EVALUATING EFFECTS OF CELLULAR ENVIRONMENT ON PROTEIN FOLDING

ALIBEK KRUGLIKOV

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the degree
Doctorate in Philosophy in Biology

Department of Biology
Faculty of Science
University of Ottawa

Thèse soumise à l'Université d'Ottawa
envers la réalization partielle des éxigeances du
Doctorat en Philosophie en Biologie

Département de Biologie
Faculté des Sciences
Université d'Ottawa

Abstract

Protein structures are fundamental to their functions and interactions, requiring extensive investigation in structural proteomics. The cellular environment, including factors such as temperature, pH, and salinity, has been shown to have a profound impact on protein folding. Despite the significance of these factors, they are frequently overlooked by protein structure models and studies. This Thesis is a collection of publications that offer insight into the complex interactions between cellular environment and protein folding dynamics.

The first publication focuses on differences in protein folding between mesophilic and thermophilic bacteria. We uncover distinctions in secondary structure when using mesophiles as expression systems for thermophiles. Understanding these variations may aid in bacterial physiology comprehension and refinement of protein structure prediction models to better account for environmental influences.

The second publication explores the relationship between amino acid sequence similarities and secondary structure variations in mammalian ACE2 proteins. Given the critical role of ACE2 in mediating coronavirus cellular entry, understanding the structural determinants governing ACE2 binding interactions is crucial. Our findings offer insights into the molecular mechanisms underlying virus-host interactions, providing additional context for therapeutic and vaccine development efforts.

The third publication investigates intrinsically disordered protein (IDP) abundance across bacteria with varying optimal growth temperatures (OGTs). IDPs play versatile roles in cellular processes, with their abundance linked to environmental adaptation. Comparing IDP abundance in mesophilic and thermophilic bacteria sheds light on potential functional implications in diverse environmental contexts. This analysis enhances our understanding of protein dynamics in response to environmental cues and provides insights into bacterial adaptation and evolution.

Through these investigations, we contribute to bridging knowledge gaps regarding the influence of cellular environment on protein folding dynamics. By unraveling the complex interplay between environmental factors and protein structure, we pave the way for more accurate predictions and manipulations of protein function.

Résumé

Les structures protéiques sont essentielles à leurs fonctions et interactions, nécessitant des études approfondies en protéomique structurale. L'environnement cellulaire, comme la température, le pH et la salinité, impacte fortement le repliement des protéines. Malgré leur importance, ces facteurs sont souvent négligés par les modèles de structure protéique. Cette thèse rassemble des publications explorant les interactions complexes entre l'environnement cellulaire et la dynamique de repliement des protéines.

La première publication examine les différences de repliement entre bactéries mésophiles et thermophiles. Nous découvrons des distinctions de structure secondaire lors de l'expression de thermophiles dans des mésophiles. Comprendre ces variations peut aider à mieux appréhender la physiologie bactérienne et affiner les modèles de prédiction structurale.

La deuxième étudie la relation entre similitudes de séquence d'acides aminés et variations structurelles de l'ACE2 chez les mammifères. Étant donné le rôle clé de l'ACE2 dans l'entrée des coronavirus, comprendre les déterminants structurels régissant sa liaison est crucial. Nos résultats éclairent les mécanismes moléculaires des interactions virus-hôte, guidant le développement thérapeutique et vaccinal.

La troisième publication analyse l'abondance des protéines intrinsèquement désordonnées (IDP) chez les bactéries à différentes températures optimales de croissance. Comparer l'abondance des IDP chez les mésophiles et thermophiles révèle leurs implications fonctionnelles dans divers environnements, éclairant l'adaptation et l'évolution bactériennes.

Ces recherches comblent les lacunes sur l'influence environnementale sur le repliement protéique. En élucidant ces interactions complexes, nous permettons de meilleures prédictions et manipulations des fonctions protéiques.

Acknowledgements

I extend my sincere gratitude to my supervisor, Dr. Xuhua Xia, for his unwavering support and invaluable guidance throughout my academic journey. His constructive feedback has been instrumental in shaping me into a better scientist, and I am truly thankful for his mentorship. Moreover, I am grateful for his flexibility in allowing me to explore my research interests and for his constant encouragement during challenging times. Dr. Xia's welcoming attitude made my transition to Canada smooth, and I am sincerely thankful for the opportunity to pursue my PhD under his mentorship at the Xia lab.

I am deeply grateful to the numerous professors who have influenced my academic journey. Their support has played a crucial role in my pursuit of a PhD and in the preparation of this thesis. Special thanks are extended to my thesis advisory committee members, Dr. Stéphane Aris-Brosou, Dr. James Cheetham, and Dr. Marcel Turcotte. Their consistent feedback and encouragement have been invaluable, and I am thankful for the opportunity they provided me to transition from the Masters to the PhD program. Their guidance has significantly enhanced both the quality of my work and the methodologies employed in my research endeavors.

