular localization and MobiDB-lite for prediction

of disorder (Zhao et al., 2021). MobiDB-lite utilizes a multitude of different predictive methods and combines the results together into a consensus prediction, resulting in a more precise prediction of disorder (Necci et al., 2017). In this study it was concluded that several cellular compartments are enriched in long disordered regions (LDRs) of more than 30 residues compared to others. The highest degree of disorder was found in the nucleus and cytoskeleton, with around 60 % of proteins containing at least one LDR. On the other hand, many cellular compartments were significantly depleted of LDRs, with lysosome, vacuolar membrane, and peroxisome being on the far end of this spectrum, with only around 20 % of proteins containing LDRs (Zhao et al., 2021). This finding is in agreement with the fact that IDRs are found mostly in proteins that moderate protein-DNA, protein-RNA and protein-protein interactions.

Many proteins can be found in more than one cellular compartment. It was concluded, that compartment specifity of a protein plays a crucial role in the degree of disorder found. While proteins associated with only one cellular compartment have a content of IDRs at 12 %, proteins associated with eight or more compartments have a 7 % degree of disorder (Zhao et al., 2021).

In summary, these findings highlight the specific functions of IDPs ~~serve~~. The enriched presence of IDPs with LDRs in the nucleus and cytoskeleton underscores their critical involvement in regulating gene expression, signal transduction, and maintaining cellular architecture. On the other hand, their smaller frequency in more metabolically focused organelles such as lysosome or peroxisome underscores the different requirements needed for enzymes (Zhao et al., 2021).

6 Conclusion

Protein disorder has been a pivotal topic of structural biology in the last two decades and research in this area has broken the structure-function dogma of protein biology. IDRs/IDPs play a crucial role in cellular signalling, protein-protein interactions, and regulation of many other cellular processes, which makes them indispensable parts of all living organisms.

While they are, to some degree, found in all forms of life, their abundance is not consistent between species and with a few exceptions, it correlates with the complexity of organisms. While prokaryotic organisms with a very low cellular complexity have a much lower degree of protein disorder than eukaryotes in general, this correlation is not visible among eukaryotic organisms, where the highest degree of disorder can be found in *Apicomplexa*. In this phylum of parasitic protists, proteins with long disordered regions can make up to 65% of the whole proteome and they play a vital role in regulating immune response of the host. As *Alveolata* are the cause of many diseases such as malaria and toxoplasmosis, research in this area could lead to a better understanding of the mechanisms of infection in these protists and potentially the development of vaccines for the diseases they cause.

In humans, aggregated IDRs are the root of many serious diseases, namely Alzheimer's and Parkinson's disease and, as they mediate cellular regulation, many different types of cancer. Deeper knowledge of these proteins and reasons leading to their malfunction is vital for their understanding and effective treatment.

One of the reasons why IDPs/IDRs were not more thoroughly studied in the past and a lot about them is still unknown today are the various complications with their detection by experimental methods and prediction *in silico*. X-ray crystallography, which is the most widely used experimental method for identification of protein structure is not suitable for IDPs. While mechanisms for protein structure prediction have improved tremendously in the past several years, their results with IDPs/IDRs are not reliable and, in many cases, predicted with very low confidence.

In conclusion, the study of intrinsically disordered proteins and intrinsically disordered regions is an important and until recently overlooked part of structural biology of major clinical significance and specific role in many cellular processes. With the emergence of new methods such as NMR spectroscopy and deep learning-based programs for protein prediction, IDPs/IDRs are slowly starting to be understood, yet there is need for further research for

programs to be specifically tailored for prediction of unstructured proteins and the way they work in living organisms.

7 Bibliography

Aderinwale, T., Bharadwaj, V., Christoffer, C., Terashi, G., Zhang, Z., Jahandideh, R., Kagaya, Y., & Kihara, D. (2022). Real-time structure search and structure classification for AlphaFold protein models. *Communications Biology*, *5*(1), 1–12. https://doi.org/10.1038/s42003-022-03261-8

Ahrens, J., Rahaman, J., & Siltberg-Liberles, J. (2018). Large-Scale Analyses of Site-Specific Evolutionary Rates across Eukaryote Proteomes Reveal Confounding Interactions between Intrinsic Disorder, Secondary Structure, and Functional Domains. *Genes*, *9*(11), 553. https://doi.org/10.3390/genes9110553

Alberts, B. (2015). *Molecular biology of the cell* (Sixth edition). Garland Science, Taylor and Francis Group.

