Exploring Feature Identification and Machine Learning in Predicting Protein-Protein Interactions of Disordered Proteins

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Gözde Kibar

Berlin, 2024

Erstgutachter: **Prof. Dr. Martin Vingron**
Zweitgutachter: **Prof. Dr. Dirk Walther**

Tag der Disputation: 16.02.2024

PUBLICATIONS

Chapter 3 contains a work that grew from a collaboration with the lab of Denes Hnisz. In this work, we developed a statistical method to identify periodic regions in transcription factors.

- Naderi J, Magalhaes A, Kibar G, Stik G, Zhang Y, Wieler H, Rossi F, Buschow R, Christou-Kent M, Alcoverro-Bertran M, Mackowiak S , Graf T, Vingron M, Hnisz D. Suboptimization of human transcription factors. (Submitted at *Nature Cell Biology*)

Chapter 6 presents a new method to tackle the challenge of protein-protein interaction prediction using intrinsically disordered regions.

- Kibar G, Vingron M. Prediction of protein-protein interactions using sequences of intrinsically disordered regions. *Proteins: Structure, Function, and Bioinformatics*. 2023 Jul;91(7):980-990. This article is licensed under a CC BY-NC-ND 4.0 license

The results from two additional collaboration projects have led to the following publications:

- Enervald E, Powell LM, Boteva L, Foti R, Blanes Ruiz N, Kibar G, Piszczek A, Cavaleri F, Vingron M, Cerase A, Buonomo SBC. RIF1 and KAP1 differentially regulate the choice of inactive versus active X chromosomes. *The EMBO journal*. 2021 Dec 15;40(24):e105862.

- Kulik M, Bothe M, Kibar G, Fuchs A, Schöne S, Prekovic S, Mayayo-Peralta I, Chung HR, Zwart W, Helsen C, Claessens F, Meijsing S H. Androgen and glucocorticoid receptor direct distinct transcriptional programs by receptor-specific and shared DNA binding sites. *Nucleic Acids Research*. 2021 Apr 19;49(7):3856-3875.

The initial collaboration with Buonomo lab involved a computational analysis of skewed X-Chromosome Inactivation in mice. The collaboration with Meijsing lab involved computational analyses of gene regulation between Androgen and Glucocorticoid receptor

using RNA-seq data. Results from these mentioned collaboration projects are not part of this thesis.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Martin Vingron for being supportive, understanding supervisor and inspirational teacher through my PhD journey. I have learned and grown a lot under his supervision. I am really grateful that I had the opportunity to pursue my PhD in such a nice and supportive atmosphere. I would like to also thank the members of my thesis advisory committee: Denes Hnisz and Dirk Walther. I appreciate their inputs during our meetings. I extend my graditude to Dirk Walther for reading and reviewing my thesis. I want to also thank the Max Planck Research School for Biology And Computation (IMPRS-BAC) of the Max Planck Institute for Molecular Genetics.

Of course, I would like thank the all amazing current and previous members of Vingron lab: Aybuge Altay, Hossein Moeinzadeh, Ela Gralinska, Mariam Ghareghani, Yan Zhao, Eldar Abdullaev, Tris Rapakoulia, Brigitte Bouman, Clemens Kohl, Ekta Shah, Ekin Aksu and all the others. I would like to give special thanks to Aybuge. No doubt, I will miss working with her in the same environment. Thanks for being such a good friend and personal consultant. I would like to thank my officemate Hossein, with whom I shared the same office for 5 years. I greatly enjoyed our discussions about science and life. Thanks to Aybuge, Ekta and Clemens for proofreading parts of this thesis.

I also want to thank the people I collaborated with during my PhD for their valuable scientific discussions: Denes Hnisz, Sebastian Meijsing, Sarah Kinkley, Julian Naderi, Francesca Rossi, Melissa Bothe, and Marina Kulik. I would like to thank our previous IMPRS-BAC PhD coordinator, Kirsten Kelleher, and her successor, Anne-Dominique Gindrat, for their help and patience during my PhD studies. I would also like to thank Martina Lorse for her consistent kindness and patience whenever we had questions.

Finally, I would like to thank my parents, Sebahat and Kazim Kibar, for their unconditional love and support, always being there for me. No words can describe how grateful I am to have them as parents. Last but not least, I would like to thank my partner, Tarik Yayla, for always believing in me and supporting me throughout this journey.

CONTENTS

1 INTRODUCTION

Today, we understand that proteins are not only composed of globular domains; there are also unstructured regions in proteins known as intrinsically disordered regions (IDRs). Protein segments referred to as IDRs lack a well-defined, stable three-dimensional structure, either entirely or partially, under physiological conditions (Babu et al., 2011). Proteins with disordered regions are termed intrinsically disordered proteins (IDPs). Traditionally, proteins were thought to solely function by adopting a single, well-defined, stable three-dimensional structure, representing the global energy minimum accessible to the polypeptide chain (Radivojac et al., 2007). However, around the early 2000s, a new understanding emerged: not all proteins can be categorized as rigid entities where polypeptide segments will adopt a stable three-dimensional structure. Unlike their more structured counterparts, IDPs do not follow the conventional funnel-shaped energy landscape with a clearly defined global energy minimum. Instead, their energy landscape is comparatively flat, with many local minima (Figure 1.1) (Uversky et al., 2008). As a consequence, their conformational changes occur rapidly (Sugase et al., 2007). While some IDPs need to remain unfolded or disordered to fulfill their functions, others only adopt a specific folded structure when interacting with particular target molecules (Wright et al., 2015).

IDRs are involved in crucial functions such as molecular recognition, signaling, binding, and transcriptional regulation (Dyson et al., 2005; Wright et al., 2015). Due to their involvement in crucial cellular processes and their prevalence within cells, understanding the fundamental roles of IDPs holds significant importance. The term 'intrinsic' disorder reflects the inherent lack of structure that can be found in the amino acid sequences of these proteins (Uversky et al., 2008). Numerous studies have shown that IDR sequences have embedded sequence features that enable them to guide their functions (Holehouse et al., 2021; Ravarani et al., 2018; Staller et al., 2018). These features include global characteristics distributed throughout the sequence as well as local attributes, such as motifs (Lu et al., 2022). Many undiscovered signatures still lie within these regions, waiting to be revealed through computational approaches. Recently, computational methods have gained considerable significance in tackling this challenge due to their capability to identify intricate attributes derived from sequences (Chong et al., 2021).

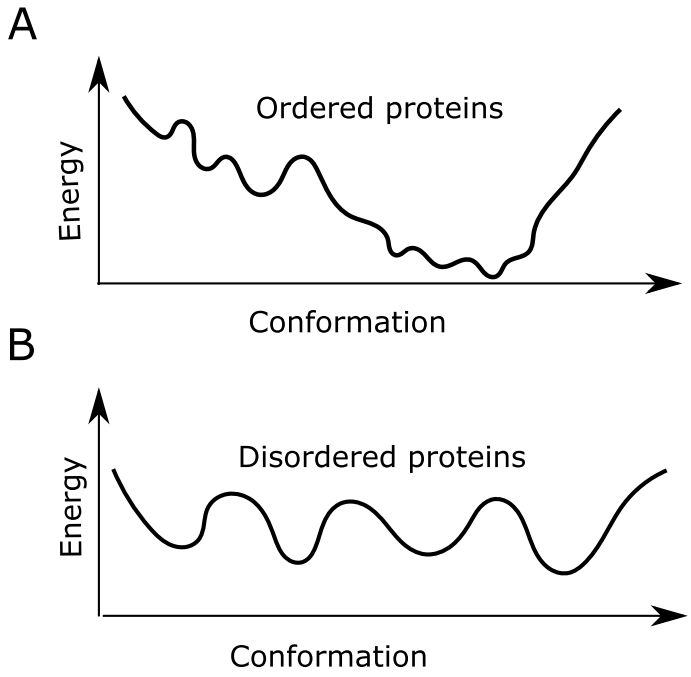


Figure 1.1: Energy landscapes of ordered and intrinsically disordered proteins. Simplified diagram illustrating the folding energy landscapes of (a) a typical globular protein and (b) a typical natively unfolded protein. Adapted from (Uversky et al., 2008).

In the recent years, considerable efforts has been devoted to the prediction of protein-protein interactions (PPIs) from amino acid sequences of proteins using computational methods. Proteins function through interactions with other proteins, making the study of how proteins interact with each other a crucial step toward uncovering the functions of proteins (Chowdhury et al., 2023; Qi et al., 2011). While established methods addressed PPI prediction (Casadio et al., 2022; Dunham et al., 2021), IDP-specific interactions remain understudied by the computational PPI prediction tools. Most of the sequence-based PPI prediction algorithms do not make a distinction between IDPs and structured proteins. On the other hand, structure-based PPI prediction algorithms base the interactions on the structures in the Protein Data Bank (Berman et al., 2000) which might not be suitable for the IDPs. Therefore, there is a need for computational approaches that links sequence information of the IDPs to their protein interactions. Leveraging diverse sequence-derived features of IDRs and adapting appropriate machine learning techniques hold promise for predicting protein-protein interactions of IDPs from their sequences.

Thesis outline

This thesis is structured as follows. Chapter 2, gives an overview of biological fundamental concepts on protein synthesis and delves into the concept of unstructural biology. We begin by introducing the concept of protein disorder. Subsequently, we explore various aspects of protein disorder and proceed to explain the PPIs. Following the introductory section, we move on with developing statistical methods that aid in identifying features and analyzing the attributes of disordered regions linked to their functionalities. In Chapter 3, we present our novel statistical method to identify aromatic periodic blocks in the human proteome. These blocks are closely associated with the phase separation behavior of disordered proteins. We demonstrate the application of this statistical method to the human proteome, with a specific focus on transcription factors (TFs), in order to identify these periodic regions. Following with Chapter 4, we present our novel statistical method for analyzing sequence features in the disordered regions of TFs that bind together on DNA elements, directing transcriptional activity. After this chapter, our focus shifts to the development of machine learning models designed to predict protein interactions. In Chapter 5, we explain the machine learning background for developing machine learning models to predict PPIs. We explain methods, features, and test and training schemes used by existing models to develop sequence-based PPI prediction models. In Chapter 6, our novel machine learning model is presented for predicting PPIs using features extracted from disordered sequences. We demonstrate how disordered segments can be used to predict interactions of disordered proteins, with an emphasis on using appropriate machine learning tools specific to the protein-protein interaction problem at hand. Finally, in Chapter 7, we introduce a novel method designed to identify protein regions with favorable interactions in protein interactions of disordered proteins. For this, we statistically analyze protein interactions between disordered proteins to extract contact potentials, whose values would indicate the interaction affinity between different amino acid groups.

2 BIOLOGICAL BACKGROUND

2.1 Proteins

Proteins play crucial roles in nearly all biological processes (Stryer, 2000). They perform essential functions within living organisms, including acting as catalysts for nearly all chemical reactions, controlling gene activity, and contributing to cellular structure (Latchman, 2002; Masulli, 2008).

The process of making proteins is a very tightly regulated process known as protein synthesis. The sequence of events involved in protein synthesis begins with DNA, which serves as the repository of genetic information in every cell (Stryer, 2000). This flow from DNA to proteins, known as the central dogma, is a two-step process in all organisms called transcription and translation, as illustrated in Figure 2.1. The first step, transcription, involves converting DNA information into messenger RNA (mRNA) within the nucleus of eukaryotic cells. Translation, the second step, takes place on ribosomes. During this process, proteins are synthesized according to the instructions encoded in the mRNA, resulting in the creation of amino acid sequences. This sequence of amino acids is then linked together through peptide bonds, ultimately forming a polypeptide chain, which consists of a unique arrangement of the 20 types of amino acids. As a result, proteins are essentially built from these fundamental building blocks known as amino acids, organized in a linear sequence. Each amino acid's distinct side chain structure and chemical properties contribute to the protein's specific chemical and physical attributes. With a total of 20 amino acids, the potential number of different possible polypeptide chains with a length of n is 20^n, resulting in a wide variety of possible chains.

After the synthesis of a polypeptide, it needs to be converted into a functional protein through a process called protein folding (Cooper, 2000). Proteins can undergo the folding process by adopting three-dimensional structures dictated by the order of the amino acid sequences in the polypeptide chain. This process is achieved through interactions between the side chains of the constituent amino acids. Any polypeptide chain's final folded configuration, or conformation, is usually the one in which the free energy is minimized (Alberts et al., 2002). Traditionally, it has been assumed that proteins are

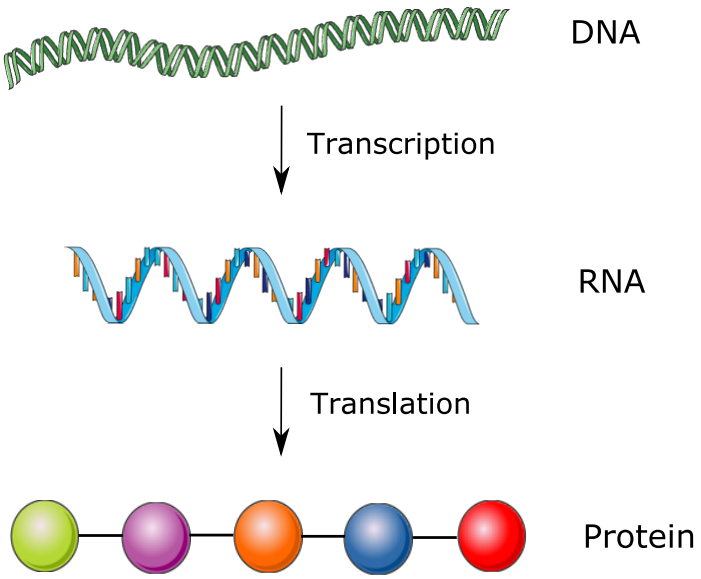


Figure 2.1: Central dogma of molecular biology. Central dogma of molecular biology dictates how the information stored in DNA gets used to make the proteins. First new RNA is made from DNA (transcription) From RNA protein chain is made (Original illustrations taken from [https://smart.servier.com/])

functional only when in a structured/folded state, giving rise to the structure-function paradigm. In the early years, the identification of a vast number of protein structures in the Protein Data Bank (PDB) (Berman et al., 2000), along with a detailed understanding of the roles of these structural proteins, including receptor signaling and transport, has further strengthened the acceptance of the structure-function paradigm (Ferreon et al., 2022; Trivedi et al., 2022). Undoubtedly, uncovering the structure of a protein provides valuable insights into its functional mechanism.

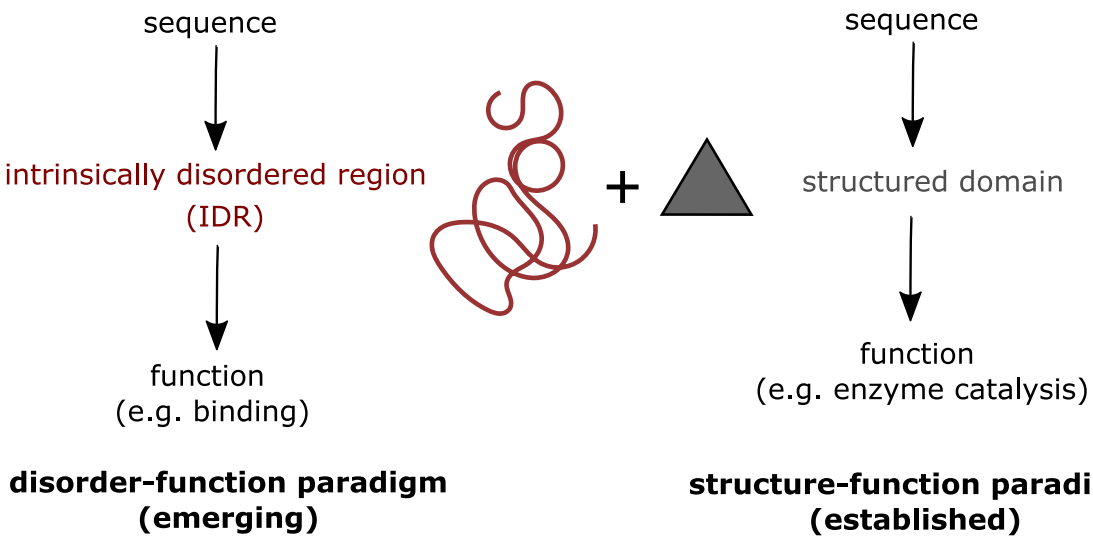


Figure 2.2: Structure-function and disorder-function paradigm. Despite lacking a stable secondary structure, IDPs have gained recognition for their numerous important functions which gave rise to the emergence of a new disorder-function paradigm. Adapted from (Babu, 2016).

2.2 Intrinsically disordered proteins

Contrary to the structure-function paradigm, a subset of proteins exists that do not fit into this model, yet remain functional. This has led to the proposition of the disorder–function paradigm (Figure 2.2). These proteins, referred to as IDPs, do not adopt a single well-defined three-dimensional structure when unbound and in solution (Fink, 2005). Instead, they exist as an ensemble of heterogeneous conformations (Tompa, 2011). In the past, IDRs were often perceived as passive segments within protein sequences, serving as "linkers" between structured domains. However, it is now widely recognized that IDRs actively engage in various protein functions (Lee et al., 2014). Despite lacking a stable secondary and tertiary structure, IDPs play a significant role in crucial cellular processes such as differentiation, transcription regulation, DNA compaction, and mRNA modification (Kosol et al., 2013).

The early 21st century introduced us more newly discovered disordered proteins (Burkart-Solyom, 2014). In the human proteome, 35 % of the total proteomic residues are in IDRs (Fukuchi et al., 2011). IDPs have unique sequence characteristics and certain types of amino acids that promote disorder are more prevalent in IDRs. These amino acids include proline, alanine, glycine, serine, glutamine, glutamic acid, lysine, and arginine (DeForte et al., 2016). IDRs also typically lack bulky hydrophobic amino acids, which would usually form a hydrophobic core in a structured domain (Lee et al., 2014). IDRs tend to have a high percentage of charged residues, and the presence of charged residues

within the IDRs leads to larger interaction surfaces. This attribute gives IDPs certain advantages in terms of interactions with their targets (Bigman et al., 2022; Morris et al., 2021).

2.2.1 Roles of disorder

During the last 20 years, much effort has been invested to understand the roles and functions of IDRs in depth. One of the important roles of IDPs is being the major players in PPIs, with IDRs serving as mediators of these interactions (Bondos et al., 2021; Chakrabarti et al., 2022; Chong et al., 2021; Uversky, 2020). The inherent absence of structure in IDPs and IDRs offers functional advantages that make them exceptionally well-suited for mediating various interaction modes. As a consequence of this conformational flexibility, IDPs can have a multitude of binding modes by acquiring different conformations based on the shape of the target protein. For example, a single IDP can fold upon binding to targets through a mechanism called disorder-to-order transition. A known example is the interaction of the KID domain of Cyclic AMP-responsive element-binding protein (CREB) and CREB-binding protein (CBP). KID domain is initially unstructured, but undergoes folding to create orthogonal helices upon binding to its target domain within CBP (Dyson et al., 2005).

While it has been proposed that one advantage of intrinsic disorder is the low-affinity, high-specificity interaction of IDPs with their targets, it has also been observed that many IDPs can bind with high affinity (Dogan et al., 2014). This dynamic binding mechanism has been proposed to enable rapid association and initiation of signaling processes, while facilitating easy dissociation once the task is accomplished (Lee et al., 2014). Two disordered proteins can also have a mutual folding upon interaction (Lindorff-Larsen et al., 2021). Another study by Borgia et al., 2018 showed two intrinsically disordered human proteins histone H1 and its nuclear chaperone prothymosin alpha interact with affinity and fully retain their structural disorder.

In recent years, the significance of disordered interaction modules encoded in the IDR sequences has become appearent. Different functional interaction modules are embedded in the IDR sequences of proteins to facilitate their interactions. These include short linear motifs (SLiMs) and molecular recognition features (MoRFs), which contribute to selective protein interactions (Lee et al., 2014). SLiMs are typically short, consisting of up to eight amino acid residues, and often display evolutionary conservation. MoRFs, on the other hand, are longer and can be 10 to 70 amino acids. MoRFs are the subregion of disordered regions and are defined as capable of undergoing a disorder-to-order transition when

binding to their partner. MoRFs may even contain SLiMs within them (Mooney et al., 2012). Additionally, Post-translational modifications (PTMs) within IDRs are frequently observed and the have been shown to regulate protein interactions by modulating the energy landscape of IDRs (Bah et al., 2016).

2.2.2 Disorder in transcription factors

Transcription factors (TFs) are proteins that bind to specific DNA sequences to regulate tissue-specific gene expression. They play a crucial role in controlling which genes are expressed and when, and they are involved in a wide range of cellular processes, including development, differentiation, and response to environmental stimuli (Weidemüller et al., 2021).

In addition to the role of IDRs in mediating protein interactions, studies have shown that IDRs in TFs can also play a role in the regulation of transcriptional processes (Lyon et al., 2021; Sabari et al., 2020). TFs have two distinct types of domains: the DNA binding domain (DBD), responsible for binding to DNA, and the activation domain (AD), responsible for facilitating the recruitment of the transcriptional machinery to gene promoter regions (Latchman, 2002). ADs, unlike structured DBDs, are often enriched in terms of disorder (Sanborn et al., 2021). An inherent lack of structure enables ADs to engage with various coactivators, thereby promoting the activation of gene expression (Scholes et al., 2016). Recently, it has been proposed that IDRs of TFs are involved in transcriptional activity by forming phase-separated condensates that regulate transcription within cells (Boija et al., 2018; Hnisz et al., 2017).

Liquid-liquid phase separation (LLPS) is a physicochemical phenomenon where molecules segregate into a dense phase and a less dense phase (Hyman et al., 2014). This interesting behavior has been observed in IDRs of TFs that interact with co-activators, forming phase-separated condensates (Boija et al., 2018; Hnisz et al., 2017). These biomolecular condensates, forming through phase separation, provide a means to compartmentalize in cellular environments without requiring a membrane (Banani et al., 2017). They dynamically assemble and are thought to facilitate cooperative transcriptional regulation (Hnisz et al., 2017). Several studies demonstrated that IDR-IDR interactions can drive this condensation behavior (Chong et al., 2018; Sabari et al., 2018).

Several additional studies have demonstrated the roles of IDRs on transcriptional regulation. Indeed, another study by Ma et al., 2021 reported an IDR-based interaction profile where the interactions between TF ADs and coactivator p300 IDRs drive condensation. Unfortunately, the knowledge of which sequence features of IDRs that promote phase

separation and underlie the transcriptional activity of TFs is very limited. In another study, Barkai et al. (2020) showed that IDRs of TFs can contribute to in vivo binding specificity by directing them to the enhancer elements. Altogether, these findings demonstrate the roles of IDRs in TFs, especially in mediating the formation of phase-separated condensates and contributing to the regulation of transcriptional processes.

2.2.3 IDPs in diseases

IDPs are associated with many diseases such as cancer, amyloidosis, diabetes, cardiovascular, and neurodegenerative diseases (Martinelli et al., 2019). One of the most well-known causes is the aggregation of disordered proteins which can result in neurological diseases (Ayyadevara et al., 2022; Breydo et al., 2011, 2012; Tsoi et al., 2023). Some of the known examples of IDPs that are associated with diseases include alpha-synuclein, tau protein, p53, and BRCA1 (Uversky et al., 2008). The exact mechanisms leading to this pathological state are still an ongoing research question.

Mutations occurring in the IDRs are recognized as one of the established factors contributing to the onset of diseases. Mutations of IDPs can affect the normal function of proteins, leading to misidentification and missignaling. IDR sequences frequently serve as loci for variants associated with disease. It has been found that more than 20% of human disease mutations occur in IDRs (Vacic et al., 2012). Recent research has demonstrated the impact of de novo frameshift variants within the IDR region of HMGB1, resulting in polydactyly and tibial aplasia syndrome, a rare and intricate malformation syndrome (Mensah et al., 2023). The authors have identified more than 600 frameshift mutations in IDR sequences. Another study conducted by (Wong et al., 2020) demonstrated a strong enrichment of missense mutations at the interface core of interacting IDRs which suggests that alterations in the interactions between IDRs can have a significant impact on protein function and cellular signaling pathways.

However, the exact mechanisms through which such variations induce diseases remain an ongoing challenge. One known mechanism for how mutations on IDRs cause diseases is the disruption or creation of motifs within the IDR sequences (Meyer et al., 2018). Additionally, it has been observed that mutant IDRs can interact with distinct partners, leading to altered interactions. For instance, recent analysis has shown that mutations located on IDR of GLUT1 can generate dileucine motifs, resulting in mislocalization of the mutated protein and causing GLUT1 deficiency syndrome (Meyer et al., 2018). Several PTMs can also lead to the formation of complex aggregates such as plaques and tangles, which in turn can lead to neurological diseases (Oldfield et al., 2014). Additionally, a

number of diseases have been associated with misregulation or formation of condensates that are governed by IDPs (Boija et al., 2021).

2.3 Protein-Protein Interactions

Proteins have evolved to interact with each other to perform certain functions, that are common to all living organisms. Which proteins interact with each other to orchestrate cellular function is a key step to improve the system-level understanding of molecular processes in cells (Levy et al., 2008). Therefore, understanding how proteins interact and map PPIs is currently an important area of both computational and experimental research (Li et al., 2020; Uversky, 2013). Several high-throughput experimental approaches are available to identify binary protein interactions. The most common experimental techniques are the yeast two-hybrid screens (Y2H) and mass spectrometric protein complex identification (MS-PCI).

YH2 is a system that is based on the interaction between AD and the DBD of GAL4 TF. If two proteins of interest interact, the interaction between these AD and DBD domains can lead to the transcriptional activation of a reporter gene in the system. To set up the YH2 system, two yeast hybrids are prepared, known as the 'bait-prey' system. In this setup, the bait protein is fused to the DBD, and the prey protein is fused to the AD. When these fused domains interact, they reconstitute the original transcription factor. If the proteins of interest interact, it results in the expression of the reporter gene (Keegan et al., 1986) (Figure 2.3).