I thank all past and present members of the Xia lab, including Parisa Aris, Mahbubeh Askari Rad, Heba Farookhi, Bosen Jia, Mohan Rakesh, Jordan Silke, and especially Yulong Wei. Their inquisitive minds and thoughtful inquiries have significantly contributed to the advancement of my projects. Their assistance and innovative ideas have been invaluable throughout my journey. As I express my appreciation, I also extend my best wishes to them for success in their future endeavors, whether they continue in scientific research or pursue other paths. May they find fulfillment and accomplishment in all their undertakings, and may their contributions continue to make a positive impact, shaping the world in meaningful ways.

I am profoundly grateful to my family for their love, sacrifices, and steadfast support throughout my academic journey. My parents' dedication to ensuring my education has been unwavering, and I am deeply thankful for their sacrifices and encouragement. My brother's constant support and assistance with my research have been invaluable, and I appreciate his belief in my abilities. Additionally, my wife's remarkable patience and constant support, including her willingness to accompany me to a foreign country, have been instrumental in my pursuit of success in Canada.

Their love and encouragement have been my guiding light, motivating me to persevere through challenges and strive for excellence. I am forever grateful for their boundless love, and I look forward to sharing many more milestones together in the future.

Finally, I extend my gratitude to the developers at OpenAI for creating and releasing the remarkable language model utilized in the refinement of the initial text. While my proficiency in English may not be impeccable, the exceptional capabilities of the language model, alongside some strategic prompt crafting on my part, have undoubtedly enhanced the readability of this thesis compared to its previous state. As a human being, I am regrettably unable to boast the linguistic prowess of a sophisticated AI language model, but I am grateful for the assistance it provides in improving communication.

Contents

List of Figures

List of Tables

Abbreviations

AA	Amino Acid
ACE2	Angiotensin-Converting Enzyme 2
ASA	Accessible Surface Area
BLAST	Basic Local Alignment Search Tool
BLASTp	Basic Local Alignment Search Tool for proteins
CASP	Critical Assessment of Structure Prediction
CB513	Test Set of 513 proteins for protein structure prediction
CD-HIT	Cluster Database at High Identity with Tolerance
COVID-19	Coronavirus Disease 2019
DAMBE	Data Analysis in Molecular Biology and Evolution
D_{KL}	Kullback–Leibler Divergence
DNA	Deoxyribonucleic Acid
FCR	Fraction of Charged Residues
FOD	Fraction of Disorder (as measured by number of IDP-coding genes)
H7N9	Avian Influenza A
HA	hemagglutinin
hACE2	Human Angiotensin-Converting Enzyme 2
IDP	Intrinsically Disordered Protein
IDR	Intrinsic Disorder Region
JSD	Jensen-Shannon Divergence
mRNA	Messenger Ribonucleic Acid

MSA	Multiple Sequence Alignment
NA	neuraminidase
NCPR	Net Charge Per Residue
nsp5	Non-structural protein 5
OGT	Optimal Growth Temperature
OLS	Ordinary Least Squares
PDB	Protein Data Bank
PDBe	Protein Data Bank Europe
PSSP	Protein Secondary Structure Prediction
PTMs	Post-Translational Modifications
Q3	PSSP accuracy metric using 3 types of SS
Q8	PSSP accuracy metric using 8 types of SS
RAPID	Rapid Accurate Protein Intrinsic Disorder prediction
RBD	Receptor Binding Domain of Spike protein
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
S	Spike protein
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SOV	Segment OVerlap score
SS	Secondary Structure
TDP-43	TAR DNA-binding protein 43

TEMPURA	Temporal RNA-Seq Unified Reader and Annotator
tRNA	Transfer Ribonucleic Acid
TS115	Test Set of 115 proteins for protein structure prediction
TS2019	Test Set for protein structure prediction updated in 2019 (261 proteins)
UniProt	Universal Protein Resource
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
ΔG	Gibbs free energy change
T_{50}	protein melting temperature (50% are irreversibly denatured)

3-SS basic protein secondary structure classification with 3 classes:

H	helix
E	sheet
C	coil

8-SS extended protein secondary structure classification with 8 classes:

H	α-helix
E	β-sheet
C	coil
I	π-helix
G	310-helix
B	β-bridge
S	Bend
T	Turn

Amino acid notation:

A	Ala	Alanine	L	Leu	Leucine
R	Arg	Arginine	K	Lys	Lysine
N	Asn	Asparagine	M	Met	Methionine
D	Asp	Aspartic acid	F	Phe	Phenylalanine
B	Asx	Asparagine or aspartic acid	P	Pro	Proline
C	Cys	Cysteine	S	Ser	Serine
Q	Gln	Glutamine	T	Thr	Threonine
E	Glu	Glutamic acid	W	Trp	Tryptophan
Z	Glx	Glutamine or glutamic acid	Y	Tyr	Tyrosine
G	Gly	Glycine	V	Val	Valine
H	His	Histidine	X	-----	Unknown
I	Ile	Isoleucine			

Chapter 1. Introduction

1.1 General Background on Protein Structure

1.1.1 Overview of Protein Structure

Proteins are fundamental biomolecules that play essential roles in various biological processes within living organisms. These versatile molecules exhibit diverse functions, ranging from catalyzing biochemical reactions to providing structural support and facilitating cellular communication.

Protein structure is central to their functionality. The structure of a protein refers to the spatial arrangement of its constituent amino acid residues in three-dimensional space. Proteins are composed of linear chains of amino acids, each having unique chemical properties and side chains. The specific sequence of amino acids determines the overall structure and function of the protein.

Protein structure can be described as organized into hierarchical levels, where each level contributes to the overall properties of the molecule. The primary structure represents the linear sequence of amino acids in the polypeptide chain (Sanger 1952). This sequence is dictated by the genetic code and serves as the foundation for higher-order structural organization. The secondary structure refers to local folding patterns within the polypeptide chain, resulting from hydrogen bonding between amino acid residues (Pauling, Corey, and Branson 1951). Common secondary structures include alpha helices and beta sheets, which are important for stability and rigidity to the protein (Sun, Foster, and Boyington 2004). Tertiary structure encompasses the overall three-dimensional folding of the entire polypeptide chain, driven by interactions between amino acid side chains. These interactions include hydrogen bonds, disulfide bonds, hydrophobic interactions, and electrostatic attractions (Vella 1992). Tertiary structure determines the specific shape and functional properties of the protein. Finally, in proteins consisting of multiple polypeptide chains, the quaternary structure describes the arrangement of these subunits relative to each other. Quaternary structure is crucial for the assembly and function of protein complexes, such as enzymes and structural proteins (E. Wood 1996).

1.1.2 Importance of Protein Structure in Biological Processes

Understanding the intricate details of protein structure is essential in unraveling their diverse functions and mechanisms of action. By interpreting protein structure, researchers gain invaluable insights into how these molecules execute their biological functions. For instance, knowledge of a protein's three-dimensional arrangement reveals its active sites, enabling predictions of substrate binding and catalytic mechanisms (Yabukarski et al. 2020). This understanding forms the foundation for biochemical and pharmacological studies aimed at manipulating protein function for therapeutic purposes (Bancet et al. 2020; Francoeur et al. 2020; S. Luo et al. 2021).

Moreover, protein structure governs their interactions with other molecules, including other proteins, nucleic acids, and small molecules. The specific spatial arrangement of amino acid residues within a protein dictates its affinity and specificity towards interacting partners (Honig and Shapiro 2020). These molecular interactions underpin essential biological processes such as signal transduction, gene regulation, and molecular transport (Berg, Tymoczko, and Stryer 2002). Consequently, untangling protein structure not only sheds light on individual protein functions but also unveils intricate networks of molecular interactions crucial for cellular homeostasis and organismal development.

Furthermore, deviations from the native protein structure, often resulting from genetic mutations or environmental factors, can lead to structural abnormalities with profound implications for health and disease. Such alterations may disrupt protein folding or compromise protein stability, culminating in a spectrum of disorders known as proteinopathies. Examples of protein misfolding diseases include Alzheimer's disease (Rahman and Lendel 2021; Y. Chen et al. 2021), Parkinson's disease (S. Hu et al. 2021; Fanning, Selkoe, and Dettmer 2020), and cystic fibrosis (Q. Chen, Shen, and Zheng 2021; Farinha and Callebaut 2022; Lewis et al. 2005), wherein aberrant protein structures contribute to pathogenic processes underlying these conditions. Understanding the relationship between protein structure and disease pathology is instrumental in the development of targeted therapies aimed at mitigating or preventing the adverse effects of protein misfolding.

1.1.3 Protein Folding Mechanisms

Protein folding is a complex process influenced by fundamental thermodynamic principles that drive the transition from a disordered state to a stable, biologically active structure. At the core of this process lies the inherent tendency of proteins to seek the lowest free energy state, which corresponds to their native three-dimensional structure. This thermodynamic principle, known as the "thermodynamic hypothesis," was first proposed by Christian B. Anfinsen in the 1970s and has since been widely accepted as a fundamental component of protein folding (Anfinsen 1973; Kaffe-Abramovich and Unger 1998).