Alderson, T. R., Pritišanac, I., Kolarić, Đ., Moses, A. M., & Forman-Kay, J. D. (2023). Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proceedings of the National Academy of Sciences*, *120*(44), e2304302120. https://doi.org/10.1073/pnas.2304302120

Arranz-Solís, D., Warschkau, D., Fabian, B. T., Seeber, F., & Saeij, J. P. J. (2023). Late Embryogenesis Abundant Proteins Contribute to the Resistance of Toxoplasma gondii Oocysts against Environmental Stresses. *mBio*, *14*(2), e02868-22. https://doi.org/10.1128/mbio.02868-22

Bank, R. P. D. (n.d.). *PDB Statistics: PDB Data Distribution by Experimental Method and Molecular Type*. Retrieved 18 March 2024, from https://www.rcsb.org/stats/summary

Barry, A. E., & Arnott, A. (2014). Strategies for Designing and Monitoring Malaria Vaccines Targeting Diverse Antigens. *Frontiers in Immunology*, *5*, 359.

https://doi.org/10.3389/fimmu.2014.00359

Bermel, W., Bertini, I., Felli, I. C., Kümmerle, R., & Pierattelli, R. (2006). Novel 13C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. *Journal of Magnetic Resonance*, *178*(1), 56–64. https://doi.org/10.1016/j.jmr.2005.08.011

Bernadó, P., & Blackledge, M. (2009). A Self-Consistent Description of the Conformational Behavior of Chemically Denatured Proteins from NMR and Small Angle Scattering. *Biophysical Journal*, *97*(10), 2839–2845. https://doi.org/10.1016/j.bpj.2009.08.044

Bondos, S. E., Dunker, A. K., & Uversky, V. N. (2022). Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Communication and Signaling*, *20*(1), 20. https://doi.org/10.1186/s12964-022-00821-7

Cheung, M. S., García, A. E., & Onuchic, J. N. (2002). Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 685–690. Scopus. https://doi.org/10.1073/pnas.022387699

Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), Article 6968. https://doi.org/10.1038/nature02261

Dosztányi, 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*(16), 3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic Disorder and Protein Function. *Biochemistry*, *41*(21), 6573–6582. https://doi.org/10.1021/bi012159+

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. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8

Dunker, A. K., Silman, I., Uversky, V. N., & Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, *18*(6), 756–764. https://doi.org/10.1016/j.sbi.2008.10.002

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology*, *6*(3), Article 3. https://doi.org/10.1038/nrm1589

Eacret, J. S., Gonzales, D. M., Franks, R. G., & Burns, J. M. (2019). Immunization with merozoite surface protein 2 fused to a Plasmodium-specific carrier protein elicits strain-specific and strain-transcending, opsonizing antibody. *Scientific Reports*, *9*(1), 9022. https://doi.org/10.1038/s41598-019-45440-4

Engelbrecht, K. C., Putonti, C., Koenig, D. W., & Wolfe, A. J. (2017). Draft Genome Sequence of Escherichia coli K-12 (ATCC 29425). *Genome Announcements*, *5*(27), e00574-17. https://doi.org/10.1128/genomeA.00574-17

Erdős, G., Pajkos, M., & Dosztányi, Z. (2021). IUPred3: Prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Research*, *49*(W1), W297–W303. https://doi.org/10.1093/nar/gkab408

Felli, I. C., Bermel, W., & Pierattelli, R. (2021). Exclusively heteronuclear NMR experiments for the investigation of intrinsically disordered proteins: Focusing on proline residues.

Magnetic Resonance, *2*(1), 511–522. https://doi.org/10.5194/mr-2-511-2021

Feng, Z.-P., Zhang, X., Han, P., Arora, N., Anders, R. F., & Norton, R. S. (2006). Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Molecular and Biochemical Parasitology*, *150*(2), 256–267. https://doi.org/10.1016/j.molbiopara.2006.08.011

Fink, A. L. (2006). The Aggregation and Fibrillation of α-Synuclein. *Accounts of Chemical Research*, *39*(9), 628–634. https://doi.org/10.1021/ar050073t

Gunasekaran, K., Tsai, C.-J., Kumar, S., Zanuy, D., & Nussinov, R. (2003). Extended disordered proteins: Targeting function with less scaffold. *Trends in Biochemical Sciences*, *28*(2), 81–85. https://doi.org/10.1016/S0968-0004(03)00003-3

He, J., Turzo, S. B. A., Seffernick, J. T., Kim, S. S., & Lindert, S. (2022). Prediction of Intrinsic Disorder Using Rosetta ResidueDisorder and AlphaFold2. *The Journal of Physical Chemistry B*, *126*(42), 8439–8446. https://doi.org/10.1021/acs.jpcb.2c05508

Hollstein, M., Sidransky, D., Vogelstein, B., & Harris, C. C. (1991). P53 mutations in human cancers. *Science (New York, N.Y.)*, *253*(5015), 49–53. https://doi.org/10.1126/science.1905840