PPIs can be classified based on different criteria, with the most common classification being between transient and stable interactions (Nooren et al., 2003). Transient interactions refer to unstable protein interactions, characterized by their ability to form and break easily (Perkins et al., 2010). Permanent PPIs on the other hand result in the formation of a stable protein complex. Both transient and permanent PPIs play crucial roles in cellular functions (Ghadie et al., 2022). Unfortunately, weak interactions are not detectable by standard methods (Nahlé et al., 2022). Unfortunately, standard affinity purification methods cannot detect weak interactions (Nahlé et al., 2022). The proximity-dependent biotin identification (BioID) approach has emerged as a promising solution to overcome this limitation. In living cells, BioID enables the detection of proteins nearby a target protein (Varnaitė et al., 2016). Future research can benefit greatly from this approach since it enables the observation of both direct and indirect PPIs.

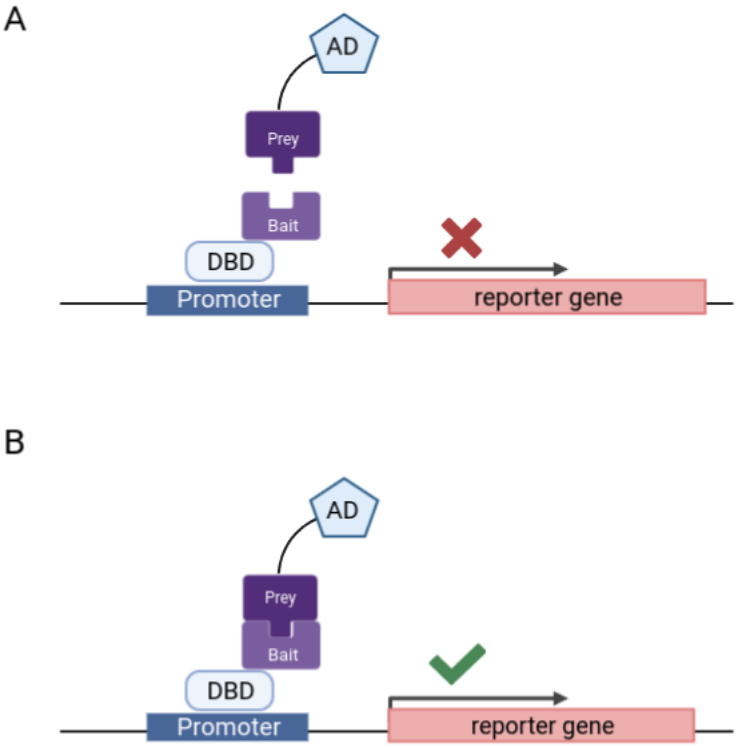


Figure 2.3: Yeast two-hybrid system. Yeast two-hybrid system. If two proteins interact, it leads to the involvement of RNA polymerase II and subsequent transcription of a reporter gene. Created with BioRender.com

A large amount of experimental data on PPIs has been generated by numerous studies. The availability of PPI data makes it easier for scientists to investigate protein interactions, understanding biological functions, and identify potential therapeutic targets. Nevertheless, these datasets may contain biases and false positives, which could have an important effect on the biological theories that are drawn from them (Grigoriev, 2003; Venkatesan et al., 2009). Many databases have been established to collect and curate PPI datasets in order to address this issue (Alanis-Lobato et al., 2017; Licata et al., 2012; Oughtred et al., 2021). These PPI databases help in organizing data from different sources and make it easier for researchers to access the PPI data.

One such database is Human Integrated Protein-Protein Interaction rEference (HIPPIE) database (Alanis-Lobato et al., 2017), which includes experimentally detected PPIs from IntAct (Toro et al., 2022), MINT (Licata et al., 2012), BioGRID (Oughtred et al., 2021), HPRD (Peri et al., 2004), DIP (Xenarios et al., 2000) , BIND (Bader et al., 2001), and MIPS (Mewes et al., 2002). HIPPIE assigns a confidence score to each PPI based on the quality

and reliability of the underlying evidence. This scoring system enables users to prioritize interactions based on their confidence levels, helping researchers focus on reliable PPI subnetworks. In their latest version, the database also has the tissue-specific PPI networks derived gene expression data from 53 healthy human tissues from GTEx (Ardlie et al., 2015).

3 IDENTIFICATION OF PERIODIC BLOCKS IN HUMAN TFS

In this chapter, we present a novel method to identify protein regions with significant, albeit not necessarily perfect periodicity in the occurrence of aromatic residues, independent of sequence length and composition. We achieve this by modeling the occurrence of aromatic residues by the Poisson process. In the following sections, we will first explain the background of aromatic periodic blocks and their link to IDRs. Then, we will introduce the Poisson distribution and explain how we can use this distribution to model the occurrence of aromatic residues. Next, we will illustrate the application of this method in the analysis of the human proteome, where we identify aromatic periodic blocks in protein sequences and explore the locations of aromatic blocks in the human proteome.

3.1 Background

Understanding the sequence characteristics of IDRs that are responsible for their wide range of behaviors remains an ongoing research question. As mentioned before in Section 2.2.2, many TFs ADs are intrinsically disordered (Baughman et al., 2022) and TF IDRs can form phase-separated condensates to faciliate transcriptional regulation (Boija et al., 2018; Hnisz et al., 2017). Additionally, TF ADs are typically characterized as the abundance of specific amino acids such acidic ADs, proline-rich or glutamine-rich (Sanborn et al., 2021). One of the sequence characteristics of TFs that is known to have functional importance is aromatic residues. A recent study by Erijman et al. (2020) trained a deep neural network on both functional and non-functional AD sequences across all TFs in budding yeast and showed that aromatic residues are highly enriched in functioning ADs. Nevertheless, sequence features of TF IDRs remain understudied through computational approaches due to limited number of studies. Therefore, there is a need to develop computational methods for identifying the specific IDR sequence features that drive this phase separation phenomenon and, in turn, regulate transcription.

There may be similarities between the sequence features involved in phase separation and transactivation of TFs and the sequence features that underlie phase separation of prion-like domains (PLDs) of RNA binding proteins. PLDs of RNA-binding proteins are

low-sequence complexity IDRs. They have become convenient domain types for decoding the underlying mechanisms of this phase-separation behavior (Holehouse et al., 2021). A recent study by Martin et al. (2020) developed a model called the “stickers and spacers model” to quantitatively understand the phase separation behavior of intrinsically disordered PLDs based on the sequence (Figure 3.1). According to the model, one can categorize parts of macromolecules as either "stickers" (regions that drive inter-molecular interactions) or "spacers" (regions that don’t). Stickers are the regions that drive inter-molecular interactions, spacers are the regions that do not contribute the interactions. Stickers could be individual residues, short linear motifs, folded domains, specific structural features in RNA, or post-translationally modified residues. They showed that uniform patterning of aromatic residues is a sequence feature that promotes LLPS. In this scenario, they found aromatic residues as stickers. They identified uniformly distributed aromatic residues along several PLDs of protein sequences.

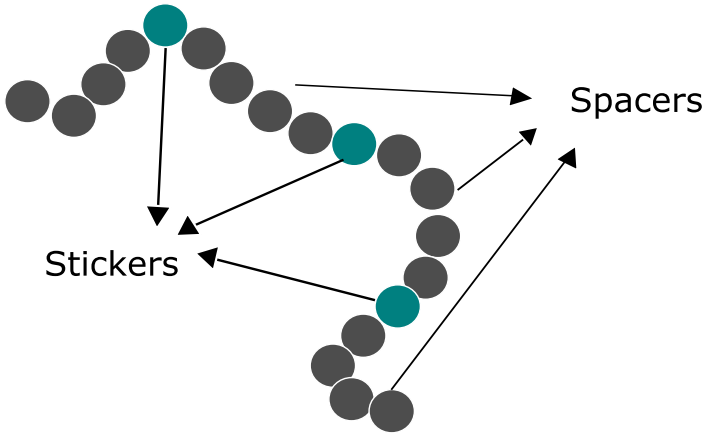


Figure 3.1: Schematic representation of spacer model. Adapted from (Martin et al., 2020).

To quantitatively measure the periodicity in protein sequences, Martin and colleagues, 2020 conducted a statistical analysis. They introduced a parameter called the ’patterning parameter’ (W_{Aro}), which assesses the probability of evenly spaced aromatic residue occurrences in a protein sequence compared to random chance.

To model the random distribution, they randomly shuffled the given sequence 10^5 times to generate random sequences with the same amino acid composition as the input sequence. Subsequently, they compared the computed W_{Aro} score to the distribution of W_{Aro} scores from these randomly generated sequences by looking at where the observed distribution of spacer lengths between aromatic residues fell on this range of random distributions. In summary, the W_{Aro} score quantifies the likelihood of the observed pattern of aromatic residues occurring by chance in their study.

As an illustrative example, in one of the PLD domains they studied, they found that aromatic residues in this PLD sequence were more uniformly spaced than 99.99% of randomly generated sequences. This finding demonstrates that the distribution of aromatic residues (or 'stickers') along the PLD sequence they studied is not random. Additionally, their study showed that the NFAT5 TF has more uniformly spaced aromatic residues in its IDR region compared to randomly generated sequences (W_{Aro} = 0.124, empirical p-value = 0)

While the method is effective, it has several drawbacks. First, it requires a substantial amount of computational time to generate expected distribution by shuffling the protein sequences 100,000 times. Secondly, the method lacks flexibility in allowing users to experiment with different interspacing lengths between stickers. This is important because protein sequences might have different spacer lengths. Finally, this study was limited to a small number of proteins. As a result, it remains uncertain whether specific protein families contain a higher number of proteins with periodically arranged aromatic residues. We were particularly interested in determining whether TFs are enriched in aromatic periodic blocks.

We developed a novel method to find those periodically arranged aromatic residues in human proteome. Our approach is efficient and straightforward. It requires only about 15 minutes to run our method over the entire human proteome, which includes more than 16,000 proteins. Overall, our approach enables us to investigate the presence of periodic aromatic blocks in the entire human proteome gaining insights into the enrichment of those periodic blocks in specific protein families of interest. Additionally, we can explore the locations of these periodic blocks within sequences and identify periodic regions that overlap with IDRs and investigate the potential link between IDRs and periodic blocks.

3.2 Methods

3.2.1 Modelling the interarrival times via Poisson distribution

The Poisson distribution is a discrete probability distribution that models the probability of a given number of events occurring in a fixed interval of time, assuming that the events occur independently and the probability of an event occurring in a given length of time does not change over time (Sinharay, 2010). Then X, the number of occurrences of a particular event in a fixed unit of time and with a constant rate λ, follows the Poisson distribution.

The probability of observing exactly k occurrences in a given interval, denoted as P(X = k), is determined by the following probability mass function:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where λ is the mean of the Poisson distribution, and it is also equal to the mean (μ) of the Poisson distribution.

In a Poisson process, the arrival times of events are considered as random variables (McCann, 2016). Specifically, the time between two consecutive events denoted as T_n (where *n* refers to the event number), is referred to as the interarrival time. For instance, the waiting time for the first event is denoted as T_1.

Let's say S_n is the total time until the occurrence of n^{th} event, then S_n is the sum of all the interarrival times up to n^{th} event:

$$S_n = T_1 + T_2 + \ldots + T_n$$

By considering the equation, we can derive an expression for the interval time, denoted as T_n, which is obtained by subtracting the previous sum:

$$T_n = S_n - S_{n-1}, \quad n \geq 1$$

Interarrival times in a Poisson process follow a geometric distribution. The geometric distribution describes the probability of waiting for the first success in a sequence of independent trials, where each trial has the same probability of success. The probability mass function (PMF) of the geometric distribution can be expressed as:

$$P(t) = (1-p)^{t-1} \cdot p$$

Here, t represents a specific interarrival time, while p represents the probability of success in a single trial, calculated using the mean of the Poisson distribution.

Assume that we're given a dataset containing n observations characterized by an unknown distribution P. Our objective is to determine whether this dataset follows a geometric distribution P_0 and choose between the following hypotheses H_0 and H_1 :

$$H_0 : P = P_0, \quad H_1 : P \neq P_0$$

In our context, where the focus is on the geometric distribution, the null hypothesis assumes that there is no significant difference between the observed dataset and the

expected geometric distribution. For such investigations, goodness-of-fit tests are commonly employed to test whether a given dataset comes from a certain distribution (Dodge, 2008). One widely recognized statistical test for this purpose is the Kolmogorov–Smirnov test (K-S) test. The K-S test is a nonparametric goodness-of-fit test that compares how the distribution of the observed sample, F_{data} aligns with a chosen theoretical distribution, F_0 (Dodge, 2008; Kendall et al., 2008; Massey, 1951).

In order the compare the empirical distribution function of the observed data F_{data} to the cumulative distribution of a hypothesized distribution F_0 (Bogacka, 2004; Massey, 1951; Stephens, 1992), the K-S test computes the maximum absolute distance between these two distributions, denoted as D_n. This measure, known as the test statistic of K-S test, can be formulated as follows:

$$D_n = \sup_x |F_0(x) - F_{\text{data}}(x)|$$

If this distance (D_n) is greater than a critical value in the K-S distribution based on given significance level (Ramachandran et al., 2014), the null hypothesis is rejected. As D_n approaches zero, it suggests that the two datasets could be drawn from the same distribution. The *p-value* is derived from the calculated distance and if it falls below the chosen significance level, the null hypothesis is rejected. Since F_0 in our case is the geometric distribution, it implies that F_{data} does not follow the geometric distribution.

3.2.2 Application to human proteome

If aromatic residues in a given protein sequence appear to occur at a certain rate, following a completely random pattern, then the occurrence of aromatic residues follows the Poisson distribution ($X \sim \text{Poisson}(\mu)$), where X represents the count of aromatic residue occurrences and μ represents the mean occurrence rate of aromatic residues in a given interval.

Since we aim to analyze the spacing between aromatic residues, we consider the presence of aromatic residues as the event of interest. Our approach involves modeling the "interarrival times," which refer to the number of residues between the instances of aromatic residues. Let T denote the number of amino acids before an aromatic residue occurs. By using the notation $P(T = t)$, we can quantify the probability of observing a specific interarrival time within the geometrically distributed interarrival times. To achieve this, our first step is to estimate the expected distribution of adjacent aromatic residues, also known as "spacer length". To model this expected geometric distribution,

we extrapolated the mean from the proportion of aromatic residues, which is the ratio of aromatic residues to the sequence length.

For the observed distribution of spacer lengths, we counted the number of residues between adjacent aromatic residues for the given protein sequence. Subsequently, we employed the K-S test to compare the observed distribution of spacer lengths with the expected geometric distribution. This comparison allows us to determine if the observed distribution aligns with the expected distribution. If the p-value is less than the chosen significance level, it indicates that the time intervals between occurrences of aromatic residues do not follow the geometric distribution. In simpler terms, the occurrences of certain amino acids are not distributed randomly as in a Poisson distribution. Therefore, the smaller the p-value, the less likely these aromatic residues occur randomly.

Next, our objective was to identify complete regions containing consecutive periodic blocks within each protein. To accomplish this, we performed K-S test for every 100-amino-acid segment within each protein using a sliding window approach. To identify complete periodic regions, we plotted p-values for every 100-amino acid segment against the position of each window and identified consecutive instances where the p-values consistently dropped below a specific threshold (=0.5 * average(p-value)), thus defining periodic regions. This process is depicted in Figure 3.2. The algorithm behind the method can be summarised as follows:

Algorithm 1: Quantification of periodicity in a given sequence

aromatic_residues = ["F", "Y", "W"]
encoded_sequence =
[1 if residue in aromatic_residues else 0 for residue in sequence]
for residue = 1 to (length(encoded_sequence) − windowsize + 1) **do**
  window_sequence = encoded_sequence[residue : (residue + windowsize − 1)]
  interarrivals = spacer lengths between aromatic residues in window_sequence
  $\lambda = \frac{\text{count(window\_sequence==aromatic\_residues)}}{\text{windowsize}}$
  perform the K-S test for the geometric distribution using the interarrivals and λ.
  return the p-value of the K-S test statistic
end for

The method was applied to every protein in the human proteome. The entire human proteome was taken from the GRCh38.p13 assembly. We extracted the protein sequences categorized as 'Ensembl canonical', which represent the most conserved and highly expressed sequences. In cases where genes lack "Ensembl canonical", we took the longest

"Genecode basic" isoform. To perform the sliding window analysis, we filtered out the sequences shorter than 100 amino acids.

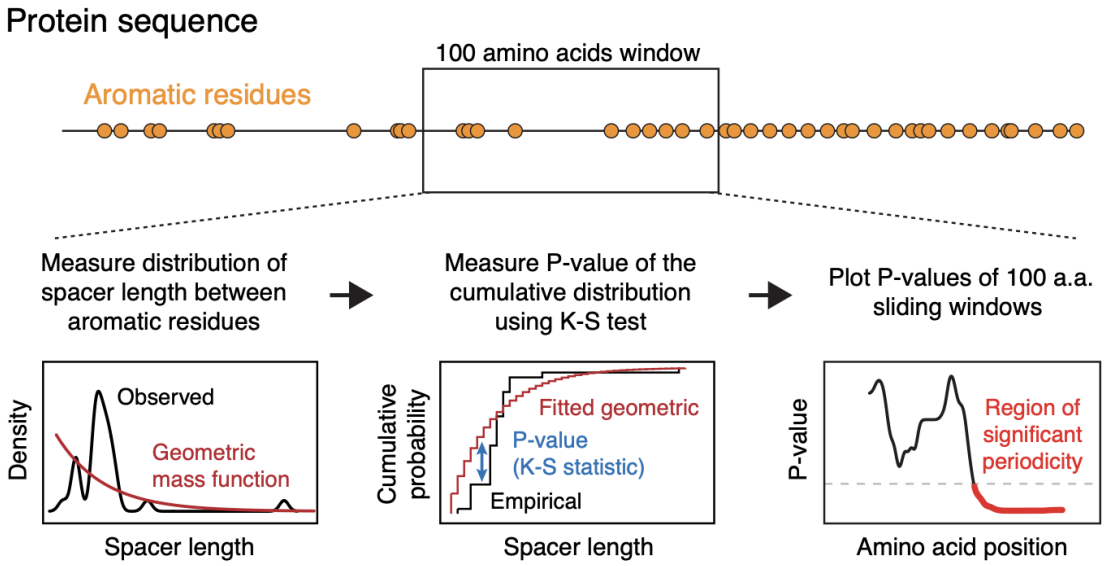


Figure 3.2: Schematic of the pipeline for the identification of regions with significant periodicity in the human proteome. For every protein sequence the distribution of spacer lengths between aromatic residues is measured. A Kolmogorov-Smirnov (K-S) test is used to measure the P-value of the cumulative distribution. The P-value is then plotted for each 100 amino acid window using a sliding window approach. Regions with significant periodicity are defined by using a P-value cutoff calculated as half of the average P-value for the entire protein sequence.

After identifying the periodic regions for each protein, we applied a minimum p-value threshold of 0.05 to identify regions with significant periodicity. Following this step, we conducted an enrichment analysis to identify the TFs and PLDs within the subset of proteins that showed significant periodicity in our dataset.

Lastly, we delved into the evolutionary aspect of periodicity by analyzing one periodic block from one of the TFs in our dataset. The aim here is to understand whether periodic regions in proteins are evolutionarily conserved. The results of these analyses are presented in the next section.

3.3 Results

3.3.1 IDR-periodicity relationship and functional annotations

After calculating the periodicity scores as p-values and identifying periodic regions across the entire human proteome, we compiled a table including the start and end positions of the periodic blocks in proteins, along with the corresponding p-value for the identified periodic region. The p-value assigned to the periodic region is determined by the minimum p-value for that entire block. We also predicted the IDRs of each protein using the disorder prediction tool Metapredict (Emenecker et al., 2021).

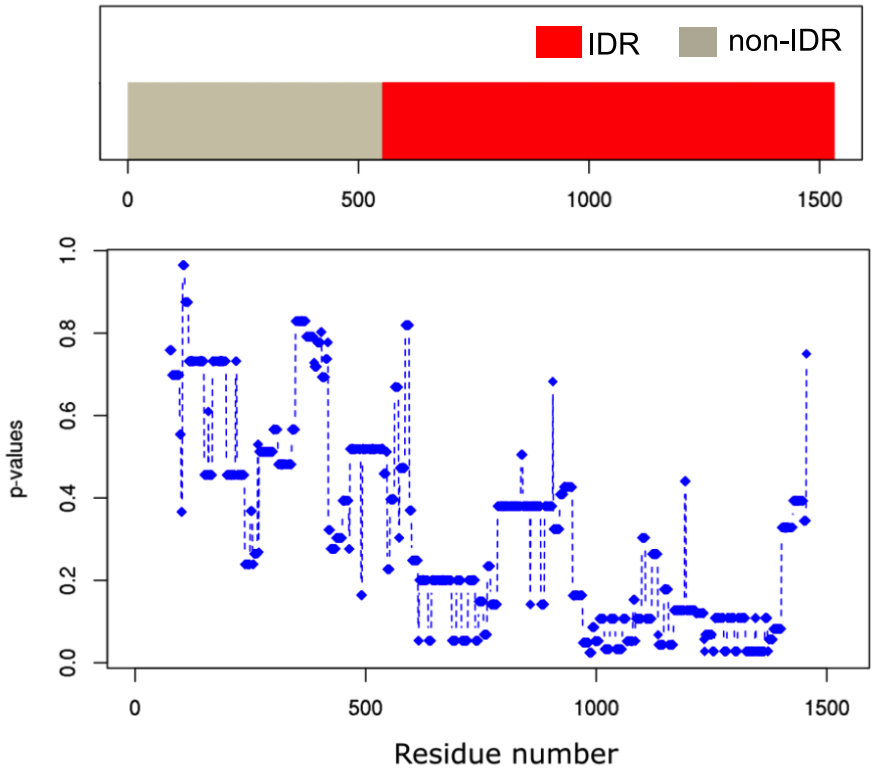


Figure 3.3: Analysis of NFAT5 sequence using our method. P-values were calculated across the NFAT5 sequence using a sliding-window approach with a window size of 150.

Firstly, we investigated proteins with known periodic IDR sequences. As mentioned previously in Section 3.1, Holehouse et al., (2020) demonstrated that the IDR sequence of NFAT5, a TF, has a high W_{Aro} score, indicating a uniform spacing between aromatic residues along the IDR sequence. We set out to determine whether our method can identify

the same IDR region. Applying our method, we plotted the p-values for each window along NFAT5's sequence (Figure 3.3). Using the sliding window approach, we were able to demonstrate a significant decline in p-values as the sliding window approaches NFAT5's IDR region (Figure 3.3). Our analysis identified the periodic block, positioned between residues 1318-1469, which aligns well with the predicted IDR region. Additionally, when we look at the actual sequence of the identified periodic block (Figure 3.4), we observe the periodic arrangement in the identified protein sequence.

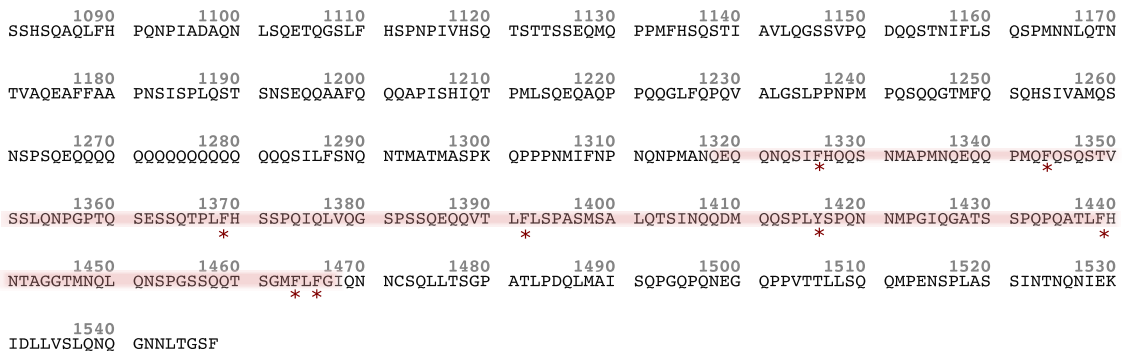


Figure 3.4: Periodic block of NFAT5. The identified periodic block from our method is indicated by the highlighted red sequence, while the asterisks represent aromatic residues. Our method effectively detects the periodic block in the NFAT5 sequence which overlaps with the IDR region.

Next, we ask the question if predicted IDR regions of proteins generally overlap with periodic blocks that we identified by our approach. Regions with significant periodicity were defined by the min p-value cutoff of 0.05. The criterion for overlap was the presence of at least one shared amino acid between the two regions. Among the 3496 significant proteins with predicted IDR regions, a total of 1665 have such overlap with the IDR regions. Next, we focused on TFs using ~1,500 previously curated human TFs (Lambert et al., 2018). 665 TFs have significant periodic regions and 360 of these TFs contained regions that overlapped with the annotated IDR regions. Remarkably, leveraging only a single sequence feature enabled us to predict over 50 percent of TF IDR regions. As a result, we have termed these regions "quasi-IDR regions". Table 3.1 shows the top 20 proteins with the highest degrees of periodicity with lowest p-values, alongside their corresponding IDR regions, whereas Table 3.2 shows the same for the TFs.

Finally, we aimed to investigate whether specific protein families including TFs show distinct enrichment patterns of periodic regions. We then categorized the proteins by annotating RNA-binding proteins, prion-like proteins, and TFs. Additionally, PLDs were identified using default settings of PLAAC (Lancaster et al., 2014). Aromatic-rich prion-

Table 3.1: The top 20 proteins with the highest levels of periodicity, along with their predicted IDR regions

Rank	Gene Symbol	Periodic Region	Min(p-value)	Overlap	IDR region(s)
1	DAZ2	117-558	1.22659E-09	yes	128-558
2	DAZ1	442-744	1.22659E-09	yes	131-195, 459-744
3	DAZ4	279-723	1.22659E-09	yes	135-195, 293-723
4	DAZ3	117-438	7.42462E-09	yes	132-387
5	GRINA	1-213	1.78098E-08	yes	1-149
6	MYO15A	146-471	3.30347E-08	no	2303-2670, 2955-3042
7	MAGED1	251-509	8.36898E-08	yes	1-465
8	POLR2A	1501-1970	7.32785E-07	yes	1485-1970
9	LMBRD2	276-559	2.09032E-06	no	581-695
10	GPR137B	1-199	2.09032E-06	no	319-399
11	SCAMP1	92-309	2.09032E-06	no	1-71
12	KRT10	1-206	2.84549E-06	yes	1-149, 455-584
13	ERICH6B	4-238	4.87161E-06	yes	1-344
14	NDFIP1	71-221	7.27159E-06	yes	1-105
15	SCAMP2	103-302	7.27159E-06	no	1-80
16	HNRNPUL2	607-747	7.27159E-06	yes	1-248, 606-747
17	CDR1	92-262	7.27159E-06	yes	1-179
18	RHBDF1	632-855	8.91194E-06	no	1-378, 496-586
19	ABCA8	990-1264	8.91194E-06	yes	1224-1277
20	EIF4B	138-380	1.56957E-05	yes	1-99, 162-611

Table 3.2: The top 20 TFs with the highest levels of periodicity, along with their predicted IDR regions

Rank	Gene Symbol	Periodic Region	Min(p-value)	Overlap	IDR region(s)
28	SON	822-1161	3.95461×10^{-05}	Yes	1-2227
64	EGR1	383-543	0.000241571	Yes	1-340, 421-543
66	ZNF768	1-213	0.000241571	Yes	1-266
68	ZNF606	275-480	0.000243978	Yes	1-70, 114-406
85	ZNF460	265-528	0.000431782	No	1-191
102	ZNF326	1-235	0.000587999	Yes	1-304
129	ZNF717	380-676	0.00090683	Yes	116-305, 663-752
135	ZNF182	172-379	0.000965554	Yes	82-191
138	ZNF83	218-419	0.000965554	No	1-107
191	ZXDA	290-506	0.001362425	No	1-259, 577-799
192	ZXDB	294-510	0.001362425	No	1-266, 581-803
195	ZNF566	72-218	0.001362425	Yes	1-189
216	RBAK	561-714	0.001841799	No	1-258
219	ZEB2	824-1045	0.001841799	Yes	1-211, 315-652, 698-1002, 1092-1214
262	ZNF132	360-651	0.002331092	No	74-285
263	ZSCAN21	242-468	0.002331092	Yes	125-275
264	ZNF57	1-173	0.002331092	Yes	1-139
265	MYNN	268-542	0.002331092	Yes	120-302
267	ZNF845	452-659	0.002331092	No	1-245

like domains were then defined from PLDs set as those that contain 10% or more aromatic content. Figure 3.5 shows the distribution of the p-values of periodic blocks in PLDs, aromatic-rich prion-like domains, and TFs. The density plots showing the p-values of the periodic blocks revealed that RNA-binding proteins and PLDs have more periodic regions compared to TFs (Figure 3.5A). Notably, our approach also identified the previously reported periodic region in HNRNPA1 from the spacer and sticker model (Martin et al., 2020), (Figure 3.5B).

Furthermore, we performed both a Gene Set Enrichment Analysis (GSEA) and a Gene Ontology (GO) analysis (Subramanian et al., 2005) for the same protein families. In this analysis, proteins were ranked based on their p-values, with lower p-values giving higher ranks. The GSEA analysis revealed that proteins containing regions of significant periodicity are enriched in RNA-binding proteins and prion-like proteins. Interestingly, this enrichment pattern is not observed in TFs to the same extent (Figure 3.6).

For the GO analysis, we focused on the top 1000 proteins with the lowest p-values. In this analysis, we examined the overrepresentation of GO terms in proteins that overlap with IDRs and those that do not (Figure 3.7). The results of the GO enrichment analysis indicate a strong association of the most significant proteins with RNA binding functions (Figure 3.7).

In conclusion, our analysis sheds light on the interplay between IDRs, periodic regions, and specific protein functions. Interestingly, our method was able to capture IDR regions in many TFs based solely on aromatic periodic regions. However, our analysis also suggests that periodic regions are not as enriched in TFs as they are in prion-like domains. Further work is required to fully understand the functional implications of periodicity in TFs. Overall, these findings highlight the strong connection between these periodic blocks and IDR sequences, contributing to a better understanding of the functional implications of periodic blocks in protein families, including TFs.

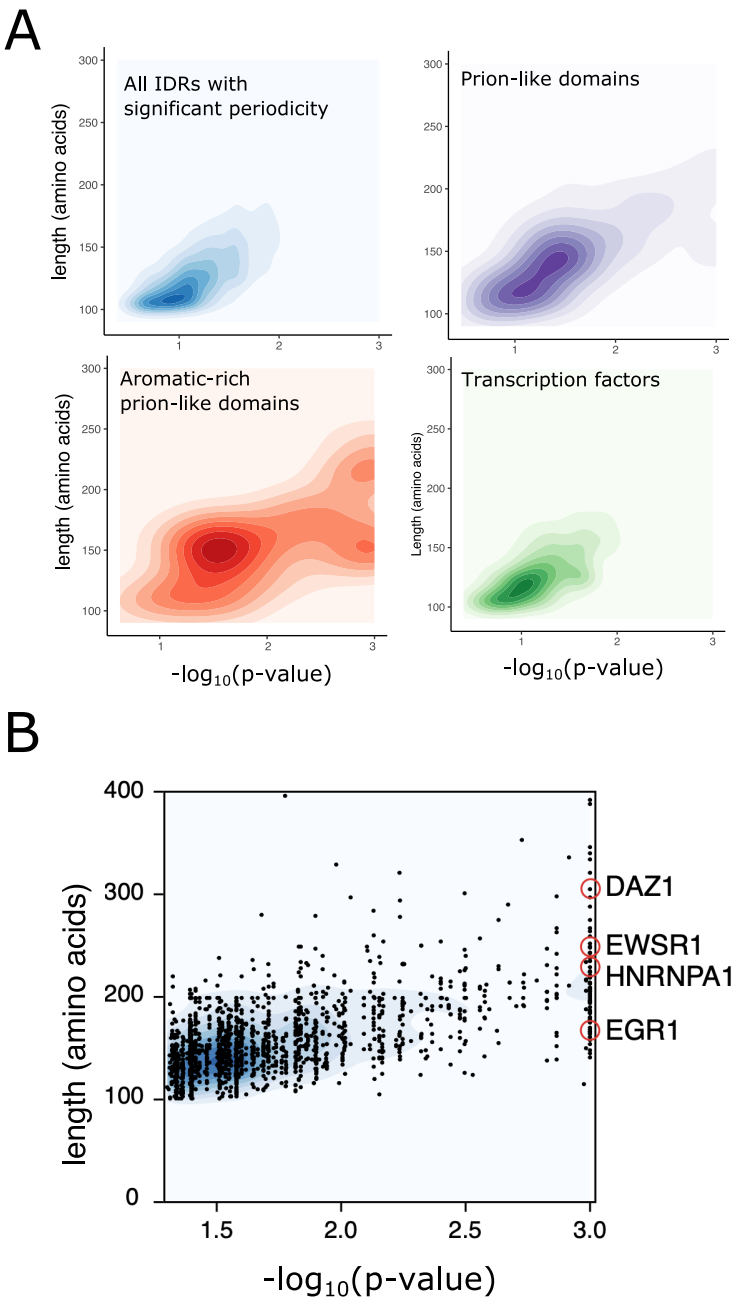


Figure 3.5: Density plot of proteins. A) Top left: All proteins Top right: Prion-like domains Bottom left: Aromatic rich prion-like domains (>10% aromatic content). Bottom right: transcription factors. For each region, the length of the region is plotted against the lowest P-value within the periodic region. The depth of the color of the cloud is proportional to the density of the dots in the area. B) Density plots for all the proteins with significant periodicity (p-value < 0.05). Each black dot represents one region, and the depth of the color of the cloud is proportional to the density of the dots in the area. The positions of the DAZ1, EWSR1, HNRNPA1, and EGR1 are highlighted with red circles.

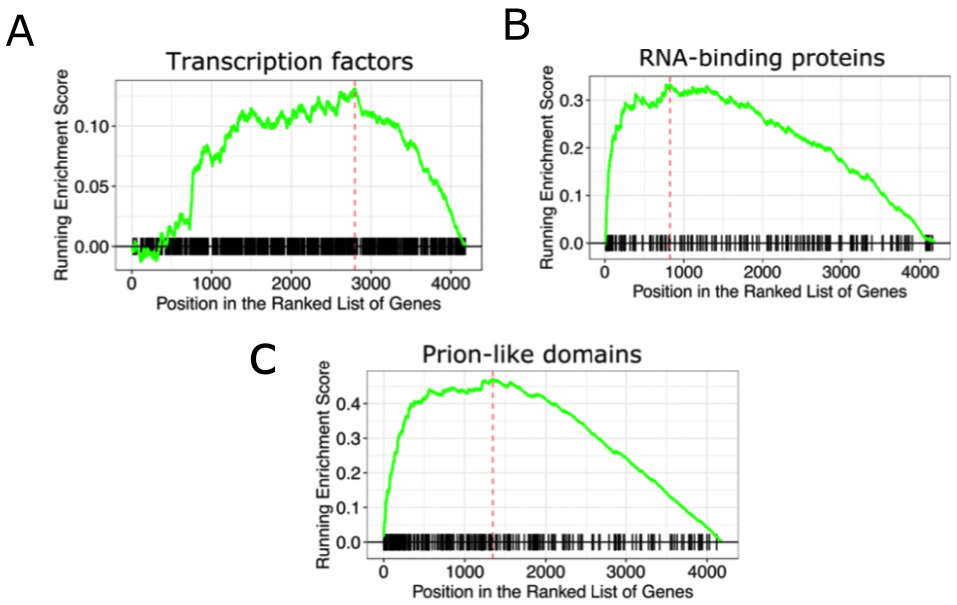


Figure 3.6: GSEA analysis for TFs, RNA-binding proteins, and prion-like domains. The GSEA algorithm computes an enrichment score that indicates the overrepresentation at the upper or lower end of a ranked list of genes from a gene set. The tick marks indicate the position of proteins in the ranked list. The Y-axis represents the enrichment score, where positive values signify gene set enrichment at the upper end of the ranked list.

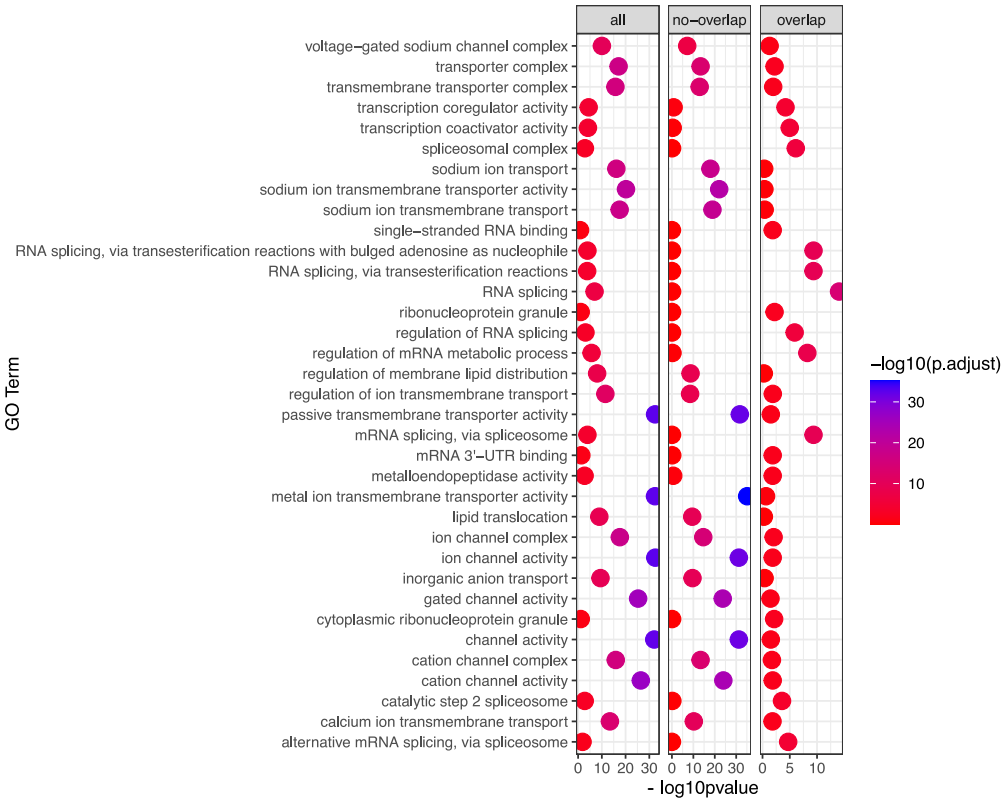


Figure 3.7: GO term enrichment analysis for proteins with significant periodic regions. Enriched GO terms for all periodic proteins, periodic proteins that have an overlap with metapredict IDR regions and have no overlap with metapredict IDR regions. The color of the circles represents the FDR value (p.adjust) of the enrichment analysis

3.3.2 Evolution of periodicity in GLI2

As shown in the previous section, periodicity is linked to disordered regions and can therefore be a functionally important characteristic of IDR regions. Due to their functional significance, we are interested in exploring the evolution of these periodic regions within protein sequences, specifically we aim to analyze how these periodic regions have evolved. In this section of the study, our focus is on the evolution of periodicity in GLI2, one of the TFs from our dataset that was identified to have a periodic block by our method.

GLI2 plays a critical role in vertebrate embryonic patterning across various regions. GLI2 is a member of the GLI family, which includes GLI1, GLI2, and GLI3. In a study conducted by Abbasi et al. (2009), the evolution of GLI proteins was studied. The authors reported that GLI1 genes have encountered distinct functional constraints, whereas GLI2 and GLI3 sequences have been subjected to similar levels of functional constraints across various lineages.

Using our method, we identified a periodic region in GLI2 between residues 1379 and 1479, with the window size of 100. Furthermore, GLI2 has a SLiM within its IDR region that may be conserved, making it an ideal candidate for studying evolutionary aspects.

The fact that GLI2 protein sequence has periodic aromatic residues in its disordered regions and that it is involved in important developmental processes suggests that periodicity may be important for its function. To investigate the evolution of periodicity in the GLI gene family, we aligned GLI protein sequences from nine different species that were used in the study. The sequences were aligned using the CLUSTAL method (Sievers et al., 2011), and a phylogenetic tree was constructed using the maximum likelihood in MEGA software (Kumar et al., 2018). We were able to recapitulate a tree with the same branching pattern as reported in their study, where vertebrate GLI2 and GLI3 genes clustered together, while GLI1 genes formed an outgroup to them (Figure 3.8).

To analyze the conservation of periodicity among GLI members, we computed p-values as periodicity scores for this periodic region in other GLI member protein sequences using our approach presented in the previous section. As illustrated in the phylogenetic tree with the corresponding p-values (Figure 3.8), ancestral *C. elegans* Tra and *Drosophilia melanogaster Ci* protein sequences have lower p-values, in contrast to GLI1 members, which show that ancestral members had an increased periodicity of aromatic residues. GLI1 members have fewer periodic aromatic residues compared to GLI2 and GLI3 members. This observation suggests that periodicity might be a dynamic trait that can be gained or lost over evolutionary time.

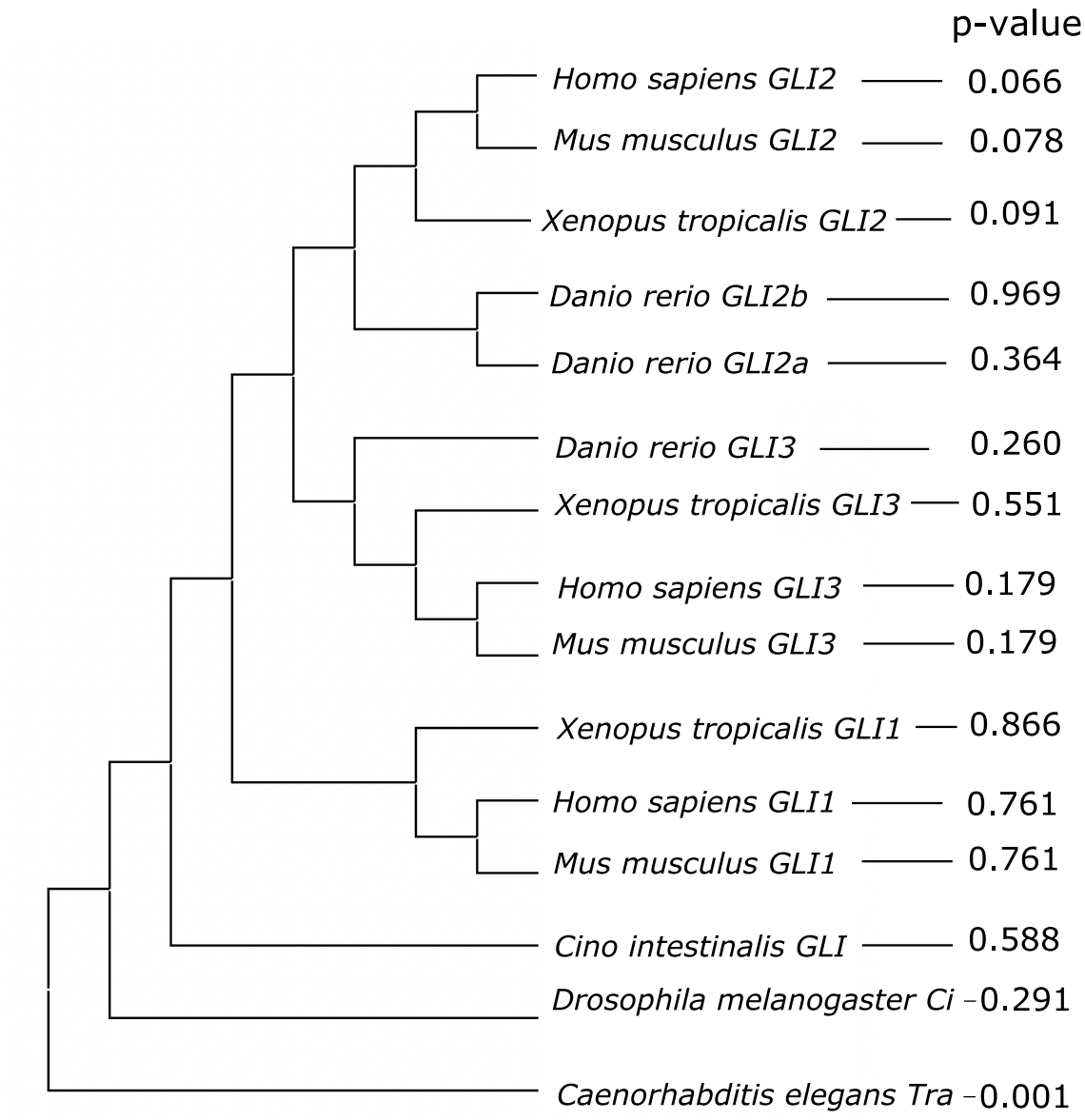


Figure 3.8: Periodicity across the evolutionary tree of GLI1, GLI2 and GLI3. Phylogenetic tree of GLI members with their corresponding p-values reflecting the periodicities. Smaller p-values indicate more periodic regions.

When we focus on the region of the periodic block in the protein sequences and highlight the aromatic residues, a dynamic pattern of periodicity becomes evident. Figure 3.9 illustrates the Jalview (Waterhouse et al., 2009) screenshot showing the CLUSTAL alignment (Sievers et al., 2011) of a specific region located within the analyzed periodic block. Interestingly, in the residue 2131, an aromatic residue is absent in the case of GLI1 members. This absence could potentially contribute to the reduced periodicity observed in GLI1 members when compared to GLI2 and GLI3 members.

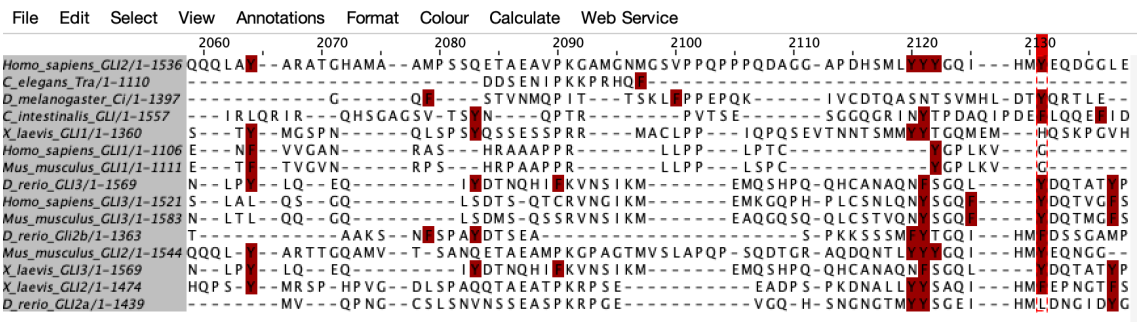


Figure 3.9: Screenshot of the Jalview visualization of alignment for the periodic block in GLI2. The screenshot displays the multiple sequence alignment of the periodic block across GLI members, with aromatic residues highlighted in red.

These findings suggest that periodicity might be a dynamic process, susceptible to gain or loss over evolutionary periods. Since we only focused on the evolutionary tree of one protein, it would be interesting to see whether other proteins involved in developmental processes also show increased periodicity in their disordered regions and whether this is a general phenomenon or specific to certain protein families.

3.4 Summary

It is still an ongoing challenge to understand the molecular grammar of IDRs guiding their transcriptional activities. In this study, we developed a statistical method to identify periodic aromatic regions in the human proteome. Our method is fast and adaptable, offering customization based on various sticker types and spacer lengths. By using our method, we investigated the enrichment of periodicity in different protein families, including TFs. We also investigated the overlap between IDRs regions and periodic regions. Surprisingly, just relying on periodicity enable us to identify predicted IDR regions in numerous proteins. This implies a relationship between these two regions, and future studies may be able to increase the precision of IDR predictions by include periodic regions.

The findings of our study have revealed that periodic regions are more abundant in PLD domains and RNA-binding proteins, while their prevalence is comparatively lower in TFs. This observation suggests that the periodicity observed in the IDRs of TFs might not be as optimized as that seen in prion-like domains. This leads to an interesting research question: understanding the functional impact of different levels of periodicity in proteins, including TFs.

We further investigated the evolutionary aspect of these periodic regions by analyzing the evolution of the periodic protein sequence from a transcription factor in our dataset. Further analyses are needed to understand the evolutionary aspect of periodicity. For instance, we need to determine if more periodic regions are conserved, or if periodic regions that overlap with IDRs are more conserved. Future analyses can also involve studying the association between protein function and the evolutionary emergence of periodicity which could provide us with a better understanding of the functionality of the periodic blocks and IDRs.

4 FEATURE IDENTIFICATION IN CO-OCCURRING TFS USING CONTINGENCY TABLES

In Chapter 2, Section 2.2.2, and Chapter 3, we discussed the presence of IDR sequences in TFs outside of their DBDs. Additionally, we highlighted findings that show the significance of IDRs in TFs and their role in mediating transcriptional regulation. An essential element of this transcriptional regulation involves the binding of multiple TFs to DNA sequences. In this chapter, we investigate IDR sequences in pairs of TFs that bind together on DNA sequences. We hypothesise that IDR sequences of co-binding TFs may be compatible in terms of their chemical properties, driving them to select each other for binding to DNA elements. We particularly focus on the IDR sequences of TFs that co-bind and evaluate the occurrence of amino acid groups within these sequences. Our aim is to investigate whether particular groups of amino acids co-occur, as these groupings could be potentially favorable for the binding decisions made by pairs of TFs. This analysis is done by analyzing the IDR sequences of TFs using contingency tables. In the following sections, we will introduce the phenomenon of TF cooperativity and the dataset we analyze. Following that, we will explain the multi-way contingency tables. Finally, we will demonstrate how we apply chi-square test statistics to these contingency tables representing TF interactions to identify combinations of amino acid groups in IDRs that might contribute to the TF cooperativity.

4.1 Background

TFs are regulatory proteins that have DNA binding domains that recognize a short specific DNA sequence to bind the DNA. TFs usually do not act alone on DNA elements, they interact with another partner TF (Banerjee, 2003). Understanding which individual TFs interact with each other to orchestrate gene expression is a crucial step in discerning mechanisms used by the biological systems to achieve tight regulation of transcription and a key step for understanding cell type diversity (Reiter et al., 2017; Reményi et al., 2004; Wright et al., 2015). It is still an ongoing challenge to understand how these TFs act together on the DNA elements (Bömmel et al., 2018; Ibarra et al., 2020).

As mentioned, IDR segments can contain features responsible for their interaction modes. A recent study by Barkai et al. (2020) demonstrated that IDRs of TFs can enhance *in vivo* binding specificity by directing them toward enhancer elements. Currently, it is unknown whether IDR segments in TFs also play a role in selecting TF partners to bind together on enhancer elements. We hypothesize that compatibility of IDR sequences between TFs may be responsible for these interactions between TFs. We therefore aim to analyze co-occurring amino acid groups throughout the IDR sequences of co-binding TFs (Figure 4.1).