Huang, Z., Liu, H., Nix, J., Xu, R., Knoverek, C. R., Bowman, G. R., Amarasinghe, G. K., & Sibley, L. D. (2022). The intrinsically disordered protein TgIST from Toxoplasma gondii inhibits STAT1 signaling by blocking cofactor recruitment. *Nature Communications*, *13*(1), 4047. https://doi.org/10.1038/s41467-022-31720-7

Jin, F., & Gräter, F. (2021). How multisite phosphorylation impacts the conformations of intrinsically disordered proteins. *PLOS Computational Biology*, *17*(5), e1008939. https://doi.org/10.1371/journal.pcbi.1008939

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalinowska, B., Banach, M., Wiśniowski, Z., Konieczny, L., & Roterman, I. (2017). Is the hydrophobic core a universal structural element in proteins? *Journal of Molecular Modeling*, *23*(7), 205. https://doi.org/10.1007/s00894-017-3367-z

Katuwawala, A., Oldfield, C. J., & Kurgan, L. (2020). Accuracy of protein-level disorder predictions. *Briefings in Bioinformatics*, *21*(5), 1509–1522. https://doi.org/10.1093/bib/bbz100

Kirkitadze, M. D., Condron, M. M., & Teplow, D. B. (2001). Identification and characterization of key kinetic intermediates in amyloid beta-protein fibrillogenesis. *Journal of Molecular Biology*, *312*(5), 1103–1119. https://doi.org/10.1006/jmbi.2001.4970

Kurgan, L. (2022). Resources for computational prediction of intrinsic disorder in proteins. *Methods*, *204*, 132–141. https://doi.org/10.1016/j.ymeth.2022.03.018

Levinthal, C. (1969). *How to fold graciously*. https://www.semanticscholar.org/paper/How-to-fold-graciously-Levinthal/1ef89dfb1e3404f4ace99399ce582b2bc982d0bf

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction: Implications for structural proteomics. *Structure (London, England: 1993)*, *11*(11), 1453–1459. https://doi.org/10.1016/j.str.2003.10.002

Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J., & Holehouse, A. S. (2024). Direct prediction of intrinsically disordered protein conformational properties from

sequence. *Nature Methods*, *21*(3), 465–476. https://doi.org/10.1038/s41592-023-02159-5

Mark, W.-Y., Liao, J. C. C., Lu, Y., Ayed, A., Laister, R., Szymczyna, B., Chakrabartty, A., & Arrowsmith, C. H. (2005). Characterization of segments from the central region of BRCA1: An intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? *Journal of Molecular Biology*, *345*(2), 275–287. https://doi.org/10.1016/j.jmb.2004.10.045

McConnell, B. S., & Parker, M. W. (2023). *Protein intrinsically disordered regions have a non-random, modular architecture* (p. 2023.05.10.539862). bioRxiv. https://doi.org/10.1101/2023.05.10.539862

Morales, R. A. V., MacRaild, C. A., Seow, J., Krishnarjuna, B., Drinkwater, N., Rouet, R., Anders, R. F., Christ, D., McGowan, S., & Norton, R. S. (2015). Structural basis for epitope masking and strain specificity of a conserved epitope in an intrinsically disordered malaria vaccine candidate. *SCIENTIFIC REPORTS*, *5*, 10103. https://doi.org/10.1038/srep10103

Naung, M. T., Martin, E., Munro, J., Mehra, S., Guy, A. J., Laman, M., Harrison, G. L. A., Tavul, L., Hetzel, M., Kwiatkowski, D., Mueller, I., Bahlo, M., & Barry, A. E. (2022). Global diversity and balancing selection of 23 leading Plasmodium falciparum candidate vaccine antigens. *PLOS Computational Biology*, *18*(2), e1009801. https://doi.org/10.1371/journal.pcbi.1009801

Necci, M., Piovesan, D., Dosztányi, Z., & Tosatto, S. C. E. (2017). MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, *33*(9), 1402–1404. https://doi.org/10.1093/bioinformatics/btx015

Panchenko, A. R., Luthey-Schulten, Z., & Wolynes, P. G. (1996). Foldons, protein structural modules, and exons. *Proceedings of the National Academy of Sciences*, *93*(5), 2008–2013. https://doi.org/10.1073/pnas.93.5.2008

Pancsa, R., & Tompa, P. (2012). Structural Disorder in Eukaryotes. *PLOS ONE*, *7*(4), e34687. https://doi.org/10.1371/journal.pone.0034687

Pearce, R., & Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, *297*(1), 100870. https://doi.org/10.1016/j.jbc.2021.100870