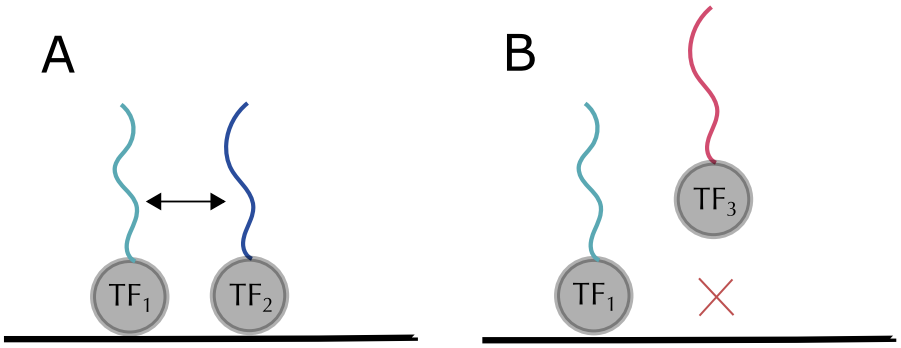


Figure 4.1: Schematic representation of co-occuring TFs. An essential part of gene regulation involves the cooperativity between TFs on DNA elements. We hypothesize that compatibility of IDR sequences between TFs may be responsible for these interactions between TFs. The colorful part of each TF represents the IDR regions. IDRs generally found outside of the DBD of TFs. A) Schematic representation of TFs that co-bind together B) Schematic representation of TFs do not bind together.

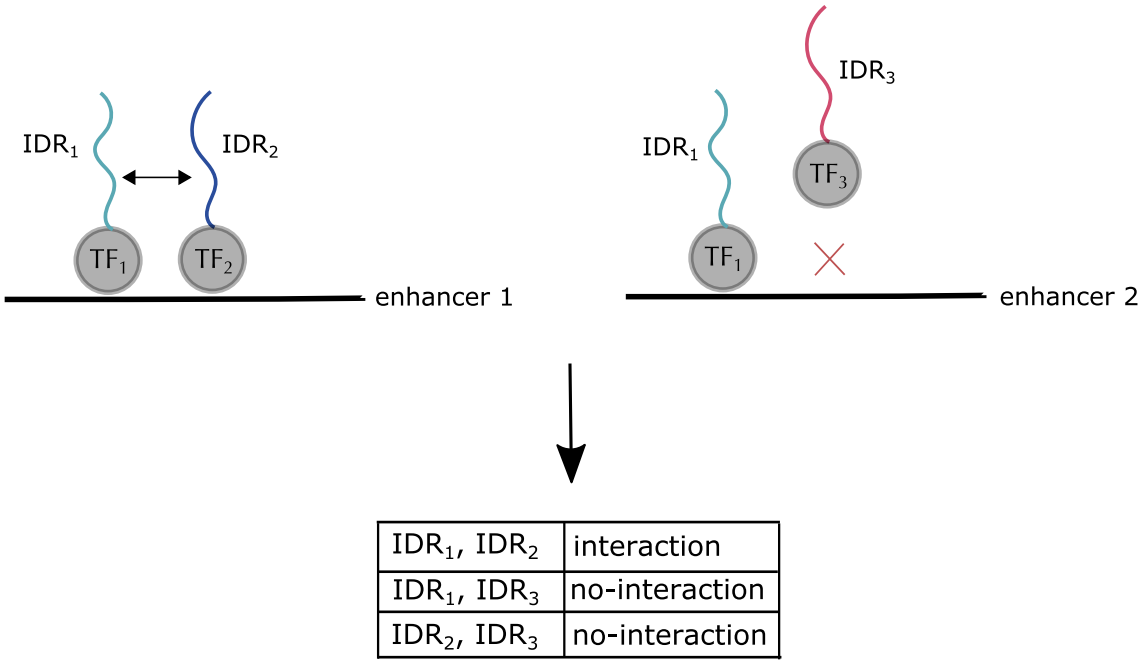
To test the hypothesis, we analyze the occurrence of pairs of amino acid groups located within IDR sequences in TFs using 2x2x2 contingency tables. Subsequently, we employed the Pearson's chi-squared test to assess the association between the occurrence of pairs of amino acid groups in the IDR sequences of TF pairs. This approach enables us to identify amino acid groups that may contribute to TF-TF interactions.

4.2 Methods

4.2.1 Dataset

We curated the dataset from Stark et al. (2015) which is derived via the enhancer complementation assays to evaluate the regulatory contributions of TFs and cofactors in the context of combinatorial enhancer control. They performed enhancer complementation assays, where they mutated enhancers and recruited 474 Drosophila TFs. In the end,

they measured the enhancer activities of these mutated enhancers after binding using luciferase assays in S2 cells. By assessing similarities in enhancer activity across different contexts, they identified 15 clusters of TFs that have distinct developmental functions. Interestingly, they found that similar TFs can substitute for each other, allowing for the reengineering of enhancers by exchanging TF motifs. They also suggested that these TFs cooperate during enhancer control and physically interact with each other.



IDR_1, IDR_2	interaction
IDR_1, IDR_3	no-interaction
IDR_2, IDR_3	no-interaction

Figure 4.2: Schematic representation of label creation. If two IDR sequences belong to TFs that bind together on enhancer elements, we classify this as an interaction; otherwise, we label it as "no-interaction".

For our analysis, we consider each TF belonging to the same cluster as a positive interaction (Figure 4.2). This resulted in a total of 3,500 TF pairs that are found in the same clusters.

4.2.2 Statistical analysis

For each pair of TFs, we counted the occurrences of residues from specific amino acid groups in IDR sequences. We used contingency tables to investigate whether the presence of certain amino acid groups in IDR regions has significance in determining the interaction between TF pairs. To employ contingency tables, we needed a binary output, which required us to establish a threshold for the occurrence counts. In other words,

we determined the minimum number of amino acids in the IDR sequences required to classify it as 'high' or 'low' for the amino acid group, depending on whether the count above or below this threshold.

As an example, consider Table 4.1, which is a 2x2 table to compare the occurrence of two specific amino acid groups (amino acid group a and amino acid group b) in IDR sequences of two TFs (IDR_1 and IDR_2) using a threshold of 10. This table shows how occurrences of specific amino acid groups are interrelated within the IDR sequences of two TFs. Each cell in this 2x2 table categorize each case as a binary outcome: whether the number of residues belonging to a certain amino acid group is higher or lower than a specific threshold which is 10 in this case.

	amino acid group $a > 10$ in IDR_1 (high)	amino acid group $a \leq$ in IDR_1 (low)
amino acid group $b > 10$ in IDR_2 (high)	Cell 1	Cell 2
amino acid group $b \leq 10$ in IDR_2 (low)	Cell 3	Cell 4

Table 4.1: Example of a two-way contingency table for two amino acid group occurrences in IDR_1 and IDR_2

In addition to this, we introduce a third variable in our analysis: 'interaction', indicating whether two TFs interact with each other or not. To summarize, we denote our variables as X, Y, and Z. For each pair of TFs in the dataset, variable X represents whether the number of residues belonging to a specific amino acid group is higher or lower than the threshold in the IDR sequence of the first TF, while variable Y indicates the same information for the second TF. In contrast, variable Z signifies the interaction between the first and second TFs, specifically whether they belong to the same cluster. Given that each categorical variable has two binary outcomes, and we have a total of three categorical variables in our analysis, the final resulting table is a 2x2x2 contingency table (Figure 4.3).

To construct the contingency tables, we generate pairwise combinations of five amino acid groups: polar, aromatic, apolar, positive, and negative residues (Table 4.2). For every

Amino acid group	**Amino Acids**
Positive	K, R, H
Negative	D, E
Polar	S, T, N, H, Q, G
Apolar	A, L, V, I, M
Aromatic	W, Y, F

Table 4.2: Amino acid classification for contingency tables

pairwise combination of amino acid groups, we construct a 2x2 table that tabulates the binary counts for interacting pairs and another for non-interacting pairs. We chose a threshold of 30 to distinguish between high and low numbers of amino acids. In this analysis, the order in which TFs are paired matters because the variables in the chi-square test are mutually exclusive. When considering each TF pair, we account for two possible orders: TF_1-TF_2 and TF_2-TF_1. As a result, every pair is counted twice in the contingency table.

As summary: For every combination a, b in aa_groups:

aa_groups = {polar, aromatic, apolar, polar, positive, negative residues}

For each pair of IDR sequences Q and L in the dataset:

$$X = \begin{cases} 0, & \text{if the number of residues belonging to amino acid group } a \text{ in sequence } Q \\ & \text{less than or equal to the threshold} \\ 1, & \text{if the number of residues belonging to amino acid group } b \text{ in sequence } Q \\ & \text{greater than the threshold} \end{cases}$$

$$Y = \begin{cases} 0, & \text{if the number of residues belonging to amino acid group } b \text{ in sequence } L \\ & \text{less than or equal to the threshold} \\ 1, & \text{if the number of residues belonging to amino acid group } b \text{ in sequence } L \\ & \text{greater than the threshold} \end{cases}$$

$$Z = \begin{cases} 0, & \text{if there is no interaction between sequence } Q \text{ and sequence } L \\ 1, & \text{if there is an interaction between sequence } Q \text{ and sequence } L \end{cases}$$

There are two 2x2 tables that make up the three-way table: an $X \times Y$ table within $Z = 0$ and an $X \times Y$ table within $Z = 1$. In our analysis, we focus on interacting TF pairs. Thus, we partition the 2x2x2 contingency table into 2x2 subtables by controlling for the levels of the interaction variable to test for independence between X and Y in interacting pairs (the red 2x2 table in Figure 4.3).

To test the relationship between two amino acid groups in interacting TF IDRs, we calculate the Pearson's chi-squared test for each contingency table corresponding to the each pairwise combinations of amino acid groups in order to identify enriched features in interacting TF IDRs ($Z = 1$). The null hypothesis is that X and Y are independent when $Z = 1$, meaning that the proportion of TF pairs with amino acid group 'a' exceeding the threshold is the same for TF pairs with amino acid group 'b' exceeding the threshold as it is for protein pairs with amino acid group 'b' not exceeding the threshold.

While the chi-square test is a useful method to understand if there is a significant association between two amino acid groups, it does not tell the nature of the association. To further validate the significant amino acid groups after the chi-square test and identify those that are abundant in both IDR sequences of interacting TFs, we visualize the

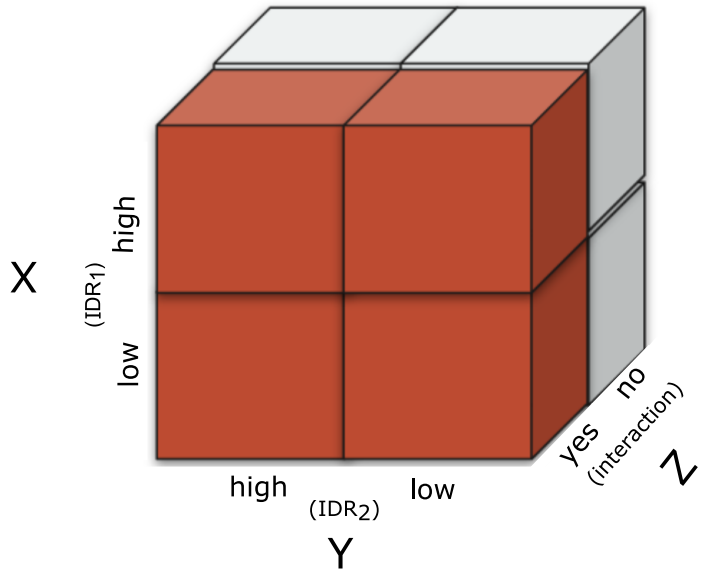


Figure 4.3: 2x2x2 contingency table. X represents the occurrence of an amino acid group in the first IDR sequence, while Y represents the same information for the second IDR sequence. Z indicates whether these two IDR sequences belong to TFs that interact with each other. High means the occurrence of the amino acid group is higher than the threshold, low means occurrence is lower than the threhold. High indicates that the occurrence of the amino acid group exceeds the threshold, while low signifies that the occurrence is below the threshold in the corresponding IDR sequence.

contingency tables of enriched amino acid groups in a mosaic plot. Mosaic plot is a useful technique to visualize Pearson residuals of computed contingency tables. Pearson residuals measure the departure of each cell from the observed and the expected count under the null model. Therefore, we use mosaic plots the interpret the occurrence of amino acid groups that are found to be significant.

4.3 Results

We extracted the longest IDR sequence for each protein by annotating the disordered sequences using the PONDR algorithm (Peng et al., 2006). Afterward, we created a contingency table for each combination of amino acid groups to categorize the abundance of pair of amino acid groups present in the IDR sequences of two TFs. Next, we applied the chi-square test to test the independence between amino acid groups. Chi-square test statistics have revealed a significant dependence in a total of five amino acid groups (p-value < 0.05). This finding implies that these specific amino acid groups might indeed play a role in interacting pairs. These groups are listed in Table 4.3.

Significant pairs of amino acid groups (p-value < 0.05)
(aromatic, apolar)
(positive, negative)
(positive, apolar)
(negative, apolar)
(polar, apolar)

Table 4.3: Significant pair of amino acid groups in co-binding TFs

After identifying the significant amino acid groups, we proceed to assess their prevalence within the IDR sequences of TFs. Subsequently, we extracted contingency tables and computed the Pearson residuals for these five pair of amino acid groups. We then used these residuals to create mosaic plots, providing a visual representation of the data which is shown in Figure 4.4. These mosaic plots revealed that, among all possible combinations, positive-negative residues were the only enriched amino acid group with positive residuals in the cell with high-high counts. This means, both positive and negative residues were found in high numbers within the interacting pairs. In the mosaic plot representing positive-negative pairs, the upper-left cell displayed a high count for both groups, resulting in a positive Pearson residual (Figure 4.4A). The positive residual in the cell with high-high counts indicates a higher than expected number of observations, suggesting a strong dependence between these two groups. Figure 4.4B shows the mosaic plot of polar-apolar amino acid groups as an example, which did not exhibit a similar pattern as positive-negative combination.

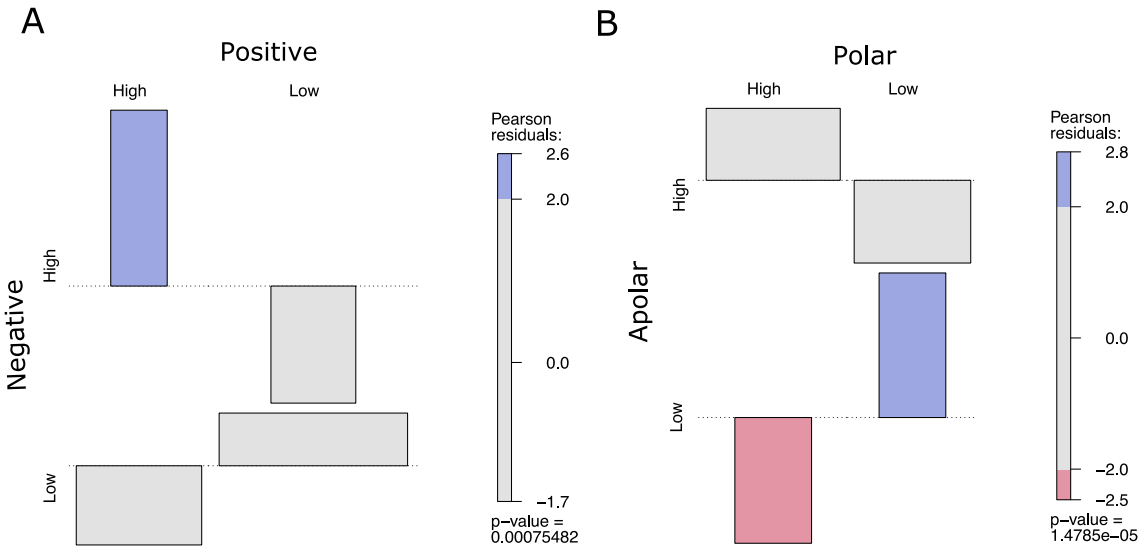


Figure 4.4: Mosaic plots based on contingency tables. Mosaic plots display Pearson residuals for each cell in the contingency table, revealing which combinations of amino acid groups contribute the most to the significance of the test of independence between (A) positive-negative amino acid groups and (B) polar-apolar amino acid groups. Blue cells (positive residuals), show that there are more observations in that cell than would be expected by the null model. On the other hand, red cells (negative residuals) show that fewer observations are present in that cell than the null model would expect.

4.4 Summary

In this chapter, we explored the dependencies between different amino acid groups in IDR sequences of TFs that bind together on enhancer elements. While significant effort has been dedicated to study the roles of DBDs of TFs in governing their transcriptional functions, less is known about the AD of TFs, which are enriched in terms of IDRs. To our knowledge, no previous research has investigated the IDR sequences of co-binding TFs.

We hypothesize that different amino acid groups in the IDR sequences may be responsible for the combinatorial binding of TFs and how they choose their enhancer partners. Our study focuses on the question of how different combinations of amino acid groups in TF IDRs influence the co-binding decisions between TFs. Our results indicate that interacting TFs have a high co-occurrence of positive and negative amino acid groups, suggesting that the high frequency of positive-negative residue combinations may indeed play a significant role in the interactions among TFs. These findings suggest a connection between the sequence of amino acids and the behavior of the TFs during interaction.

A recent study by Vandel et al. (2019) has shown that TF binding combinations in enhancers, compared to promoters, are more cell-type specific. Therefore, enhancer complementation assay can offer valuable data for combinations of TFs. On the other hand, we analyzed a cluster dataset, from which we extracted interactions. It's worth noting that these interactions may lack specificity since they don't inherently occur as pairwise interactions. An important application of this research is to identify important features for TF co-binding and input them into machine learning models for predicting these interactions or cell-type-specific responses. Overall, these results provide the potential significance of specific combinations of amino acid groups in the context of TF interactions.

5 MACHINE LEARNING BACKGROUND FOR PPI PREDICTION

In this chapter, we will review the literature on machine learning background for PPI prediction and explore some fundamental concepts that we will later apply in our work. Firstly, we will review available machine learning models, beginning with an explanation of machine learning methods. Next, we will explain the features mainly used in the protein world. Afterward, we will elaborate on the available methods and how they combine different protein features and machine learning methods. Developing machine learning models for PPI prediction comes with its own set of challenges because it involves pair of inputs and there are specific challenges associated with this pair prediction. In the end, we will discuss these machine learning challenges related to pair prediction.

5.1 Machine learning methods

Machine learning methods can be categorized as supervised or unsupervised learning. Typically, predicting PPIs involves a binary classification task in which the model estimates the probability of interaction between two given proteins (Hu et al., 2022). Therefore, supervised learning is the most relevant approach for PPI prediction. In this chapter, our focus will be on supervised machine learning methods for PPI prediction. Some of the most popular methods employed in PPI prediction models include support vector machines, random forests, and Bayesian inference.

The primary objective of machine learning methods is to learn a mapping function based on the provided training data (Sarker, 2021). This function takes into account various features of a protein pair from the training set and generates a prediction score. This score, ranging from 0 to 1, represents the probability of protein interaction. Additionally, deep learning methods have become increasingly common in PPI prediction models. In the following sections, we will first explain the random forest machine learning model and then briefly touch upon deep learning methods.

5.1.1 Random forest

Random forest (RF) is a supervised learning algorithm that combines individual decision trees to form an ensemble learning to enhance prediction and accuracy (Figure 5.1). A decision tree can be seen as a step-by-step decision-making process with a binary tree structure. Each decision tree consists of a root node, intermediate nodes, leaf nodes, and decision rules. The root node signifies the tree's starting point, and as additional decision rules are added, the tree generates intermediate and leaf nodes. Leaf nodes serve as stopping points where the tree completes its growth and produces final predictions.

Given the dataset D, the process of building the RF model begins by creating an ensemble of decision trees. We construct B decision trees, denoted as $T_1, T_2, \ldots, T_B$, where B represents the desired number of trees. To create each decision tree T_i, a random subset of the training data D_i is selected using a technique called bootstrapping. This involves randomly sampling observations from D with replacement. The selection of D_i is performed using a random vector θ_i, which ensures that the values are chosen independently and with replacement while being different from the $m-1$ previously generated vectors (Algorithm 2).

Each decision tree T_i is built using the bootstrap sample D_i. At each node of the tree T_i, we further introduce randomness by randomly selecting a subset of m features from the total d features. The random feature selection is determined by the random vector θ_i. With the selected features at each node, we determine the best split based on the Gini index, which measures the impurity of the node. This split helps in determining the optimal partitioning of the data based on the selected features at that node. Once all B decision trees are constructed they are combined to have the ensemble of trees. The classification of a new instance $x \in X$ is performed using a majority voting mechanism. Each decision tree casts a vote for the class of the instance x, and the final prediction is based on a majority vote over trees $T_1, T_2, \ldots, T_B$. The classification output is presented as a probability value, representing the fraction of trees that voted in favor of the predicted class (Figure 5.1).

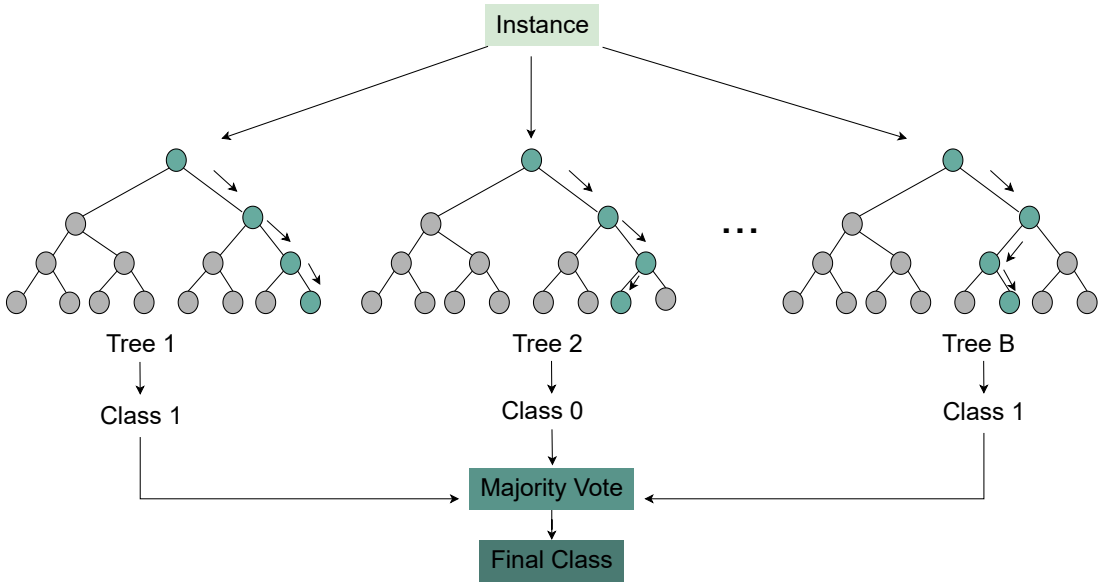


Figure 5.1: Random forest model. Each of the B decision trees assigns a label to a new data instance based on its path through the tree and the terminal node it reaches. The final class assigned to the new data instance is determined by taking the majority vote of the labels assigned by the individual trees.

Algorithm 2: Random Forest Algorithm

for $b = 1$ to B **do**
  Generate a random vector θ_b.
  Choose a bootstrap sample D_b from the observations D based on θ_b.
  Construct a decision tree T_b from the bootstrap sample D_b.
  for each node in the tree T_b **do**
    Randomly select m features from the total d features based on θ_b.
    Find the best split at the node based on the Gini index using the selected features.
  end for
  Add the constructed tree T_b to the forest collection.
end for
Return the forest collection, which contains B decision trees.

Two sources of randomness—bootstrap aggregation and random feature selection—minimize correlation among the trees and introduce variability in the choice of features at each node, leading to reduced inter-tree dependency. An additional advantage of this feature selection approach is the ability to quantify the feature importance of the variables

which is typically done by looking at the improvement or loss in accuracy after excluding or including a specific variable from the trees.

Ensemble methods offer several advantages, primarily in reducing errors through the combination of multiple trees. When using a single tree, the model tends to have high variance, making it sensitive to the specific arrangement of data points within the bootstrapped dataset. This variance arises from the fact that each tree is built using only a subset of the data, leading to high variability among individual trees. After combining the predictions, the overall ensemble is less prone to the high variance exhibited by individual trees. This reduction in variance contributes to improved generalization and stability of the final estimator, making ensemble methods a powerful approach in machine learning.

The RF algorithm has two key parameters namely the number of trees (B) and the number of features considered for splitting at each node of the tree (m). In other words, only a subset of features is utilized for determining the optimal splits. If we have a feature vector x_i with d dimensions, it is advisable to choose values for B and m such that they are less than or equal to the total number of features (d), as this helps maintain diversity and prevent overfitting. Finding an appropriate balance between these parameters is crucial for achieving optimal results.

5.1.2 Deep learning methods

Deep learning methods are neural networks with multiple layers of interconnected computational nodes, referred to as neurons, capable of computing input-output mappings (O'shea et al., 2015). Different classes of deep neural networks are defined based on the types of layers and connection topologies, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks, and Siamese Neural Networks.

CNNs are one of the most commonly used deep learning architectures, typically involving convolutional layers, pooling layers, and fully connected layers (Yamashita et al., 2018). The distinctive feature of CNNs is the convolutional layer, which performs convolution operations. In CNNs, convolution is used as an operator, where a small array of numbers, known as a kernel, moves across the input, performing element-wise multiplication to create feature maps that represent various characteristics of the input dataset.

Recently, the Siamese architecture has become a common choice as a deep learning method in many PPI prediction methods (Chen et al., 2022; Czibula et al., 2021). This architecture consists of two identical submodules that share the same configuration and

weights. Each submodule can be trained on one input from the pair, allowing them to learn the data individually. Therefore, this model can be a good choice for predicting PPI in pairs. Each submodule generates high-dimensional feature representations for its input. Subsequently, these outputs are combined into a final vector representing the protein pair. These final representations are then used to calculate an interaction score and predict the probability of interaction (Figure 5.2).

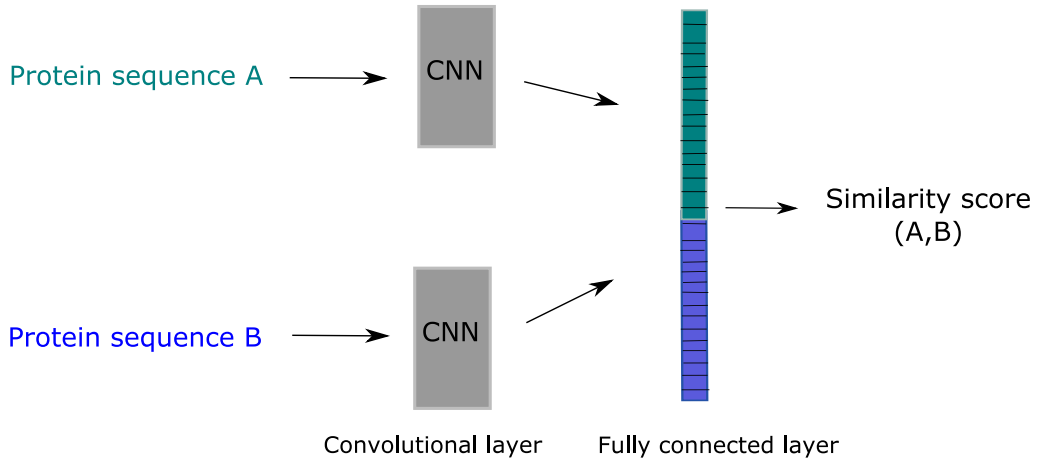


Figure 5.2: Overview of the Siamese-Joint architecture. Siamese-Joint architecture has two identical submodules, consisting of convolutional neural networks (CNNs). The CNNs process input sequences, extracting high-level features through convolutional layers. These features are then passed through a fully connected layer, merging the outputs of the CNNs. Adapted from Chira et al., 2021

There are some limitations of deep learning methods. Firstly, training deep learning models can be a slow process due to the large amount of datasets required and the inherent complexity of the models. Second, often the interpretability of deep learning models in the context of PPI prediction remains a challenge, as they often function as black boxes, making it difficult to understand the underlying mechanisms driving their predictions.

5.2 Protein input features

In this section, we will discuss the available and commonly used features in PPI prediction methods. Once the machine learning method has been selected, we can proceed to select the features to input into our machine learning model. In the field of PPI prediction, machine learning methods can be broadly categorized into sequence-based and structure-based approaches (Soleymani et al., 2022). The choice of approach significantly influences

the features we extract. In this section, we will focus on the features derived from protein sequences.

To begin, the first step is selecting an appropriate PPI dataset from which we extract proteins and their interactions. Among PPI datasets, HIPPIE (Alanis-Lobato et al., 2017) and HPRD (Peri et al., 2004) are commonly used resources for PPI data. After getting the proteins and their interactions from the database of choice, one needs to collect the sequence data for all the proteins that are involved in the interactions. It is common practice to collect protein sequences from the Uniprot database (Bateman et al., 2021). After selecting the dataset, researchers often use tools such as CD-HIT (Li et al., 2006) or BlasClust (Wei et al., 2012) to remove similar sequences by clustering them based on their similarity. Once the protein sequences have been collected and processed, features can be extracted.

Many PPI prediction tools employ numerical feature vectors that are extracted from protein sequences. The process of converting protein sequences into numerical vectors is referred to as feature extraction in the context of protein sequences. Machine learning tools use various feature extraction methods to develop algorithms for PPI prediction. These methods are based on different types of features, the most common features include physicochemical features of amino acids, evolutionary information, domain annotations, GO annotations, and direct sequences (Hashemifar et al., 2018; Mu et al., 2021; Wang et al., 2020).

The physicochemical properties of amino acids have been widely used by several prediction methods. These properties include amino acid characteristics such as residue hydrophobicity, charge, polarity, volume, and conformational propensities. They are commonly sourced from databases like AAindex (Kawashima et al., 2000) or obtained through dimensionality reduction techniques applied to precomputed residue properties. These features can also be structural features involving protein flexibility and disorder. By incorporating these properties into the models, the aim is to capture essential information about the chemical and structural features of amino acids, enabling more accurate predictions of protein properties or interactions.

Evolutionary information is also used by many studies and provides insights into the functional importance of the residues which has proven to be also valuable in predicting PPIs. Highly conserved positions are indicative of critical functional or structural roles in protein interactions. Position-specific scoring matrices (PSSMs) can be derived from sequence alignments of homologous proteins using multiple sequence alignment (MSA) in position-specific iterative BLAST (PSI-BLAST) (Altschul et al., 1997). It can be also

obtained from databases such as Pfam (Mistry et al., 2021) or NCBI's Conserved Domain Database (Marchler-Bauer et al., 2015).

Domain annotations are another feature based on the evolutionary conservation of proteins, as domains are predominantly conserved throughout evolution (Chen et al., 2005). Protein domains are also reliable indicators of protein functions. Pfam (Mistry et al., 2021), InterPro (Blum et al., 2021), and Prosite (Sigrist et al., 2002) are the commonly used databases to get the domain annotations for proteins. Domain information is usually represented either directly or by creating feature vectors that indicate the presence of specific domains in protein pairs.

Finally, the semantic similarity using GO annotations of proteins is also recognized as a important characteristic for predicting PPIs (Ieremie et al., 2022). This approach is based on the idea that interacting protein pairs may engage in interactions within the same cellular location and participate in similar biological processes. As a result, these two forms of interaction share similar GO terms, which can be valuable features in PPI prediction.

5.3 Available models for PPI prediction

Various machine learning models have been developed so far for PPI prediction. Machine learning models for PPI prediction are generally distinguished by the machine learning methods they use and the features they employ. In this section, we will discuss some of the most popular machine learning tools developed for PPI prediction and the methods and features they use.

Some of the models use evolutionary profiles of proteins in combination with supervised machine learning models. In 2015, Rost and Hamp introduced their novel method based on the PSSMs as an evolutionary profile of proteins, using support vector machines in their model (Hamp et al., 2015). Later, Hashemifar et al., (2018) used the PSSMs based on this study but employed Siamese deep neural networks to train their model. As mentioned, some of the studies use domain annotations as a proxy for evolutionary information. For instance, in a PPI prediction model developed by Xue-Wen Chen and Mei Liu in 2005, domain annotations of proteins are used as features, and a RF is used as the machine learning model. In their model, each feature vector consists of values, namely 0, 1, or 2, reflecting domain knowledge. If a protein pair does not contain a specific domain, the corresponding value for that domain is set to 0. If one of the proteins in the pair

possesses the domain, the value is set to 1. Finally, if both proteins contain the domain, the value is assigned as 2.

Table 5.1: Available models and input features

Publication	Method	Input Features
Sledzieski et al., (2021)	Deep Neural Network	Pre-trained language model embeddings (Bepler et al., 2019)
Chen et al., (2019)	Siamase Neural Network	Similarity of electrostaticity and hydrophobicity
Perovic et al., (2018)	Random Forest	Pseudo amino acid composition, dipeptide composition
Hashemifar et al., (2018)	Siamase Neural Network	PSSM
Du et al., (2017)	Deep Neural Network	Amino acid composition, dipeptide composition, composition, transitions, and distributions of residue along the sequence, pseudo-amino acid composition
Rost and Hamp, (2015)	Support Vector Machine	PSSM
Chen et al., (2005)	Random Forest	Domain annotations

Many machine learning models use physicochemical features of amino acids, including amino acid composition and dipeptide composition. DeepPPI (Du et al., 2017) is one of these methods; in their study, they use a deep neural network along with the physicochemical features of proteins, which are shown in Table 5.1. IDPpi is another PPI prediction model, a RF model, that uses pseudo-amino acid composition and dipeptide composition as a feature set as well. In the next chapters, we will explain IDPpi in more detail, as well as one of the state-of-the-art algorithms in the field, D-SCRIPT.

5.3.1 IDPpi

IDPpi (Perovic et al., 2018) is a RF model that uses pseudo amino acid composition (PAAC) and dipeptide composition (DC) as feature sets, which are extracted from the entire protein sequences. The construction of the PAAC composition involves incorporating amino acid propensity scales into the calculations (Table A1). These scales, including the TOP-IDP scale, B-values, the FoldUnfold scale, and the DisProt scale, assess the order or

disorder tendencies of amino acids. By incorporating these scales, IDPpi aims to capture the degree of disorder in relation to neighboring residues.

IDPpi has a specific focus on studying interaction networks involving IDPs. It relies on features extracted from entire protein sequences. For the PPI data curation, they analyzed the partners of IDPs comes from DisProt database (Piovesan et al., 2017). The curated PPI dataset consisted of 19,837 interactions of 5,989 human proteins. The method reached an averaged AUC score of 0.74 in the original paper on 5 cross-validation test sets

5.3.2 D-SCRIPT

Deep sequence contact residue interaction prediction transfer (D-SCRIPT) (Sledzieski et al., 2021) is a recently published state-of-the-art approach which is a deep learning method developed for predicting PPIs.

In the process, a pre-trained language model (Bepler et al., 2019) is incorporated to extract features for individual proteins, generating low-dimensional embeddings. These embeddings are then used in the projection and convolution layers to make final predictions. The intermediate representation of this deep learning approach serves as a contact module, creating an inter-protein contact map. This contact map predicts the probability of interaction between all pairs of residues in a given protein pair.

To obtain PPI data, D-SCRIPT has extracted the interactions from the STRING database (Szklarczyk et al., 2015) and employed a 150-dimensional layered deep learning framework for predicting protein interactions. D-SCRIPT was trained originally on 38,345 human proteins and reached an AUC score of 0.833 on the human PPI network. Users can either choose to train the model from scratch using their training dataset or use a pre-trained human model.

5.4 Pair prediction

All of these methods, along with the problem we are addressing, are not just simple machine learning tasks; they share a characteristic feature of being a pair prediction problem. This means, we are not just developing a tool that deals with single inputs, as with other machine learning methods; our methods operate on pairs of objects. This aspect of the problem has a significant impact on various layers of the machine learning methods, differentiating it from the process of developing a model that operates on single

inputs across multiple steps. This inherent characteristic of paired datasets requires more careful considerations compared to other machine learning applications.

Until now, significant efforts have been devoted to standardizing techniques in machine learning development for common single inputs. However, most of these techniques fails to operate on paired inputs due to their complex structure. Furthermore, the paired nature of the dataset is not limited to PPIs but is present in many important biological machine learning applications, such as drug-drug interaction prediction or drug-target prediction. However, to our knowledge, there is no golden standard or comprehensive guide for handling paired inputs in machine learning models, addressing all the associated challenges that arise throughout the entire modeling process. Indeed, there exist important papers that have addressed the influence of this characteristic on specific aspects of the machine learning models, which are presented in the below sections (Hamp et al., 2015; Park et al., 2011, 2012; Yu et al., 2010). In the following sections, we will address the challenges and issues in the field of pair prediction.

5.4.1 Feature combination

In Section 5.2, we have demonstrated how each protein can be represented with a feature vector. However, these features are not inherent attributes of the pairs and rather belong to individual proteins. The first challenge that comes with the paired nature of inputs is the necessity to fabricate new features to represent the pairs for which interactions are predicted. The common approach is the fusion of feature vectors obtained from two individual proteins to create a combined feature representation for each protein pair. This approach comes with an inherent challenge regarding how to fuse the features. Again, there is no accepted standard for feature fusion. Various approaches have used different techniques to date.

Concatenation and combination are two commonly used approaches that are employed for feature fusion in PPI prediction (Chen et al., 2019). Concatenation is a approach where the feature vectors of two proteins are combined into a single vector, as illustrated in Figure 5.3. This approach is also widely used in deep learning approaches (Hu et al., 2022). It's important to note that concatenation is non-commutative, meaning that the resulting vector depends on the order of pairs. Additionally, concatenating feature vectors doubles the number of dimensions, potentially making it more challenging for machine learning models to process the data. Another limitation of concatenation is that it does not create a unique feature vector; instead, it simply combines the vectors. Several studies have

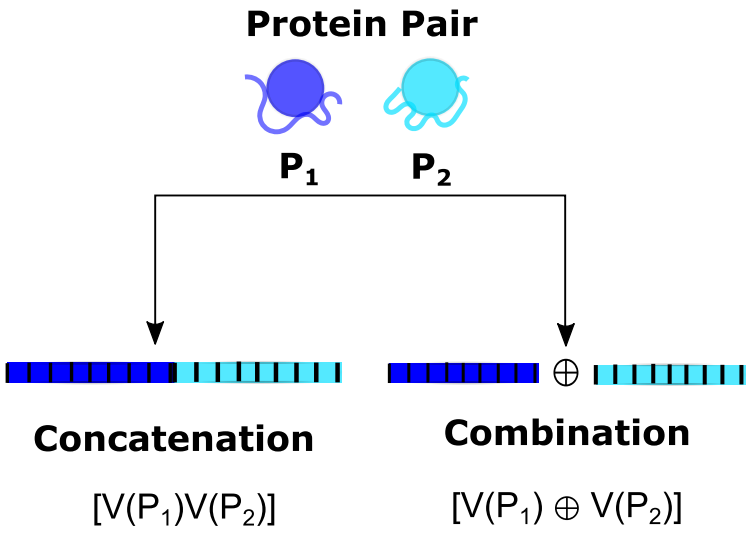


Figure 5.3: Feature combination techniques for machine learning in PPI prediction. Two approaches are typically used which are concatenation and combination. Concatenation involves fusing vectors one after another. On the other hand, combination involves applying a specific operation in an element-wise fashion between feature vectors.

recommended using feature fusion techniques to create a unique feature vector for pairs when dealing with two fixed-length feature vectors (Ross, 2009; Zhang et al., 2014).

Another approach is combination, which involves element-wise operations such as addition or subtraction. In this method, the corresponding elements of the feature vectors are objected to a operation like addition or subtraction. This approach can give the differences or similarities between the individual features of the protein pairs. The combination approach allows for preserving the original feature dimensions and creates a unique feature vector for pairs.

5.4.2 Sampling strategies for negative training dataset

This section discusses another aspect of the pair prediction task, which involves selecting non-interacting PPI data (negative examples) for machine learning methods in PPI prediction. The availability of non-interacting PPI data is just as crucial as the interacting PPI data for machine learning methods. While databases typically provide information about interacting proteins, access to non-interacting PPI data is relatively limited and not widely accessible. It has become a common practice to assume that a pair of proteins for which no interaction is reported are accepted as non-interacting pairs (Park et al., 2011). Naturally, the number of negative examples is generally much larger than that of positive examples, resulting in an imbalanced dataset.

Of course, dealing with imbalanced datasets is not only unique to PPI machine learning tasks. In classical scenarios, researchers often employ precision-recall curves to deal with imbalanced datasets, or they downsample the dataset by selecting a reasonable number of points from the dataset to create a balanced one. Subsampling is a routine practice to balance the dataset by obtaining an equal number of negative and positive interactions. However, in our case, we are not dealing with individual nodes; instead, we have a graph representing our protein interaction network, where enodes represent proteins while edges represent interactions. This raises the question of how to subsample the graph. As expected, the process becomes more intricate when dealing with paired inputs and selecting data points for downsampling. In our case, we are not only selecting individual nodes; rather, we must carefully select pairs of nodes from this non-interactions network that accurately represent the intricacies of non-interacting proteins in our network. The subsampled version of the set is expected to "represent" the entire set of non-interactions. Exactly how one subsamples a negative training set has a large impact on the performance evaluation of the machine learning models which makes choosing the appropriate subsampling technique crucial.

Another issue we should consider when we subsample from a PPI network is the potential risk of creating dependencies between the subsampled negative PPI and the positive dataset. Given that this is paired data, it's possible that in either the negative or positive part of the dataset, some proteins may have more interactions and, consequently, more edges. These proteins, often referred to as overstudied proteins, may become overrepresented in the positive or negative part of the data.

In most of the PPI machine learning models, a random sampling approach has been used for subsampling the negative PPI dataset. As the name suggests, this approach involves the random selection of a subset from the entire negative dataset to reduce the number of negative pairs to match the number of positive pairs. However, due to the paired nature of the PPI dataset, this approach ignores the dependencies between the positive and negative training sets.

To overcome this issue, Westhead and colleagues (2010) come up with an approach called "balanced sampling". Balanced sampling aims to create a negative set where each node (protein) has an equal number of degrees (interactions) as in the positive set. This is achieved by subsampling a negative training dataset from the non-interacting pairs, ensuring that the degree of each protein in the positive dataset equals to that in the negative dataset.

An example dataset showing how balanced sampling works is provided in Table 5.3. In the example the negative training dataset is created using a balanced sampling approach. Training pairs are composed of interacting protein pairs as positive examples and non-interacting protein pairs as negative examples. The training dataset consists of four interactions among seven proteins. Non-interacting protein pairs are created using a balanced sampling approach, ensuring that each protein has the same number of interactions in both the interacting and non-interacting parts of the training data. For example, protein 1 (P1) has two interactions in both the negative and positive portions of the training data, whereas protein 2, protein 3, protein 5, protein 6, and protein 7 each have only one interaction.

Of note, Park and Marcotte (2011) suggested that the strategies used for creating a negative training dataset and a negative test set should differ. They did not recommend balanced sampling for selecting negative test pairs, but only for the negative training dataset. In the next section, we will discuss the strategies for selecting test sets.

5.4.3 Testing schemes

It is a common practice to evaluate the model's predictive performance using the test data that the model has never seen before. The standard procedure for this testing is cross-validation (CV). In CV, the dataset is split into k folds. Then for each fold, the model is trained on the remaining $k-1$ part of the original data while one fold is set aside as a temporary test set to test the accuracy. This method overcomes the risk of overfitting and gives more reliable accuracies. In k-fold cross-validation, the splits between folds are performed randomly. Typically, researchers split the available data into a training set and a test set, with a ratio of approximately 80% for training and 20% for testing (Joseph, 2022). They then repeat this process to perform 5 or 10-fold cross-validation.

However, when it comes to applying these techniques to paired datasets, things become more challenging yet again. It is unrealistic to expect similar outcomes when these methods are applied to paired datasets. Many machine learning models that utilize PPI inputs adopt a similar approach, treating them as single-nature inputs. These models use a random sampling approach, selecting both positive and negative test pairs randomly from the initial positive and negative datasets to create test sets. It's important to note that this is no longer the original cross-validation method when applied to paired-nature inputs. In traditional CV, the aim is to test the model on unseen data. But, we cannot expect this effect when we apply this technique to paired-nature inputs. Due to the paired nature of the data, there is a inherent dependency between training and test data.

Therefore, when random sampling is applied, we introduce randomness into the selection process for test pairs which leads to potential biases due to the overlap between training and test pairs. We can apply CV to obtain more reliable accuracy scores, however, it is important to note that this application differs from traditional CV methods.

Park and Marcotte, (2012) addressed this issue by suggesting the partitioning of test pairs into three distinct classes: C1, C2, and C3, based on their component-level overlap with the training set. Essentially, pairs can be categorized into these three types of features according to their overlap with the training set

- C1 has test pairs sharing both proteins with the training set.
- C2 has test pairs sharing only one protein with the training set.
- C3 has test pairs sharing neither protein with the training set.

Table 5.3: Example dataset: Training Pairs

Interacting protein pairs	**Non-interacting protein pairs**
P1-P7	P1-P6
P7-P5	P1-P5
P2-P6	P2-P7
P1-P3	P6-P3

Table 5.4: Example dataset: Test Pairs

Class	**Example**
C1	P2-P3
C2	P3-P8
C3	P8-P9

Example test pairs for these three classes can be found in Table 5.4. As expected, their study revealed that C3 is the most challenging class to predict when compared to C1 and C2. They observed that relying solely on cross-validation with C1 test pairs might not accurately reflect the model's performance at the population level. As a result, they recommend reporting the performance of machine learning models separately for each category of test pairs.

5.5 Evaluation

As mentioned in previous sections, the conventional methods for assessing model performance fall short when dealing with PPI data characterized by paired inputs. Therefore, one must already take all the considerations we discussed into account in the model development process to assess its final performance for pair prediction.

Evaluating the performance of the final classifier is crucial for understanding the model parameters and assessing its generalizability on the test set.

To assess the final performance of machine learning classifiers, multiple statistical performance metrics are generally employed, including accuracy, recall, precision, and F1 score. To calculate each metric, prediction results of the unseen test dataset are typically used to construct a 2x2 confusion matrix. In the case of binary PPI prediction scenarios, outputs are categorized as either positive interactions or negative interactions (non-PPI). The 2x2 confusion matrix for PPI prediction is shown in Table 5.5

Table 5.5: Performance metrics for ML classifiers

	Predicted : PPI (+)	**Predicted: Non-PPI (-)**
Actual: PPI (+)	True Positive (TP)	False Negative (FN)
Actual: Non-PPI (-)	False Positive (FP)	True Negative (TN)

In the context of a 2x2 confusion matrix for PPI, true positives (TP) represents cases where the actual value is 'PPI' and the predicted class is also 'PPI'; false positives (FP) represents cases where the actual value is 'Non-PPI' but the predicted class is 'PPI'. False negatives (FN) occurs when the actual value is 'PPI' but the predicted outcome is 'Non-PPI', while true negatives (TN) refers to cases where both the actual and predicted results are 'Non-PPI'.

Once the counts for each of the four prediction cases are determined and summarized in a 2x2 table, the performance metrics can be calculated using the following equations:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} \tag{5.1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.2}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.3}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

The area under the ROC curve (AUC) is another commonly used performance, the AUC is computed by summing the products of the differences in rank (R) and the sum of True Positives (TP) for adjacent thresholds, divided by twice the total number of positive instances multiplied by the total number of negative instances.

When analyzing the performance results, several important factors need to be considered; this is a standard practice in machine learning. Firstly, the dataset's class distribution plays a significant role. If one class is substantially larger than the other, relying solely on certain metrics may not provide an accurate assessment of the model's performance. For example, in cases where the data contains more positive examples and the model consistently predicts the positive class while assigning a 0 output to the negative class, the accuracy can be misleadingly high, which shows accuracy might be not the best metric to evaluate the model. Another consideration is the potential imbalance in the prediction outputs. It is important to be aware of the trade-off between different performance measurements. For instance, even if the initial data is balanced, a model that consistently predicts the negative class may exhibit high precision but low recall. In summary, when analyzing performance results, it is critical to consider the dataset's class distribution, be cautious of metrics that may be misleading due to imbalanced data or skewed predictions, and understand the trade-offs between different performance measurements.

The choice of the best method ultimately depends on the specific use case. Different use cases may require different methods for pair prediction. Furthermore, the nature of the pair-prediction problem we are dealing with plays a significant role. For instance, Park and Marcotte, (2012) discussed the applicability of random sampling when predicting C1 class pairs, which might be a limited study. Therefore, the selection of methods heavily relies on the specific type of pair prediction we aim to achieve and the type of pairs we want to predict. Park and Marcotte, (2012) primarily addressed this issue in the context of selecting test pairs. However, the considerations that must be taken into account during the model development phase represent another facet of this issue. To address this, we need to have a clear understanding of the specific pair prediction task or question we are targeting. How we define the pair prediction problem types and tackle each problem type is explained in detail in our study, specifically in Chapter 6, Definition of the problems.

6 PREDICTION OF PROTEIN-PROTEIN INTERACTIONS OF IDPS

Please refer to the Publisher Version (https://doi.org/10.1002/prot.26486) for access to chapter 6.

7 INFERRING NOVEL IDP-SPECIFIC AMINO ACID CONTACT POTENTIALS

In this study, our aim is to derive novel contact potentials between amino acid groups, specifically computed for interactions of disordered proteins. This is achieved through a statistical analysis of IDP-IDP interaction network. To illustrate our method, we utilize these contact potentials to generate contact maps of interacting protein pairs in our dataset, investigating subregions that could be crucial for their interactions. Generated contact maps of the interactions enable us to identify protein regions and prioritize IDRs with favorable interactions. We hypothesize that these identified protein regions may play a role in interactions or might be located in the interaction surfaces.

7.1 Background

As discussed in Chapters 5 and 6, predicting PPIs involving IDPs is crucial for gaining essential insights into their roles and functions. As mentioned earlier, PPI prediction typically focuses on a binary interaction task, where the model predicts whether two given proteins interact. However, this prediction does not explain which domains IDPs interact through or the strength of their interactions. Therefore, in addition to this binary information, it is important to identify the interaction interfaces and pinpoint the protein regions in the IDPs that could be important for the interaction.

PPIs depend on a certain degree of 'compatibility' between interacting proteins, including their structural, electrostatic, and other attributes. These characteristics can be found in the interaction interface of these proteins (Chothia et al., 1975; Jones et al., 1996; Livingstone et al., 1993). To gain a better understanding of these interactions involving IDPs and their interfaces, it is important to determine which amino acids or amino acid groups have more favorable interactions with each other in interacting IDPs when compared to their non-interacting counterparts. This insight can provide valuable clues about the essential properties driving interaction dynamics and potentially reveal the specific amino acids or groups present on the interaction surface of these interactions.