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, *7*(1), 208. https://doi.org/10.1186/1471-2105-7-208

PENG, K., VUCETIC, S., RADIVOJAC, P., BROWN, C. J., DUNKER, A. K., & OBRADOVIC, Z. (2011). OPTIMIZING LONG INTRINSIC DISORDER PREDICTORS WITH PROTEIN EVOLUTIONARY INFORMATION. *Journal of Bioinformatics and Computational Biology*. https://doi.org/10.1142/S0219720005000886

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., & Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1687–1699. https://doi.org/10.1002/prot.26171

Schad, E., Tompa, P., & Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, *12*(12), R120. https://doi.org/10.1186/gb-2011-12-12-r120

Schlessinger, A., Liu, J., & Rost, B. (2007). Natively Unstructured Loops Differ from Other Loops. *PLoS Computational Biology*, *3*(7), e140.

https://doi.org/10.1371/journal.pcbi.0030140

Schütz, S., Bergsdorf, C., Hänni-Holzinger, S., Lingel, A., Renatus, M., Gossert, A. D., & Jahnke, W. (2024). Intrinsically Disordered Regions in the Transcription Factor MYC:MAX Modulate DNA Binding via Intramolecular Interactions. *Biochemistry*, *63*(4), 498–511. https://doi.org/10.1021/acs.biochem.3c00608

Takala, S. L., & Plowe, C. V. (2009). Genetic diversity and malaria vaccine design, testing, and efficacy: Preventing and overcoming “vaccine resistant malaria”. *Parasite Immunology*, *31*(9), 560–573. https://doi.org/10.1111/j.1365-3024.2009.01138.x

Tompa, P., Dosztányi, Z., & Simon, I. (2006). Prevalent Structural Disorder in E. coli and S. cerevisiae Proteomes. *Journal of Proteome Research*, *5*(8), 1996–2000. https://doi.org/10.1021/pr0600881

Tovchigrechko, A., & Vakser, I. A. (2001). How common is the funnel-like energy landscape in protein-protein interactions? *Protein Science*, *10*(8), 1572–1583. https://doi.org/10.1110/ps.8701

Uversky, V. N. (2002). What does it mean to be natively unfolded? *European Journal of Biochemistry*, *269*(1), 2–12. https://doi.org/10.1046/j.0014-2956.2001.02649.x

Uversky, V. N. (2009). Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding. *The Protein Journal*, *28*(7), 305–325. https://doi.org/10.1007/s10930-009-9201-4

Uversky, V. N. (2015). Functional roles of transiently and intrinsically disordered regions within proteins. *The FEBS Journal*, *282*(7), 1182–1189. https://doi.org/10.1111/febs.13202

Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K., & Zaslavsky, B. (2015). Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Letters*, *589*(1), 15–22. https://doi.org/10.1016/j.febslet.2014.11.028

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, *20*(13), 2138–2139. https://doi.org/10.1093/bioinformatics/bth195

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. *Journal of Molecular Biology*, *337*(3), 635–645. https://doi.org/10.1016/j.jmb.2004.02.002

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *JOURNAL OF MOLECULAR BIOLOGY*, *293*(2), 321–331. https://doi.org/10.1006/jmbi.1999.3110

Wu, H., Wolynes, P. G., & Papoian, G. A. (2018). AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *The Journal of Physical Chemistry B*, *122*(49), 11115–11125. https://doi.org/10.1021/acs.jpcb.8b05791

Wu, H., & Wu, D. Y. (1925). NATURE OF HEAT DENATURATION OF PROTEINS. *Journal of Biological Chemistry*, *64*(2), 369–378. https://doi.org/10.1016/S0021-9258(18)84930-4

Xue, B., Dunker, A. K., & Uversky, V. N. (2012). Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics*, *30*(2), 137–149. https://doi.org/10.1080/07391102.2012.675145

Yao, J., Dyson, H. J., & Wright, P. E. (1997). Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. *FEBS Letters*, *419*(2), 285–289. https://doi.org/10.1016/S0014-5793(97)01474-9

Zhao, B., Katuwawala, A., Uversky, V. N., & Kurgan, L. (2021). IDPology of the living cell: Intrinsic disorder in the subcellular compartments of the human cell. *CELLULAR AND MOLECULAR LIFE SCIENCES*, *78*(5), 2371–2385. https://doi.org/10.1007/s00018-020-03654-0

Zhu, J., Li, Z., Tong, H., Lu, Z., Zhang, N., Wei, T., & Chen, H.-F. (2023). Phanto-IDP: Compact model for precise intrinsically disordered protein backbone generation and enhanced sampling. *Briefings in Bioinformatics*, *25*(1), bbad429. https://doi.org/10.1093/bib/bbad429