Amino acid contact potentials offer a useful method for understanding how amino acids prefer to interact with one another. Traditionally contact potentials are derived for

amino acids by analyzing a vast database of known protein structures and examining the frequency of occurrence of different amino acid pairs in close proximity to each other (Khatun et al., 2004; Pokarowski et al., 2005). The assumption is that amino acid pairs that occur more frequently in close contact with experimentally determined protein structures are likely to exhibit favorable interactions. This information is usually derived from statistical analyses of experimentally determined protein structures. Statistical analysis is carried out by comparing the observed frequency of an amino acid pair with the expected frequency based on the occurrence of each amino acid individually, assuming no preference for specific pairs. This comparison is generally made using protein complex structures found in structural databases (Holland et al., 2022). The resulting potentials at the end of this analysis describe the interaction energies between the 20 amino acids and can be visualized as a 20x20 matrix. Each element of this matrix represents the interaction strength between a pair of amino acids in contact (Buchete et al., 2008). Amino acids that are in close proximity are expected to have positive scores with higher potentials indicating favorable interactions (Eyal, 2005).

Previous studies have demonstrated the usefulness of contact potentials in characterizing the physical driving forces involved in protein interactions (Pokarowski et al., 2005). However, it's important to note that most existing amino acid contact potentials are derived from experimentally determined protein structures and may not accurately capture the unique characteristics of IDPs and their interactions (Khatun et al., 2004). By deriving novel IDP-specific potentials through the analysis of their interactions, we can identify where these favorable interactions are likely to occur. IDPs can interact with each other using their different IDR segments. This information would help us prioritize the IDR segments in proteins responsible for their interaction dynamics.

In this work, we predicted a novel contact potential matrix between amino acid groups solely based on interactions between IDPs. To do this, we have employed a greedy statistical approach to compare the occurrence of amino acid groups between interacting and non-interacting IDPs. Following the identification of the IDP-specific contact potentials between amino acid groups, we visualized the heatmaps of contact maps between protein pairs to identify subregions that are important for the interactions.

7.2 Dataset

For the analysis, we employed interactions of IDPs that we curated for developing machine learning model to predict PPIs, as described in Section 6.3.1. In this study, we extract entire protein sequences from this dataset. To create a non-interacting PPI dataset, we

use the BRS-nonint program (Yu et al., 2010) again, providing the entire protein-protein interactions as input. As a result, we achieved a balanced ratio of 1:1 between negative and positive interactions.

7.3 Methods

The matrix we develop is a 5x5 contact potential matrix between amino acid groups, categorizing amino acids into five groups: positive, negative, polar, apolar, and aromatic (Table 7.1). The values in this matrix indicate how favorable is the interaction between every pair of amino acid groups. In this study, amino acid groups that occurred together in frequently interacting pairs compared to non-interacting IDPs are the favorable interactions for our IDP-IDP network.

We employ a greedy-type statistical approach to infer the contact potentials between amino acid groups. This approach involves initializing a random contact potential matrix for each iteration and evaluating its performance in reflecting the differences in the occurrences of amino acid groups between interacting and non-interacting proteins. At the end of the analysis, we select and retain the contact potential matrix that best represents the differences in the occurrences of amino acid groups.

First, we initialize a 5x5 contact potential matrix with random values ranging from -2 to 2 in multiples of 0.5. After extracting entire protein sequences of proteins from the dataset, we encode each protein residue in the protein sequence to one of the five categories based on their amino acid groups. Then, we use this randomly generated contact potential matrix to score interactions between proteins by constructing dot-matrix representations for each protein pair. These dot-matrix representations allow us to compare proteins along the vertical and horizontal axes. This dot-plot is a n x m matrix for each protein pair, m is the length of first protein sequence and n is the length of the second protein sequence in the protein pair. Given D is the dot-matrix, traditionally $D(i, j) = 1$, if the

Amino acid group	Amino Acids
Positive	K, R, H
Negative	D, E
Polar	S, T, C, Q, N
Apolar	A, L, V, I, M, G, P
Aromatic	W, F, Y

Table 7.1: Amino acid classification for contact potentials

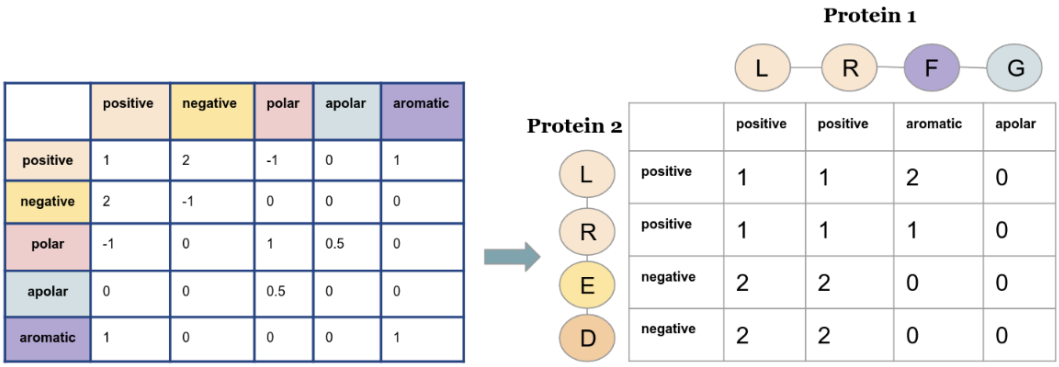


Figure 7.1: Scoring each protein sequence pair using random matrix. First, a scoring matrix with random values ranging from -2 to 2 is initiated. Then we use a dot-plot to represent each protein pair in our network and fill the matrix with the scores coming from the initial scoring matrix.

amino-acid position *i* in the first sequence is the same as the amino-acid at position *j* in the second sequence. In this scenario diagonals from top left to bottom right correspond to regions that are identical in both sequences. In our case, we fill the values in this matrix with the scores from this random scoring matrix based on the amino acid groups at each position. An example of how we fill a dot matrix based on the generated scoring matrix can be found in Figure 7.1.

To score the interaction between the two proteins, we analyze the dot-matrix diagonals. We sum the values along each diagonal larger than 12 cells and normalize the sum by the length of the diagonal. Diagonals with high scores indicate continuous regions where the amino acid groups have a higher propensity to interact. For each pair and corresponding dot matrix, we return the sum of the highest-scored diagonal as a final value representing the score of the interaction. After creating the dot matrix for each protein pair for both interacting and non-interacting protein pairs using the given scoring matrix, we compiled a list of scores for interacting and non-interacting IDPs. To test if there is a significant difference between the scores of these groups, we conduct the Mann-Whitney U test and return the p-value.

We repeated this entire process, generating a new random matrix each time and saving the p-value of the matrix and the matrix itself. We stopped the process after 300 iterations and returned the contact potentials matrix with the lowest p-value. The lower the p-value, the more significant the corresponding matrix is in separating interacting and non-interacting protein pairs. The overall workflow is illustrated in Figure 7.2. Traditionally, once the amino acid contact potentials are derived, they can be used in protein-protein

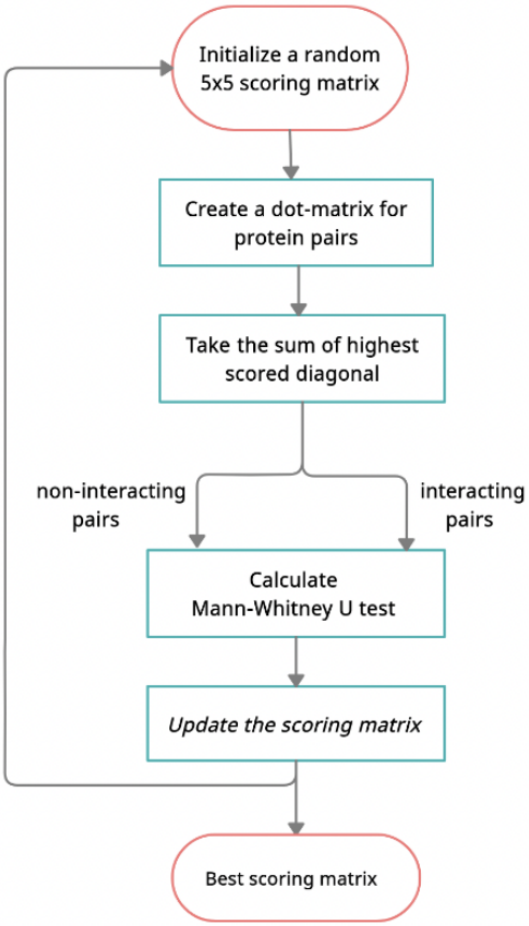


Figure 7.2: Workflow for deriving the contact matrix. Greedy type iterative approach to find best scoring matrix using Mann-Whitney U test.

docking simulations or protein-protein interaction interfaces. Similarly, we use the best contact potentials to score the interactions in our dataset and identify the interaction interfaces.

7.4 Results

In order to identify the most informative contact potential matrix between the occurrences of amino acid groups in IDP networks, we performed a statistical comparison of amino acid group occurrences in interacting IDPs versus non-interacting IDPs . The analysis reveals a scoring matrix that gives the lowest p-value, representing the contact potentials between distinct amino acid groups. The best IDP-specific scoring matrix with the lowest p-value (p-value<0.01), is presented in Table 7.2.

The table clearly shows that aromatic-aromatic, polar-polar, and negative-negative combinations have the highest potentials. This finding is consistent with previous studies that highlight the important role of aromatic interactions in protein-protein interfaces (Lanzarotti et al., 2020) and the involvement of aromatic residues in the binding and function of intrinsically disordered proteins (Espinoza-Fonseca, 2011).

Table 7.2: Best scoring matrix

	positive	negative	polar	apolar	aromatic
positive	0	0	-0.5	-1	-1
negative	0	2	-1.5	-0.5	-1.5
polar	-0.5	-1.5	2	1	0
apolar	-1	-0.5	1	-1	0.5
aromatic	-1	-1.5	0	0.5	2

Next, we sought to identify particular subregions within interactions characterized by high-value enrichments. To do this, we used the identified contact potential matrix to generate contact maps for the interactions in our dataset. By visualizing these contact maps as heatmaps, we can understand where the enriched pair of amino acid groups are located. Additionally, we overlaid the IDR regions of proteins onto the heatmap to identify which specific IDR regions displayed higher potentials. IDR regions are predicted using PONDR (Peng et al., 2006).

One of the protein pairs for which we analyzed the interaction is the interaction between FUS and SF3A2. To gain insights into the subregions with enriched amino acid groups, we generated a heatmap of the interaction matrix between these two protein sequences. The visualization of the interaction matrix heatmap indicates higher potentials between the N-terminal IDR region (depicted as the blue block in Figure 7.3A) of FUS and the C-terminal IDR region of SF3A2 (depicted as the red block in Figure 7.3A). To further

understand and potentially validate our observations, we used the structural prediction tool Alphafold2 to predict the structure of the same interaction (Figure 7.3B) (Jumper et al., 2021). The predicted structure suggests that the IDR segments identified through the heatmap analysis might be located at the interface of the protein complex.

Another protein interaction included in our analysis is the self-interaction of PHF13. PHF13 is a disordered protein that functions as a chromatin remodeler and has a self-aggregation ability. In their study, Kinkley lab demonstrated that deletion of ordered regions of PHF13 promotes LLPS (Chong et al., 2018; Rossi et al., 2022). Their research focuses on exploring the molecular mechanisms underlying PHF13 phase transitions.

To identify the subregions responsible for self-interaction and prioritize IDR regions in PHF13, we visualized the PHF13-PHF13 contact map using our method. PHF13 has five IDR regions in total. Our in silico prediction, based on the heatmap, suggests that IDR region 1 and IDR region 3 have higher potentials, suggesting that these regions might be important for the self-interaction of PHF13 (Figure 7.4). Interestingly, both IDR region 1 and IDR region 3 regions overlap with the PEST domain of the PHF13 protein. The PEST domain contains various phosphorylation sites. These IDR sites may be more important for the self-interaction of PHF13 which might influence the choice between LLPS and a polymer-polymer phase separation state.

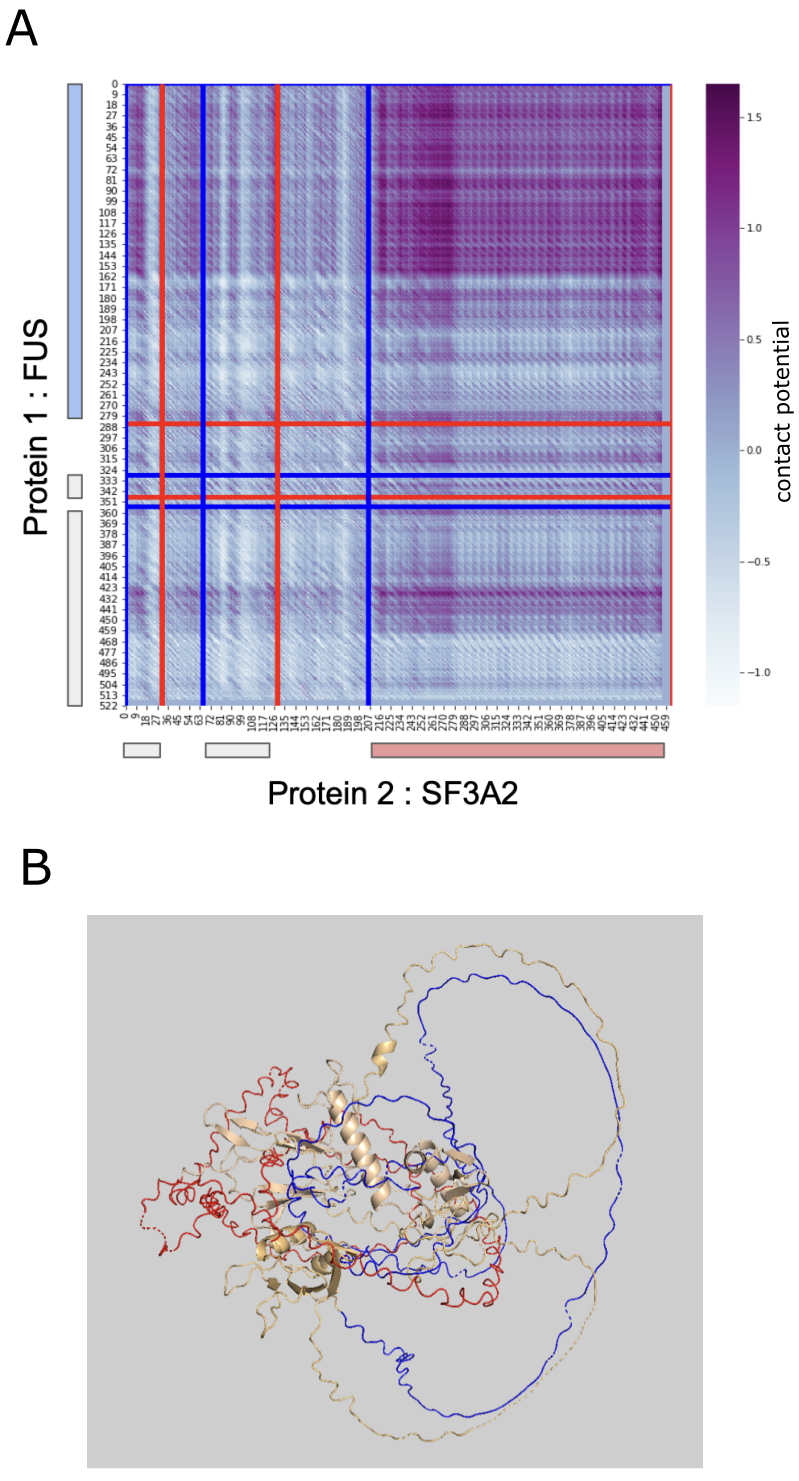


Figure 7.3: The interaction of between FUS and SF3A2. A) The heatmap visualization of the interaction matrix generated using our method. Blue lines represent the start of an IDR block, red lines represent the end of an IDR block B) Alphafold2 prediction of the interaction. The blue color represents the N-terminal IDR of FUS and the red color represents the terminal IDR of SF3A2

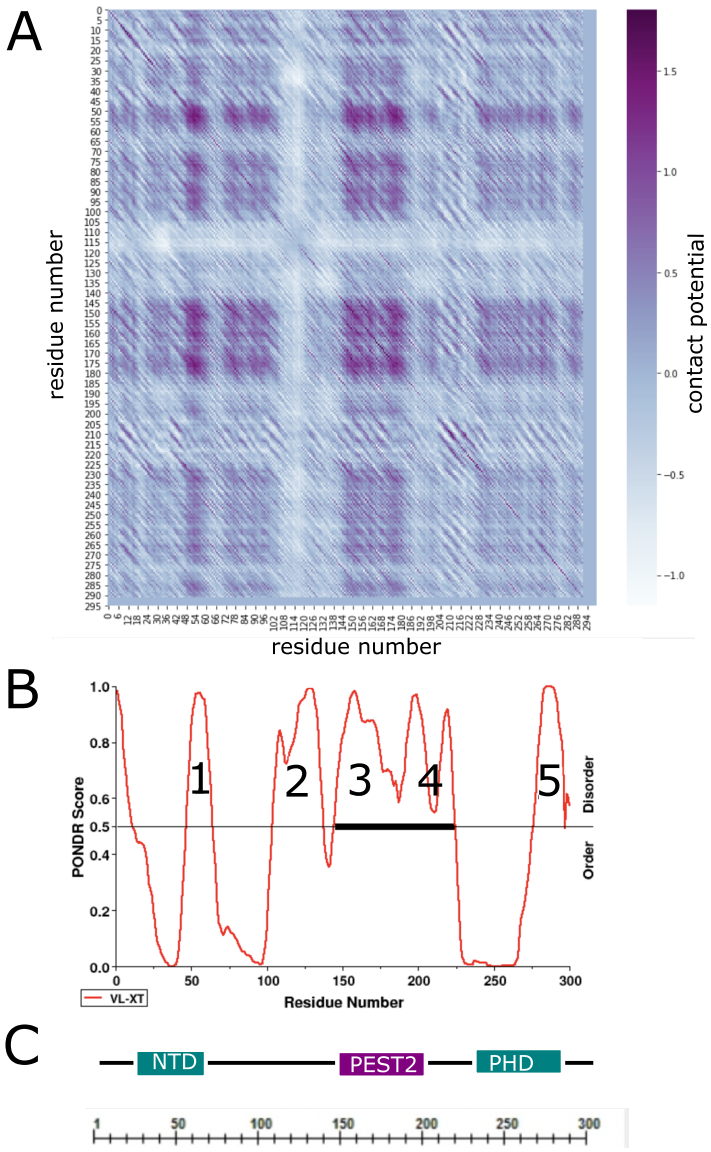


Figure 7.4: The self interaction map of PHF13. A) The heatmap of the contact matrix of PHF13-PHF13 computed using our method B) Disorder prediction of the PHF13 using PONDR C) Functional domains of PHF13.

7.5 Summary

This study presents a novel approach to investigate the interaction dynamics of IDPs by computing contact potentials between amino acid groups to capture the unique interaction characteristics of disordered proteins. The resulting contact potentials revealed that combinations involving aromatic-aromatic, polar-polar, and negative-negative amino acids showed the highest interaction potentials, consistent with previous findings highlighting the importance of aromatic interactions in protein-protein interfaces. To visualize these protein interactions, we generated heatmaps of contact matrices for protein pairs, thereby spotlighting potential interaction hotspots. These contact matrices can help to identify interaction interfaces within protein pairs. However, it's important to note that our study here represents a preliminary investigation into interaction sites, and further research is needed to deepen our understanding of these interactions and validate our findings.

In future analyses, these contact potentials can be integrated into machine learning models for PPI prediction. Another future direction is combining such statistical analysis with the protein structure prediction tool AlphaFold2 (Jumper et al., 2021). AlphaFold2 has recently been able to predict the structure of protein interactions. Our contact potentials inform us about amino acid preferences and interaction hotspots, while AlphaFold2 can provide detailed structural models of protein complexes and where the identified regions fall in the predicted structure. By combining different insights, we can enhance protein contact prediction and deepen our understanding of their interaction preferences.

8 DISCUSSION AND CONCLUSION

In this thesis, we first demonstrated several statistical methods for investigating IDRs based on their primary amino acid sequences, with the goal of identifying the sequence features responsible for their functions. In Chapters 3 and 4, we introduced statistical approaches with a specific focus on TFs to identify their associated protein amino acid features of IDRs underlie their functionality.

In Chapter 3, we demonstrated a novel approach to statistically model the occurrence of aromatic periodic blocks, which are known to play a role in the phase separation behavior of proteins. Using our approach, we quantified the periodicity of aromatic residues in human proteins. This was achieved by modeling the occurrences of aromatic residues using a Poisson process which enables efficient pinpointing of regions with significant periodicity in the human proteome. We identified proteins with significant periodic regions and found that many human TFs have periodic regions in their IDR regions. Our analysis indicates a connection between periodic aromatic blocks and IDRs, suggesting that these periodic regions could potentially serve as functional sites in IDRs responsible for the transcriptional activity of TFs. Furthermore, we demonstrated that these regions are particularly enriched in PLDs. It remains a challenge to understand how increasing or decreasing the degree of periodicity in these regions affects TF functions, including phase separation and transactivation capabilities. To gain insights into the evolution of these regions in proteins, we conducted a preliminary analysis to detect evolutionary changes in periodic protein sequences. Our results hint at the possibility that periodicity is a dynamic trait that changes across the evolution. Further research is needed to investigate more proteins with significant periodicity to understand how their periodic regions evolve by analyzing the periodicity scores across different species.

In Chapter 4, we continued to explore protein sequences of TF IDRs, aiming to gain an understanding of the compatibility of sequence features in the IDRs of TFs that bind to enhancer elements. We performed a statistical analysis to identify the amino acid groups that co-occur in IDRs of TFs binding together to enhancer elements. By comparing the prevalence of different combinations of amino acid groups between co-binding TFs and non co-binding TFs, we illustrated that TFs that co-bind together tend to be enriched with positively and negatively charged amino acids. Our preliminary results suggest that

the co-occurrence of positive and negative charged groups may play a role in the binding decisions between TFs, by attracting them to each other.

In the second part of the thesis, we focus on developing machine learning algorithm to predict PPIs of IDPs. In Chapter 5, we provided a background on machine learning for PPI prediction and highlighted the challenges associated with developing a machine learning model for pair prediction. These challenges set machine learning for pair prediction apart from traditional supervised machine learning models.

In Chapter 6, we introduced our machine learning model for PPI prediction. In the framework of our method, we addressed the challenges related to pair prediction and explained how we tackled these challenges in our approach. We defined two subproblems in pair prediction, namely the asymmetric and symmetric problems. Subsequently, we introduced our framework for PPI prediction, taking appropriate precautions for each problem type, and developing distinct random forest models tailored to each of these problems. To the best of our knowledge, this is the first machine learning model for PPI prediction where the architecture of the machine learning model was carefully designed according to the nature of the interaction problem.

We based our PPI prediction model on IDR sequences. To assess the predictive capabilities of IDRs, we developed a machine learning model that utilizes amino acid characteristics within IDRs for the purpose of predicting PPIs. Our findings indicated that IDR sequences are better predictors compared to non-IDR and entire sequences in predicting PPIs, supporting the notion that IDR sequences provide crucial information for interaction prediction.

In Chapter 7, we demonstrated a study to identify the subregions in proteins that are important for interactions by generating contact maps of IDPs. While this study offers insights into the significance of IDR regions in IDP interactions, further confirmatory research is required for the identified subregions.

Unfortunately, there isn't a universally accepted benchmark or standard approach for developing machine learning models for predicting PPIs, and various machine learning models have employed different approaches in their development. We believe our proposals regarding pair prediction schemes have enhanced our ability to understand and evaluate PPI models. Our testing and training schemes are rigorous and help maintain the accuracy of our work. We have designed both training and test scenarios to be as suitable as possible for the specific problem type at hand.

We showed that state-of-the-art methods perform much worse than originally reported when we evaluated them with correct test datasets without changing their training schemes used for the PPI prediction task. The differences between our results and originally reported performances can be explained by using more rigorous testing scenarios. This observation indicates once again the importance of evaluating the performance of the models separately on different test cases. Since pair prediction models involve many layers and have their own complexities, it becomes challenging to disentangle the impact of these approaches on different layers leading to final predictions. Additionally, this complexity also makes it difficult for users to select the machine learning model that best suits their needs.

Although our results hint at the capability of IDR to predict PPI networks, we cannot predict the test pairs that are completely new to the model with high accuracy. This raises the question if it is due to the architecture of our model. However, we also observed that other methods also struggle with solving the symmetric problem, despite being less cautious with their training strategies. This is also in line with the results by Dunham and colleagues (Dunham et al., 2021) who reported that most of the PPI tools perform much lower than originally reported. Future work could be aimed at improving the performance of the asymmetric model or symmetric model. Given that we demonstrated symmetric problem poses a significant challenge in the field of PPI, it becomes an interesting one to tackle.

In summary, we have applied various statistical methodologies to identify essential sequence features in IDR regions and showed how IDR sequences can be employed to predict PPIs. Our findings provide compelling evidence regarding the sequence characteristics of IDR sequences and their role in predicting interactions. By prioritizing these key sequence attributes, we can gain a better understanding of IDR behaviors. In summary, this thesis offers valuable insights into IDR sequence features across diverse functional contexts and contribute to the development of suitable machine learning models for addressing challenges in protein interaction prediction.

ABBREVIATIONS

IDRs intrinsically disordered regions

IDPs intrinsically disordered proteins

PPIs protein-protein interactions

PDB Protein Data Bank

SLiMs short linear motifs

MoRFs molecular recognition features

HIPPIE Human Integrated Protein-Protein Interaction rEference

LLPS liquid-liquid phase separated condansates

Y2H yeast two-hybrid screens

TFs Transcription factors

BioID proximity-dependent biotin identification

AD activation domain

DBD DNA binding domain

PLDs prion-like domains

PMF probability mass function

mRNA messenger RNA

K-S Kolmogorov–Smirnov test

LLPS Liquid-liquid phase separation

GO Gene Ontolog

PSSMs Position-specific scoring matrices

PAAC pseudo amino acid composition

DC dipeptide composition

AC Moreau-Broto autocorrelation

QSO Quasi-sequence-order

D-SCRIPT Deep sequence contact residue interaction prediction transfer

PTMs Post-translational modifications

GSEA Gene Set Enrichment Analysis

GO Gene Ontology

MS-PCI mass spectrometric protein complex identification

CV cross-validation

CNNs Convolutional Neural Networks

AUC The area under the ROC curve

MSA multiple sequence alignment

PSI-BLAST position-specific iterative BLAST

RF Random forest

LIST OF FIGURES

LIST OF TABLES

A APPENDIX

A.1 Supplementary Tables

Table A1: Feature encodings of IDPpi

Scale	W	F	Y	I	M	L	V	N	C	T	A	G	R	D	H	Q	K	S	E	P
Top-IDP	-0.884	-0.697	-0.51	-0.486	-0.397	-0.326	-0.121	0.007	0.02	0.059	0.06	0.166	0.18	0.192	0.303	0.318	0.586	0.341	0.736	0.987
B-value	0.938	0.934	0.981	0.977	0.963	0.982	0.968	1.022	0.939	0.998	0.994	1.018	1.026	1.022	0.967	1.041	1.029	1.025	1.052	1.05
FoldUnfold	28.48	27.18	25.93	25.71	24.82	25.36	23.93	18.49	23.52	19.81	19.89	17.11	21.03	17.41	21.72	19.23	18.19	17.67	17.46	17.43
Net charge	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	1	0	-1	0
DisProt	-0.465	-0.381	-0.427	-0.393	0.197	-0.26	-0.302	-0.106	-0.546	-0.116	0.042	0.095	0.211	0.127	-0.127	0.381	0.37	0.201	0.469	0.419

BIBLIOGRAPHY

Abbasi, Amir Ali, Debbie K. Goode, Saneela Amir, and Karl Heinz Grzeschik (May 2009). "Evolution and Functional Diversification of the GLI Family of Transcription Factors in Vertebrates." In: *Evolutionary Bioinformatics Online* 5 (5), p. 5. ISSN: 11769343. DOI: 10.4137/EBO.S2322.

Alanis-Lobato, Gregorio, Miguel A. Andrade-Navarro, and Martin H. Schaefer (Jan. 2017). "HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks." In: *Nucleic Acids Research* 45 (Database issue), p. D408. ISSN: 13624962. DOI: 10.1093/NAR/GKW985.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter (2002). "Analyzing Protein Structure and Function." In.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (Sept. 1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." In: *Nucleic acids research* 25 (17), pp. 3389–3402. ISSN: 0305-1048. DOI: 10.1093/NAR/25.17.3389.

Ardlie, Kristin G. et al. (May 2015). "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans." In: *Science (New York, N.Y.)* 348 (6235), pp. 648–660. ISSN: 1095-9203. DOI: 10.1126/SCIENCE.1262110.

Ayyadevara, Srinivas, Akshatha Ganne, Meenakshisundaram Balasubramaniam, and Robert J. Shmookler Reis (Jan. 2022). "Intrinsically disordered proteins identified in the aggregate proteome serve as biomarkers of neurodegeneration." In: *Metabolic Brain Disease* 37 (1), p. 147. ISSN: 15737365. DOI: 10.1007/S11011-021-00791-8.

Babu, M. Madan (Oct. 2016). "The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease." In: *Biochemical Society Transactions* 44 (5), pp. 1185–1200. ISSN: 14708752. DOI: 10.1042/BST20160172.

Babu, M. Madan, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer (June 2011). "Intrinsically disordered proteins: regulation and disease." In: *Current Opinion in Structural Biology* 21 (3), pp. 432–440. ISSN: 0959-440X. DOI: 10.1016/J.SBI.2011.03.011.

Bader, Gary D., Ian Donaldson, Cheryl Wolting, B. F.Francis Ouellette, Tony Pawson, and Christopher W.V. Hogue (Jan. 2001). "BIND—The Biomolecular Interaction Network

Database." In: *Nucleic Acids Research* 29 (1), p. 242. ISSN: 03051048. DOI: 10.1093/NAR/29.1.242.

Bah, Alaji and Julie D. Forman-Kay (Mar. 2016). "Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications." In: *The Journal of Biological Chemistry* 291 (13), p. 6696. ISSN: 1083351X. DOI: 10.1074/JBC.R115.695056.

Banani, Salman F., Hyun O. Lee, Anthony A. Hyman, and Michael K. Rosen (Feb. 2017). "Biomolecular condensates: organizers of cellular biochemistry." In: *Nature Reviews Molecular Cell Biology 2017 18:5* 18 (5), pp. 285–298. ISSN: 1471-0080. DOI: 10.1038/nrm.2017.7.

Banerjee, N. (2003). "Identifying cooperativity among transcription factors controlling the cell cycle in yeast." In: *Nucleic Acids Research* 31.23, 7024–7031. DOI: 10.1093/nar/gkg894.

Bateman, Alex et al. (Jan. 2021). "UniProt: the universal protein knowledgebase in 2021." In: *Nucleic Acids Research* 49 (D1), pp. D480–D489. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA1100.

Baughman, Hannah E.R., Dominic Narang, Wei Chen, Amalia C. Villagrán Suárez, Joan Lee, Maxwell J. Bachochin, Tristan R. Gunther, Peter G. Wolynes, and Elizabeth A. Komives (Sept. 2022). "An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFB p50/RelA." In: *Journal of Biological Chemistry* 298 (9), pp. 102349–102350. ISSN: 1083351X. DOI: 10.1016/j.jbc.2022.102349.

Bepler, Tristan and Bonnie Berger (Feb. 2019). "Learning protein sequence embeddings using information from structure." In: *7th International Conference on Learning Representations, ICLR 2019*.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne (Jan. 2000). "The Protein Data Bank." In: *Nucleic acids research* 28 (1), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/NAR/28.1.235.

Bigman, Lavi S., Junji Iwahara, and Yaakov Levy (July 2022). "Negatively Charged Disordered Regions are Prevalent and Functionally Important Across Proteomes." In: *Journal of molecular biology* 434 (14). ISSN: 1089-8638. DOI: 10.1016/J.JMB.2022.167660.

Blum, Matthias et al. (Jan. 2021). "The InterPro protein families and domains database: 20 years on." In: *Nucleic Acids Research* 49 (D1), pp. D344–D354. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA977.

Bogacka, Barbara (2004). *Chapter 3 Kolmogorov-Smirnov Tests*.

Boija, Ann, Isaac A. Klein, and Richard A. Young (Feb. 2021). "Biomolecular Condensates and Cancer." In: *Cancer cell* 39 (2), pp. 174–192. ISSN: 1878-3686. DOI: 10.1016/J.CCELL.2020.12.003.

Boija, Ann et al. (Dec. 2018). "Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains." In: *Cell* 175 (7), 1842–1855.e16. ISSN: 1097-4172. DOI: 10.1016/J.CELL.2018.10.042.

Bondos, Sarah E., A. Keith Dunker, and Vladimir N. Uversky (Dec. 2021). "On the roles of intrinsically disordered proteins and regions in cell communication and signaling." In: *Cell Communication and Signaling* 19 (1), pp. 1–9. ISSN: 1478811X. DOI: 10.1186/S12964-021-00774-3/TABLES/2.

Borgia, Alessandro et al. (Feb. 2018). "Extreme disorder in an ultrahigh-affinity protein complex." In: *Nature 2018 555:7694* 555 (7694), pp. 61–66. ISSN: 1476-4687. DOI: 10.1038/nature25762.

Breydo, Leonid and Vladimir N. Uversky (Nov. 2011). "Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases." In: *Metallomics* 3 (11), pp. 1163–1180. ISSN: 1756-5901. DOI: 10.1039/C1MT00106J.

Breydo, Leonid, Jessica W. Wu, and Vladimir N. Uversky (Feb. 2012). "-Synuclein misfolding and Parkinson's disease." In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1822 (2), pp. 261–285. ISSN: 0925-4439. DOI: 10.1016/J.BBADIS.2011.10.002.

Brodsky, Authors Sagie, Tamar Jana, Karin Mittelman, Michal Chapal, Divya Krishna Kumar, Miri Carmi, Sagie Brodsky, and Naama Barkai (2020). "Intrinsically Disordered Regions Direct Transcription Factor Innbsp;Vivo Binding Specificity." In: *Molecular Cell* 79, 459–471.e4. DOI: 10.1016/j.molcel.2020.05.032.

Broto, P., G. Moreau, and C. Vandycke (1984). "Molecular structures: perception, autocorrelation descriptor and sar studies. Autocorrelation descriptor." In: *undefined*.

Buchete, N. V., J. E. Straub, and D. Thirumalai (Jan. 2008). "Dissecting contact potentials for proteins: relative contributions of individual amino acids." In: *Proteins* 70 (1), pp. 119–130. ISSN: 1097-0134. DOI: 10.1002/PROT.21538.

Burkart-Solyom, Zsofia (Nov. 2014). "NMR methods for intrinsically disordered proteins : application to studies of NS5A protein of hepatitis C virus." In.

Bömmel, Alena van, Michael I. Love, Ho Ryun Chung, and Martin Vingron (Aug. 2018). "coTRaCTE predicts co-occurring transcription factors within cell-type specific enhancers." In: *PLOS Computational Biology* 14 (8), e1006372. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1006372.

Casadio, Rita, Pier Luigi Martelli, and Castrense Savojardo (Nov. 2022). "Machine learning solutions for predicting protein–protein interactions." In: *WIREs Computational Molecular Science* 12 (6). ISSN: 1759-0876. DOI: 10.1002/wcms.1618.

Chakrabarti, Pinak and Devlina Chakravarty (Apr. 2022). "Intrinsically disordered proteins/regions and insight into their biomolecular interactions." In: *Biophysical Chemistry* 283, p. 106769. ISSN: 0301-4622. DOI: 10.1016/J.BPC.2022.106769.

Chen, Kuan Hsi, Tsai Feng Wang, and Yuh Jyh Hu (June 2019). "Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme." In: *BMC Bioinformatics* 20 (1), pp. 1–17. ISSN: 14712105. DOI: 10.1186/S12859-019-2907-1/TABLES/10.

Chen, Wenqi, Shuang Wang, Tao Song, Xue Li, Peifu Han, and Changnan Gao (Dec. 2022). "DCSE:Double-Channel-Siamese-Ensemble model for protein protein interaction prediction." In: *BMC Genomics* 23 (1), pp. 1–14. ISSN: 14712164. DOI: 10.1186/S12864-022-08772-6/FIGURES/11.

Chen, Xue Wen and Mei Liu (Dec. 2005). "Prediction of protein–protein interactions using random decision forest framework." In: *Bioinformatics* 21 (24), pp. 4394–4400. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTI721.

Chong, Shasha and Mustafa Mir (June 2021). "Towards Decoding the Sequence-Based Grammar Governing the Functions of Intrinsically Disordered Protein Regions." In: p. 166724. ISSN: 00222836. DOI: 10.1016/j.jmb.2020.11.023.

Chong, Shasha et al. (July 2018). "Imaging dynamic and selective low-complexity domain interactions that control gene transcription." In: *Science (New York, N.Y.)* 361 (6400). ISSN: 1095-9203. DOI: 10.1126/SCIENCE.AAR2555.

Chothia, Cyrus and Joël Janin (1975). "Principles of protein-protein recognition." In: *Nature* 256 (5520), pp. 705–708. ISSN: 00280836. DOI: 10.1038/256705a0.

Chou, Kuo Chen (Nov. 2000). "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect." In: *Biochemical and biophysical research communications* 278 (2), pp. 477–483. ISSN: 0006-291X. DOI: 10.1006/BBRC.2000.3815.

– (May 2001). "Prediction of protein cellular attributes using pseudo-amino acid composition." In: *Proteins* 43 (3), pp. 246–255. ISSN: 0887-3585. DOI: 10.1002/PROT.1035.

Chowdhury, Aritra, Daniel Nettels, and Benjamin Schuler (May 2023). "Interaction Dynamics of Intrinsically Disordered Proteins from Single-Molecule Spectroscopy." In: *Annual review of biophysics* 52 (1). ISSN: 1936-1238. DOI: 10.1146/ANNUREV-BIOPHYS-101122-071930.

Cooper, Geoffrey M (2000). "The Cell." In: 8, pp. 103–108.

Czibula, Gabriela, Alexandra Ioana Albu, Maria Iuliana Bocicor, and Camelia Chira (June 2021). "Autoppi: An ensemble of deep autoencoders for protein–protein interaction prediction." In: *Entropy* 23 (6). ISSN: 10994300. DOI: 10.3390/E23060643.

DeForte, Shelly and Vladimir N. Uversky (Aug. 2016). "Order, Disorder, and Everything in Between." In: *Molecules* 21 (8). ISSN: 14203049. DOI: 10.3390/MOLECULES21081090.

Dodge, Yadolah (2008). "The concise encyclopedia of statistics." In: p. 616.

Dogan, Jakob, Stefano Gianni, and Per Jemth (Mar. 2014). "The binding mechanisms of intrinsically disordered proteins." In: *Physical Chemistry Chemical Physics* 16 (14), pp. 6323–6331. ISSN: 1463-9084. DOI: 10.1039/C3CP54226B.

Du, Xiuquan, Shiwei Sun, Changlin Hu, Yu Yao, Yuanting Yan, and Yanping Zhang (June 2017). "DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks." In: *Journal of Chemical Information and Modeling* 57 (6), pp. 1499–1510. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00028.

Dunham, Brandan and Madhavi K. Ganapathiraju (Dec. 2021). "Benchmark Evaluation of Protein–Protein Interaction Prediction Algorithms." In: *Molecules* 27 (1), p. 41. ISSN: 1420-3049. DOI: 10.3390/molecules27010041.

Dyson, H. Jane and Peter E. Wright (Mar. 2005). "Intrinsically unstructured proteins and their functions." In: *Nature reviews. Molecular cell biology* 6 (3), pp. 197–208. ISSN: 1471-0072. DOI: 10.1038/NRM1589.

Emenecker, Ryan J., Daniel Griffith, and Alex S. Holehouse (Oct. 2021). "Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure." In: *Biophysical Journal* 120 (20), pp. 4312–4319. ISSN: 0006-3495. DOI: 10.1016/J.BPJ.2021.08.039.

Erijman, Ariel, Lukasz Kozlowski, Salma Sohrabi-Jahromi, James Fishburn, Linda Warfield, Jacob Schreiber, William S. Noble, Johannes Söding, and Steven Hahn (June 2020). "A High-Throughput Screen for Transcription Activation Domains Reveals Their Sequence Features and Permits Prediction by Deep Learning." In: *Molecular cell* 78 (5), 890–902.e6. ISSN: 1097-4164. DOI: 10.1016/J.MOLCEL.2020.04.020.

Espinoza-Fonseca, L. Michel (Dec. 2011). "Aromatic residues link binding and function of intrinsically disordered proteins." In: *Molecular BioSystems* 8 (1), pp. 237–246. ISSN: 1742-2051. DOI: 10.1039/C1MB05239J.

Eyal, Eran (2005). *Predicting side chain conformations and stability of mutants using contact surface areas.*

Ferreon, C, Allan Chris, M Ferreon, Rakesh Trivedi, and Hampapathalu Adimurthy Nagarajaram (Nov. 2022). "Intrinsically Disordered Proteins: An Overview." In: *International Journal of Molecular Sciences 2022, Vol. 23, Page 14050* 23 (22), p. 14050. ISSN: 1422-0067. DOI: 10.3390/IJMS232214050.

Fink, Anthony L. (2005). "Natively unfolded proteins." In: *Current opinion in structural biology* 15 (1), pp. 35–41. ISSN: 0959-440X. DOI: 10.1016/J.SBI.2005.01.002.

Fukuchi, Satoshi, Kazuo Hosoda, Keiichi Homma, Takashi Gojobori, and Ken Nishikawa (2011). "Binary classification of protein molecules into intrinsically disordered and ordered segments." In: *BMC Structural Biology* 11. ISSN: 14726807. DOI: 10.1186/1472-6807-11-29.

Ghadie, Mohamed Ali and Yu Xia (Apr. 2022). "Are transient protein-protein interactions more dispensable?" In: *PLOS Computational Biology* 18 (4), e1010013. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1010013.

Grantham, R. (1974). "Amino acid difference formula to help explain protein evolution." In: *Science (New York, N.Y.)* 185 (4154), pp. 862–864. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.185.4154.862.

Grigoriev, Andrei (July 2003). "On the number of protein-protein interactions in the yeast proteome." In: *Nucleic Acids Res.* 31 (14), pp. 4157–4161. ISSN: 03051048. DOI: 10.1093/nar/gkg466.

Guo, Yanzhi, Lezheng Yu, Zhining Wen, and Menglong Li (May 2008). "Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences." In: *Nucleic acids research* 36 (9), pp. 3025–3030. ISSN: 1362-4962. DOI: 10.1093/NAR/GKN159.

Hamp, Tobias and Burkhard Rost (June 2015). "Evolutionary profiles improve protein–protein interaction prediction from sequence." In: *Bioinformatics* 31 (12), pp. 1945–1950. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTV077.

Hashemifar, Somaye, Behnam Neyshabur, Aly A Khan, and Jinbo Xu (2018). "Predicting protein-protein interactions through sequence-based deep learning." In: DOI: 10.1093/bioinformatics/bty573.

Hnisz, Denes, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp (Mar. 2017). "A Phase Separation Model for Transcriptional Control." In: *Cell* 169 (1), pp. 13–23. ISSN: 10974172. DOI: 10.1016/J.CELL.2017.02.007.

Holehouse, Alex S, Garrett M Ginell, Daniel Griffith, and Elvan Böke (2021). "Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates." In: 60, p. 34. DOI: 10.1021/acs.biochem.1c00465.

Holland, Jack and Gevorg Grigoryan (Apr. 2022). "Structure-conditioned amino-acid couplings: How contact geometry affects pairwise sequence preferences." In: *Protein Science* 31 (4), pp. 900–917. ISSN: 1469-896X. DOI: 10.1002/PRO.4280.

Hu, Xiaotian, Cong Feng, Tianyi Ling, and Ming Chen (Jan. 2022). "Deep learning frameworks for protein–protein interaction prediction." In: *Computational and Structural*

Biotechnology Journal 20, pp. 3223–3233. ISSN: 2001-0370. DOI: 10.1016/J.CSBJ.2022.06.025.

Hyman, Anthony A., Christoph A. Weber, and Frank Jülicher (2014). "Liquid-liquid phase separation in biology." In: *Annual review of cell and developmental biology* 30, pp. 39–58. ISSN: 1530-8995. DOI: 10.1146/ANNUREV-CELLBIO-100913-013325.

Ibarra, Ignacio L., Nele M. Hollmann, Bernd Klaus, Sandra Augsten, Britta Velten, Janosch Hennig, and Judith B. Zaugg (Jan. 2020). "Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions." In: *Nature Communications 2020 11:1* 11 (1), pp. 1–16. ISSN: 2041-1723. DOI: 10.1038/S41467-019-13888-7.

Ieremie, Ioan, Rob M. Ewing, and Mahesan Niranjan (Apr. 2022). "TransformerGO: predicting protein-protein interactions by modelling the attention between sets of gene ontology terms." In: *Bioinformatics (Oxford, England)* 38 (8), pp. 2269–2277. ISSN: 1367-4811. DOI: 10.1093/BIOINFORMATICS/BTAC104.

Jo, Yongsang, Jinyoung Jang, Daesun Song, Hyoin Park, and Yongwon Jung (2022). "Determinants for intrinsically disordered protein recruitment into phase-separated protein condensates." In: *Chemical Science* 13 (2), pp. 522–530. ISSN: 2041-6520. DOI: 10.1039/D1SC05672G.

Jones, Susan and Janet M. Thornton (Jan. 1996). "Principles of protein-protein interactions." In: *Proceedings of the National Academy of Sciences of the United States of America* 93 (1), pp. 13–20. ISSN: 0027-8424. DOI: 10.1073/PNAS.93.1.13.

Joseph, V. Roshan (Aug. 2022). "Optimal ratio for data splitting." In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15 (4), pp. 531–538. ISSN: 1932-1872. DOI: 10.1002/SAM.11583.

Jumper, John et al. (July 2021). "Highly accurate protein structure prediction with AlphaFold." In: *Nature 2021 596:7873* 596 (7873), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/S41586-021-03819-2.

Kawashima, Shuichi and Minoru Kanehisa (Jan. 2000). "AAindex: Amino Acid index database." In: *Nucleic Acids Research* 28 (1), pp. 374–374. ISSN: 0305-1048. DOI: 10.1093/NAR/28.1.374.

Keegan, Liam, Grace Gill, and Mark Ptashne (1986). "Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein." In: *Science (New York, N.Y.)* 231 (4739), pp. 699–704. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.3080805.

Kendall, K and Maurice George (Feb. 2008). "Kolmogorov–Smirnov Test." In: *The Concise Encyclopedia of Statistics*, pp. 283–287. DOI: 10.1007/978-0-387-32833-1_214.

Khatun, Jainab, Sagar D. Khare, and Nikolay V. Dokholyan (Mar. 2004). "Can Contact Potentials Reliably Predict Stability of Proteins?" In: *Journal of Molecular Biology* 336 (5), pp. 1223–1238. ISSN: 0022-2836. DOI: 10.1016/J.JMB.2004.01.002.

Kosol, Simone, Sara Contreras-Martos, Cesyen Cedeño, and Peter Tompa (Sept. 2013). "Structural characterization of intrinsically disordered proteins by NMR spectroscopy." In: *Molecules* 18 (9), pp. 10802–10828. ISSN: 14203049. DOI: 10.3390/MOLECULES180910802.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura (June 2018). "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." In: *Molecular biology and evolution* 35 (6), pp. 1547–1549. ISSN: 1537-1719. DOI: 10.1093/MOLBEV/MSY096.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch (Feb. 2018). "The Human Transcription Factors." In: *Cell* 172 (4), pp. 650–665. ISSN: 10974172. DOI: 10.1016/J.CELL.2018.01.029/ATTACHMENT/EDE37821-FD6F-41B7-9A0E-9D5410855AE6/MMC2.XLSX.

Lancaster, Alex K., Andrew Nutter-Upham, Susan Lindquist, and Oliver D. King (Sept. 2014). "PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition." In: *Bioinformatics* 30 (17), pp. 2501–2502. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTU310.

Lanzarotti, Esteban, Lucas A. Defelipe, Marcelo A. Marti, and Adrián G. Turjanski (May 2020). "Aromatic clusters in protein-protein and protein-drug complexes." In: *Journal of cheminformatics* 12 (1). ISSN: 1758-2946. DOI: 10.1186/S13321-020-00437-4.

Latchman, David S. (2002). "Gene regulation : a eukaryotic perspective." In: p. 323.

Lee, Robin Van Der et al. (July 2014). "Classification of intrinsically disordered regions and proteins." In: *Chemical reviews* 114 (13), pp. 6589–6631. ISSN: 1520-6890. DOI: 10.1021/CR400525M.

Levy, Emmanuel D. and Jose B. Pereira-Leal (June 2008). "Evolution and dynamics of protein interactions and networks." In: *Current Opinion in Structural Biology* 18 (3), pp. 349–357. ISSN: 0959-440X. DOI: 10.1016/J.SBI.2008.03.003.

Li, Weizhong and Adam Godzik (July 2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." In: *Bioinformatics (Oxford, England)* 22 (13), pp. 1658–1659. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTL158.

Li, Yiwei and Lucian Ilie (2020). "Predicting Protein-Protein Interactions Using SPRINT." In: *Methods in molecular biology (Clifton, N.J.)* 2074, pp. 1–11. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-9873-9_1.

Licata, Luana et al. (Jan. 2012). "MINT, the molecular interaction database: 2012 update." In: *Nucleic Acids Research* 40 (D1), pp. D857–D861. ISSN: 0305-1048. DOI: 10.1093/NAR/GKR930.

Lindorff-Larsen, Kresten and Birthe B. Kragelund (Oct. 2021). "On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins." In: *Journal of molecular biology* 433 (20). ISSN: 1089-8638. DOI: 10.1016/J.JMB.2021.167196.

Livingstone, Craig D. and Geoffrey J. Barton (1993). "Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation." In: *Computer applications in the biosciences : CABIOS* 9 (6), pp. 745–756. ISSN: 0266-7061. DOI: 10.1093/BIOINFORMATICS/9.6.745.

Lu, Alex X., Amy X. Lu, Iva Pritišanac, Taraneh Zarin, Julie D. Forman-Kay, and Alan M. Moses (June 2022). "Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning." In: *PLOS Computational Biology* 18 (6), e1010238. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1010238.

Lyon, Andrew S., William B. Peeples, and Michael K. Rosen (Mar. 2021). "A framework for understanding the functions of biomolecular condensates across scales." In: *Nature reviews. Molecular cell biology* 22 (3), pp. 215–235. ISSN: 1471-0080. DOI: 10.1038/S41580-020-00303-Z.

Ma, Liang, Zeyue Gao, Jiegen Wu, Bijunyao Zhong, Yuchen Xie, Wen Huang, and Yihan Lin Correspondence (2021). "Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics." In: DOI: 10.1016/j.molcel.2021.01.031.

Marchler-Bauer, Aron et al. (Jan. 2015). "CDD: NCBI's conserved domain database." In: *Nucleic Acids Research* 43 (D1), pp. D222–D226. ISSN: 0305-1048. DOI: 10.1093/NAR/GKU1221.

Martin, Erik W., Alex S. Holehouse, Ivan Peran, Mina Farag, J. Jeremias Incicco, Anne Bremer, Christy R. Grace, Andrea Soranno, Rohit V. Pappu, and Tanja Mittag (Feb. 2020). "Valence and patterning of aromatic residues determine the phase behavior of prion-like domains." In: *Science (New York, N.Y.)* 367 (6478), pp. 694–699. ISSN: 1095-9203. DOI: 10.1126/SCIENCE.AAW8653.

Martin, Shawn, Diana Roe, and Jean Loup Faulon (Jan. 2005). "Predicting protein-protein interactions using signature products." In: *Bioinformatics (Oxford, England)* 21 (2), pp. 218–226. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTH483.

Martinelli, Anne H.S., Fernanda C. Lopes, Elisa B.O. John, Célia R. Carlini, and Rodrigo Ligabue-Braun (Mar. 2019). "Modulation of Disordered Proteins with a Focus on Neu-

rodegenerative Diseases and Other Pathologies." In: *International Journal of Molecular Sciences* 20 (6). ISSN: 14220067. DOI: 10.3390/IJMS20061322.

Massey, Frank J. (Mar. 1951). "The Kolmogorov-Smirnov Test for Goodness of Fit." In: *Journal of the American Statistical Association* 46 (253), p. 68. ISSN: 01621459. DOI: 10.2307/2280095.

Masulli, F (2008). "Marketa Zvelebil and Jeremy Baum, Understanding Bioinformatics , Garland Science—Taylor Francis Group (2007) 800 pp., Paperback, ISBN: 0-8153-4024-9/ISBN-13: 978-0-8153-4024-9, 111.00 US (Amazon.com), 41.99 £ (amazon. UK), 64.99 EUR (amazon.de)." In: *Computer Methods and Programs in Biomedicine* 91 (2), pp. 182–182. ISSN: 01692607.

McCann, J. F. (2016). "Arrivals and waiting times." In.

Mensah, Martin A. et al. (Feb. 2023). "Aberrant phase separation and nucleolar dysfunction in rare genetic diseases." In: *Nature* 614 (7948), pp. 564–571. ISSN: 1476-4687. DOI: 10.1038/S41586-022-05682-1.

Mewes, H. W., D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil (Jan. 2002). "MIPS: a database for genomes and protein sequences." In: *Nucleic Acids Research* 30 (1), p. 31. ISSN: 03051048. DOI: 10.1093/NAR/30.1.31.

Meyer, Katrina, Marieluise Kirchner, Bora Uyar, Sebastian Diecke, Juan M Pascual, and Matthias Selbach Correspondence (2018). "Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs." In: *Cell* 175, 239–253.e17. DOI: 10.1016/j.cell.2018.08.019.

Mistry, Jaina et al. (Jan. 2021). "Pfam: The protein families database in 2021." In: *Nucleic Acids Research* 49 (D1), pp. D412–D419. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA913.

Mooney, Catherine, Gianluca Pollastri, Denis C. Shields, and Niall J. Haslam (Jan. 2012). "Prediction of Short Linear Protein Binding Regions." In: *Journal of Molecular Biology* 415 (1), pp. 193–204. ISSN: 0022-2836. DOI: 10.1016/J.JMB.2011.10.025.

Morris, Owen Michael, James Hilary Torpey, and Rivka Leah Isaacson (Oct. 2021). "Intrinsically disordered proteins: modes of binding with emphasis on disordered domains." In: *Open Biology* 11 (10). ISSN: 20462441. DOI: 10.1098/RSOB.210222.

Mu, Zengchao, Ting Yu, Xiaoping Liu, Hongyu Zheng, Leyi Wei, and Juntao Liu (Dec. 2021). "FEGS: a novel feature extraction model for protein sequences and its applications." In: *BMC Bioinformatics* 22 (1), pp. 1–15. ISSN: 14712105. DOI: 10.1186/S12859-021-04223-3/TABLES/1.

Müller-Späth, Sonja, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rüegger, Luc Reymond, Daniel Nettels, and Benjamin Schuler (Aug. 2010). "Charge interactions can dominate the dimensions of intrinsically disordered proteins." In: *Proceedings*

of the National Academy of Sciences 107 (33), pp. 14609–14614. ISSN: 0027-8424. DOI: 10.1073/pnas.1001743107.

Nahlé, Sarah, Laura Quirion, Jonathan Boulais, Halil Bagci, Denis Faubert, Anne Claude Gingras, and Jean François Côté (Mar. 2022). "Defining the interactomes of proteins involved in cytoskeletal dynamics using high-throughput proximity-dependent biotinylation in cellulo." In: *STAR Protocols* 3 (1), p. 101075. ISSN: 26661667. DOI: 10.1016/j.xpro.2021.101075.

Nooren, Irene M.A. and Janet M. Thornton (July 2003). "Diversity of protein-protein interactions." In: *The EMBO journal* 22 (14), pp. 3486–3492. ISSN: 0261-4189. DOI: 10.1093/EMBOJ/CDG359.

Oldfield, Christopher J. and A. Keith Dunker (2014). "Intrinsically disordered proteins and intrinsically disordered protein regions." In: *Annual review of biochemistry* 83, pp. 553–584. ISSN: 1545-4509. DOI: 10.1146/ANNUREV-BIOCHEM-072711-164947.

O'shea, Keiron and Ryan Nash (2015). "An Introduction to Convolutional Neural Networks." In.

Oughtred, Rose et al. (Jan. 2021). "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions." In: *Protein Science : A Publication of the Protein Society* 30 (1), p. 187. ISSN: 1469896X. DOI: 10.1002/PRO.3978.

Park, Yungki and Edward M. Marcotte (Nov. 2011). "Revisiting the negative example sampling problem for predicting protein–protein interactions." In: *Bioinformatics* 27 (21), pp. 3024–3028. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTR514.

– (Dec. 2012). "A flaw in the typical evaluation scheme for pair-input computational predictions." In: *Nature methods* 9 (12), p. 1134. ISSN: 15487091. DOI: 10.1038/NMETH.2259.

Peng, Kang, Radivojac P, Vucetic S, Dunker AK, and Obradovic Z (2006). "Length-dependent prediction of protein intrinsic disorder." In: *BMC bioinformatics* 7 (1), pp. 208–212. ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-208.

Peri, Suraj et al. (Jan. 2004). "Human protein reference database as a discovery resource for proteomics." In: *Nucleic Acids Research* 32 (Database issue), p. D497. ISSN: 03051048. DOI: 10.1093/NAR/GKH070.

Perkins, James R., Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, and Christine Orengo (Oct. 2010). "Transient Protein-Protein Interactions: Structural, Functional, and Network Properties." In: *Structure* 18 (10), pp. 1233–1243. ISSN: 0969-2126. DOI: 10.1016/j.str.2010.08.007.

Perovic, Vladimir, Neven Sumonja, Lindsey A. Marsh, Sandro Radovanovic, Milan Vukicevic, Stefan G. E. Roberts, and Nevena Veljkovic (Dec. 2018). "IDPpi: Protein-Protein

Interaction Analyses of Human Intrinsically Disordered Proteins." In: *Scientific Reports* 8 (1), p. 10563. ISSN: 2045-2322. DOI: 10.1038/s41598-018-28815-x.

Piovesan, Damiano et al. (Jan. 2017). "DisProt 7.0: a major update of the database of disordered proteins." In: *Nucleic acids research* 45 (D1), pp. D219–D227. ISSN: 1362-4962. DOI: 10.1093/NAR/GKW1056.

Pokarowski, Piotr, Andrzej Kloczkowski, Robert L. Jernigan, Neha S. Kothari, Maria Pokarowska, and Andrzej Kolinski (Apr. 2005). "Inferring Ideal Amino Acid Interaction Forms From Statistical Protein Contact Potentials." In: *Proteins* 59 (1), p. 49. ISSN: 08873585. DOI: 10.1002/PROT.20380.

Qi, Yanjun and William Stafford Noble (2011). "Protein interaction networks: Protein domain interaction and protein function prediction." In.

Radivojac, Predrag, Lilia M. Iakoucheva, Christopher J. Oldfield, Zoran Obradovic, Vladimir N. Uversky, and A. Keith Dunker (2007). "Intrinsic disorder and functional proteomics." In: *Biophysical journal* 92 (5), pp. 1439–1456. ISSN: 1542-0086. DOI: 10.1529/BIOPHYSJ.106.094045.

Ramachandran, Kandethody M. and Chris P. Tsokos (Sept. 2014). "Mathematical Statistics with Applications in R: Second Edition." In: *Mathematical Statistics with Applications in R: Second Edition*, pp. 1–800. DOI: 10.1016/C2012-0-07341-3.

Ravarani, Charles NJ, Tamara Y Erkina, Greet De Baets, Daniel C Dudman, Alexandre M Erkine, and M Madan Babu (May 2018). "High-throughput discovery of functional disordered regions: investigation of transactivation domains." In: *Molecular Systems Biology* 14 (5), e8190. ISSN: 1744-4292. DOI: 10.15252/MSB.20188190.

Reiter, Franziska, Sebastian Wienerroither, and Alexander Stark (Apr. 2017). "Combinatorial function of transcription factors and cofactors." In: *Current opinion in genetics development* 43, pp. 73–81. ISSN: 1879-0380. DOI: 10.1016/J.GDE.2016.12.007.

Reményi, Attila, Hans R. Schöler, and Matthias Wilmanns (Sept. 2004). "Combinatorial control of gene expression." In: *Nature structural molecular biology* 11 (9), pp. 812–815. ISSN: 1545-9993. DOI: 10.1038/NSMB820.

Ross, Arun (2009). "Fusion, Feature-Level." In: *Encyclopedia of Biometrics*, pp. 597–602. DOI: 10.1007/978-0-387-73003-5_157.

Rossi, Francesca et al. (Mar. 2022). "Connecting the Dots: PHF13 and cohesin promote polymer-polymer phase separation of chromatin into chromosomes." In: *bioRxiv*, p. 2022.03.04.482956. DOI: 10.1101/2022.03.04.482956.

Sabari, Benjamin R., Alessandra Dall'Agnese, and Richard A. Young (Nov. 2020). "Biomolecular Condensates in the Nucleus." In: *Trends in biochemical sciences* 45 (11), pp. 961–977. ISSN: 0968-0004. DOI: 10.1016/J.TIBS.2020.06.007.

Sabari, Benjamin R. et al. (July 2018). "Coactivator condensation at super-enhancers links phase separation and gene control." In: *Science (New York, N.Y.)* 361 (6400). ISSN: 1095-9203. DOI: 10.1126/SCIENCE.AAR3958.

Sanborn, Adrian L., Benjamin T. Yeh, Jordan T. Feigerle, Cynthia V. Hao, Raphael J.L. Townshend, Erez Lieberman Aiden, Ron O. Dror, and Roger D. Kornberg (Apr. 2021). "Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to mediator." In: *eLife* 10. ISSN: 2050084X. DOI: 10.7554/ELIFE.68068.

Sarker, Iqbal H. (May 2021). "Machine Learning: Algorithms, Real-World Applications and Research Directions." In: *SN Computer Science* 2 (3), pp. 1–21. ISSN: 26618907. DOI: 10.1007/S42979-021-00592-X/FIGURES/11.

Schneider, G. and P. Wrede (1994). "The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site." In: *Biophysical journal* 66 (2 Pt 1), pp. 335–344. ISSN: 0006-3495. DOI: 10.1016/S0006-3495(94)80782-9.

Scholes, Natalie S. and Robert O.J. Weinzierl (May 2016). "Molecular Dynamics of "Fuzzy" Transcriptional Activator-Coactivator Interactions." In: *PLOS Computational Biology* 12 (5), e1004935. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1004935.

Shen, Juwen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang (Mar. 2007). "Predicting protein-protein interactions based only on sequences information." In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (11), pp. 4337–4341. ISSN: 0027-8424. DOI: 10.1073/PNAS.0607879104.

Sievers, Fabian et al. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." In: *Molecular systems biology* 7. ISSN: 1744-4292. DOI: 10.1038/MSB.2011.75.

Sigrist, Christian J.A., Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher (2002). "PROSITE: a documented database using patterns and profiles as motif descriptors." In: *Briefings in bioinformatics* 3 (3), pp. 265–274. ISSN: 1467-5463. DOI: 10.1093/BIB/3.3.265.

Sinharay, S. (Jan. 2010). "Discrete Probability Distributions." In: *International Encyclopedia of Education, Third Edition*, pp. 132–134. ISSN: 16166361. DOI: 10.1016/B978-0-08-044894-7.01721-8.

Sledzieski, Samuel, Rohit Singh, Lenore Cowen, and Bonnie Berger (Oct. 2021). "D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions." In: *Cell Systems* 12 (10), 969–982.e6. ISSN: 24054712. DOI: 10.1016/j.cels.2021.08.010.

Soleymani, Farzan, Eric Paquet, Herna Viktor, Wojtek Michalowski, and Davide Spinello (Jan. 2022). "Protein–protein interaction prediction with deep learning: A comprehensive review." In: *Computational and Structural Biotechnology Journal* 20, pp. 5316–5341. ISSN: 2001-0370. DOI: 10.1016/J.CSBJ.2022.08.070.

Staller, Max V., Alex S. Holehouse, Devjanee Swain-Lenz, Rahul K. Das, Rohit V. Pappu, and Barak A. Cohen (Apr. 2018). "A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain." In: *Cell systems* 6 (4), 444–455.e6. ISSN: 2405-4712. DOI: 10.1016/J.CELS.2018.01.015.

Stampfel, Gerald, Tomáš Kazmar, Olga Frank, Sebastian Wienerroither, Franziska Reiter, and Alexander Stark (Dec. 2015). "Transcriptional regulators form diverse groups with context-dependent regulatory functions." In: *Nature* 528 (7580), pp. 147–151. ISSN: 1476-4687. DOI: 10.1038/NATURE15545.

Stephens, M. A. (1992). "Introduction to Kolmogorov (1933) On the Empirical Determination of a Distribution." In: pp. 93–105. DOI: 10.1007/978-1-4612-4380-9_9.

Stryer, Lubert (2000). *Biochemistry*. 4th ed. Eighth printing.

Subramanian, Aravind et al. (Oct. 2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (43), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/PNAS.0506580102.

Sugase, Kenji, H. Jane Dyson, and Peter E. Wright (June 2007). "Mechanism of coupled folding and binding of an intrinsically disordered protein." In: *Nature* 447 (7147), pp. 1021–1025. ISSN: 1476-4687. DOI: 10.1038/NATURE05858.

Szklarczyk, Damian et al. (Jan. 2015). "STRING v10: protein-protein interaction networks, integrated over the tree of life." In: *Nucleic acids research* 43 (Database issue), pp. D447–D452. ISSN: 1362-4962. DOI: 10.1093/NAR/GKU1003.

Tompa, Peter (June 2011). "Unstructural biology coming of age." In: *Current opinion in structural biology* 21 (3), pp. 419–425. ISSN: 1879-033X. DOI: 10.1016/J.SBI.2011.03.012.

Toro, Noemi del et al. (Jan. 2022). "The IntAct database: efficient access to fine-grained molecular interaction data." In: *Nucleic Acids Research* 50 (D1), pp. D648–D653. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAB1006.

Trivedi, Rakesh and Hampapathalu Adimurthy Nagarajaram (Nov. 2022). "Intrinsically Disordered Proteins: An Overview." In: *International journal of molecular sciences* 23 (22). ISSN: 1422-0067. DOI: 10.3390/IJMS232214050.

Tsoi, Phoebe S., My Diem Quan, Josephine C. Ferreon, and Allan Chris M. Ferreon (Feb. 2023). "Aggregation of Disordered Proteins Associated with Neurodegeneration." In:

International journal of molecular sciences 24 (4). ISSN: 1422-0067. DOI: 10.3390/IJMS24043380.

Uversky, Vladimir N. (May 2013). "Intrinsic disorder-based protein interactions and their modulators." In: *Current pharmaceutical design* 19 (23), pp. 4191–4213. ISSN: 1873-4286. DOI: 10.2174/1381612811319230005.

– (2020). "Analyzing IDPs in Interactomes." In: *Methods in molecular biology (Clifton, N.J.)* 2141, pp. 895–945. ISSN: 1940-6029. DOI: 10.1007/978-1-0716-0524-0_46.

Uversky, Vladimir N., Christopher J. Oldfield, and A. Keith Dunker (2008). "Intrinsically disordered proteins in human diseases: introducing the D2 concept." In: *Annual review of biophysics* 37, pp. 215–246. ISSN: 1936-122X. DOI: 10.1146/ANNUREV.BIOPHYS.37.032807.125924.

Vacic, Vladimir, Phineus R.L. Markwick, Christopher J. Oldfield, Xiaoyue Zhao, Chad Haynes, Vladimir N. Uversky, and Lilia M. Iakoucheva (2012). "Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder." In: *PLoS Computational Biology* 8 (10). ISSN: 15537358. DOI: 10.1371/JOURNAL.PCBI.1002709.

Vandel, Jimmy, Océane Cassan, Sophie Lèbre, Charles Henri Lecellier, and Laurent Bréhélin (Feb. 2019). "Probing transcription factor combinatorics in different promoter classes and in enhancers." In: *BMC Genomics* 20 (1). ISSN: 14712164. DOI: 10.1186/S12864-018-5408-0.

Varnaitė, Renata and Stuart A. MacNeill (Oct. 2016). "Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID." In: *Proteomics* 16 (19), pp. 2503–2518. ISSN: 1615-9861. DOI: 10.1002/PMIC.201600123.

Venkatesan, Kavitha et al. (2009). "An empirical framework for binary interactome mapping." In: *Nature methods* 6 (1), pp. 83–90. ISSN: 1548-7105. DOI: 10.1038/NMETH.1280.

Wang, Yanbin, Zhuhong You, Liping Li, and Zhanheng Chen (Aug. 2020). "A survey of current trends in computational predictions of protein-protein interactions." In: *Frontiers of Computer Science* 14 (4). ISSN: 20952236. DOI: 10.1007/S11704-019-8232-Z.

Waterhouse, Andrew M., James B. Procter, David M.A. Martin, Michèle Clamp, and Geoffrey J. Barton (May 2009). "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." In: *Bioinformatics* 25 (9), pp. 1189–1191. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTP033.

Wei, Dan, Qingshan Jiang, Yanjie Wei, and Shengrui Wang (July 2012). "A novel hierarchical clustering algorithm for gene sequences." In: *BMC Bioinformatics* 13 (1), p. 174. ISSN: 14712105. DOI: 10.1186/1471-2105-13-174.

Weidemüller, Paula, Maksim Kholmatov, Evangelia Petsalaki, and Judith B. Zaugg (Dec. 2021). "Transcription factors: Bridge between cell signaling and gene regulation." In: *Proteomics* 21 (23-24). ISSN: 1615-9861. DOI: 10.1002/PMIC.202000034.

Wong, Eric T.C., Victor So, Mike Guron, Erich R. Kuechler, Nawar Malhis, Jennifer M. Bui, and Jörg Gsponer (Aug. 2020). "Protein–Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations." In: *Biomolecules* 10 (8), pp. 1–19. ISSN: 2218273X. DOI: 10.3390/BIOM10081097.

Wright, Peter E. and H. Jane Dyson (Dec. 2015). "Intrinsically Disordered Proteins in Cellular Signaling and Regulation." In: *Nature reviews. Molecular cell biology* 16 (1), p. 18. ISSN: 14710080. DOI: 10.1038/NRM3920.

Xenarios, Ioannis, Danny W. Rice, Lukasz Salwinski, Marisa K. Baron, Edward M. Marcotte, and David Eisenberg (Jan. 2000). "DIP: the Database of Interacting Proteins." In: *Nucleic Acids Research* 28 (1), p. 289. ISSN: 03051048. DOI: 10.1093/NAR/28.1.289.

Xiao, Nan, Dong Sheng Cao, Min Feng Zhu, and Qing Song Xu (June 2015). "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences." In: *Bioinformatics (Oxford, England)* 31 (11), pp. 1857–1859. ISSN: 1367-4811. DOI: 10.1093/BIOINFORMATICS/BTV042.

Yamashita, Rikiya, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi (Aug. 2018). "Convolutional neural networks: an overview and application in radiology." In: *Insights into Imaging* 9 (4), pp. 611–629. ISSN: 18694101. DOI: 10.1007/S13244-018-0639-9/FIGURES/15.

Yu, Jiantao, Maozu Guo, Chris J. Needham, Yangchao Huang, Lu Cai, and David R. Westhead (Aug. 2010). "Simple sequence-based kernels do not predict protein-protein interactions." In: *Bioinformatics (Oxford, England)* 26 (20), pp. 2610–2614. ISSN: 1367-4811. DOI: 10.1093/BIOINFORMATICS/BTQ483.

Zhang, Shao Wu, Li Yang Hao, and Ting He Zhang (Feb. 2014). "Prediction of Protein–Protein Interaction with Pairwise Kernel Support Vector Machine." In: *International Journal of Molecular Sciences* 15 (2), p. 3220. ISSN: 14220067. DOI: 10.3390/IJMS15023220.

ABSTRACT

Intrinsically disordered regions (IDRs) in proteins have been linked to many crucial functions, including mediating protein-protein interactions (PPIs), despite lacking a single invariant three-dimensional structure. This growing recognition has led to an increased demand for computational studies that focus on the amino acid sequences corresponding to proteins to identify crucial sequence characteristics in IDRs and their connections to diverse cellular functions. In the first part of this thesis, we have put forward two statistical methods to identify sequence features responsible for IDR functions. We introduce a statistical approach for quantifying the periodicity of aromatic residues in the human proteome by modeling their occurrence using a Poisson process. Next, we introduce another statistical analysis of IDR sequences to identify co-occurring amino acid groups in transcription factors (TFs) that co-bind to enhancer elements. In the second part of the thesis, our focus shifts to predicting PPIs using only protein sequences. In this thesis, we present a novel method to address PPI prediction challenge using IDR sequences. We encountered challenges while developing a PPI prediction model because our task essentially involves making predictions based on pairs of input data. In this regard, we present two distinct machine learning algorithms to address two different types of PPI prediction problems, namely, asymmetric and symmetric problems. For the asymmetric problem, where one of the proteins has already been included in the classifier, we develop a method to predict disordered protein partners of the known proteins in our dataset. On the other hand, for the symmetric problem, we implement another approach to predict entirely novel PPIs. Furthermore, we explore whether IDR amino acid sequences outperform other sequence components, including entire sequences and non-IDR regions, in predicting PPIs. Our findings led us to the conclusion that disordered regions are particularly valuable in predicting interactions between intrinsically disordered proteins. In summary, this thesis provides insights into dealing with paired nature datasets when developing machine learning models for PPI prediction and demonstrates how statistical approaches can be used to investigate IDR sequences for feature identification and predict PPIs based on IDR sequences.

ZUSAMMENFASSUNG

Intrinsically disordered regions (IDRs) in Proteinen wurden mit vielen wichtigen Funktionen assoziiert, obwohl ihnen eine einzelne unveränderliche 3-dimensionale Struktur fehlt, unter anderem die Vermittlung von Protein-Protein-Interaktionen (PPIs). Die wachsende Erkenntnis über die Bedeutung von IDRs hat zu einer erhöhten Nachfrage nach computergestützten Studien geführt, die sich auf die Aminosäuresequenzen von Proteinen konzentrieren, um entscheidende Sequenzmerkmale in IDRs und ihre Verbindungen zu verschiedenen zellulären Funktionen zu identifizieren. Im ersten Teil dieser Arbeit stellen wir zwei statistische Methoden zur Identifikation von Sequenzmerkmalen vor, die für IDR-Funktionen verantwortlich sind. Wir präsentieren einen statistischen Ansatz zur Quantifizierung der Periodizität aromatischer Rückstände im menschlichen Proteom durch Modellierung ihres Auftretens anhand eines Poisson-Prozesses. Außerdem führen wir eine weitere statistische Analyse von IDR-Sequenzen ein, um gemeinsam auftretende Aminosäuregruppen in Transkriptionsfaktoren (TFs) zu entdecken, die zusammen an Enhancer-Elemente binden. Im zweiten Teil der Arbeit liegt unser Fokus auf der Vorhersage von PPIs nur aus Proteinsequenzen. Hier präsentieren wir eine neue Methode, um die Herausforderung der PPI-Vorhersage unter Verwendung von IDR-Sequenzen anzugehen. Wir stießen bei der Entwicklung eines PPI-Vorhersagemodells auf Herausforderungen, da unsere Aufgabe im Prinzip darin besteht, Vorhersagen auf der Grundlage von Paaren von Eingabedaten zu treffen. In diesem Zusammenhang stellen wir zwei unterschiedliche Algorithmen für maschinelles Lernen vor, um zwei PPI-Vorhersageproblemen zu lösen, nämlich asymmetrische und symmetrische Probleme. Für das asymmetrische Problem, bei dem eines der Proteine bereits im Klassifizierer enthalten ist, entwickeln wir eine Methode zur Vorhersage ungeordneter Proteinpartner bekannter Proteine in unserem Datenset. Für das symmetrische Problem implementieren wir hingegen einen anderen Ansatz, um völlig neue PPIs vorherzusagen. Zudem prüfen wir, ob IDR-Aminosäuresequenzen andere Sequenzkomponenten, einschließlich ganzer Sequenzen und Nicht-IDR-Regionen, in der PPI-Vorhersage übertreffen. Unsere Ergebnisse führen zu der Schlussfolgerung, dass ungeordnete Regionen besonders wertvoll für die Vorhersage von Interaktionen zwischen intrinsisch ungeordneten Proteinen sind. Zusammenfassend liefert diese Arbeit Erkenntnisse über den Umgang mit gepaarten Datensätzen bei der Entwicklung von maschinellen Lernmodellen für die PPI-Vorhersage. Wir zeigen, wie statistische Ansätze verwendet werden können, um IDR-Sequenzen für die Merkmalsidentifizierung zu untersuchen und PPIs basierend auf IDR-Sequenzen vorherzusagen.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Name: Kibar
Vorname: Gözde

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Berlin, 2024

Gözde Kibar