The folding process involves the optimization of various non-covalent interactions, including hydrogen bonding, van der Waals forces, hydrophobic interactions, and electrostatic interactions. These forces collectively contribute to the stabilization of the protein's native structure, which represents the thermodynamically favored state (J. Wang et al. 2016; Ross and Rekharsky 1996). Through a series of stochastic events, proteins explore a vast conformational space, sampling a myriad of intermediate states before settling into their energetically preferred structure. This process is often referred to as the "folding funnel," wherein the landscape of potential protein conformations narrows as folding progresses, ultimately leading to the native state (Socci, Onuchic, and Wolynes 1998).

Entropy, a measure of disorder or randomness, plays a crucial role in protein folding dynamics. As a polypeptide chain folds into its thermodynamically optimal structure, the conformational entropy decreases due to the reduction in the number of possible conformations. The free energy of unfolding (ΔG) represents the energy required to disrupt the native structure and can be used as a measure of protein stability (Teufl, Zajc, and Traxlmayr 2022; Stadler et al. 2016). Proteins with higher ΔG values are more resistant to unfolding and exhibit greater thermodynamic stability. Another way to quantify the average thermodynamic stability of a protein domain is using melting temperature (T_{50}), the temperature at which half of the protein population is in irreversibly denatured. For mesophilic proteins, which thrive in moderate temperature environments, T_{50} typically is below 50°C. In contrast, thermophilic proteins, adapted to survive in high-temperature environments, exhibit higher T_{50} values, often exceeding 80°C (Teufl, Zajc, and Traxlmayr 2022). This increased thermostability is achieved through various structural

adaptations, such as enhanced hydrophobic interactions, increased rigidity, and optimized electrostatic interactions.

In the crowded and dynamic cellular environment, protein folding is often susceptible to errors and interruptions. To mitigate these challenges, cells employ a sophisticated network of molecular chaperones, specialized proteins that assist in the correct folding of nascent polypeptides. Chaperones bind to unfolded or misfolded protein intermediates, shielding them from inappropriate interactions and facilitating their transition to the native state (Balchin, Hayer-Hartl, and Hartl 2020). By providing a supportive environment conducive to folding, chaperones enhance the efficiency and fidelity of protein folding pathways. Notably, chaperone proteins play essential roles in preventing protein aggregation, a characteristic of numerous neurodegenerative diseases, by promoting the correct folding and assembly of misfolded protein species (Sakahira et al. 2002; Schlee and Reinstein 2002; Giffard et al. 2004).

In addition to thermodynamic and chaperone-mediated mechanisms, the folding dynamics of proteins are profoundly influenced by physiochemical properties inherent to their surrounding environment. Factors such as pH, temperature, ionic strength, and the presence of cofactors or ligands exert significant effects on protein folding kinetics and stability. For instance, variations in pH can alter the protonation state of ionizable amino acid residues, thereby modulating electrostatic interactions critical for folding (Platzer, Okon, and McIntosh 2014; Konermann 2012). Similarly, changes in temperature can disrupt hydrogen bonding and hydrophobic interactions, leading to protein denaturation or aggregation (Konermann 2012; Soulages et al. 2002). Furthermore, the presence of specific ions or small molecules can serve as cofactors or stabilizers, promoting the adoption of specific protein conformations essential for biological function (Jaenicke, n.d.).

The interplay between thermodynamic principles, molecular chaperones, and physiochemical influences shapes the intricate landscape of protein folding, underscoring its essential role in cellular physiology and disease pathogenesis. Explaining the mechanisms governing protein folding not only enhances our understanding of fundamental biological processes but also holds promise for the development of novel therapeutic strategies targeting protein misfolding disorders. By unraveling the complexities of protein folding mechanisms, researchers continue to

unlock the secrets of life's molecular choreography, paving the way for transformative advances in biomedicine and beyond.

1.2 Cellular Environment and Protein Folding

1.2.1 Environmental Factors Influencing Protein Folding

Environmental factors exert profound effects on protein folding dynamics and stability, playing critical roles in shaping the functional properties of proteins within the cellular milieu. Among these factors, temperature, pH, and osmotic pressure stand out as key modulators of protein folding processes, influencing protein stability, conformational changes, and interactions.

Temperature variations represent one of the most significant environmental factors impacting protein folding. Proteins are exquisitely sensitive to changes in temperature, with alterations in thermal conditions affecting their stability and folding kinetics. At higher temperatures, proteins are more prone to denaturation (S. Bondos and Matthews 2021; Jaenicke, n.d.), wherein they lose their native structure and functional integrity due to disruption of non-covalent interactions that maintain their three-dimensional conformation. Conversely, lowering the temperature can slow down protein folding kinetics (Privalov 1990; Xiong 1997), leading to the accumulation of folding intermediates and increased susceptibility to misfolding and aggregation. Moreover, extreme temperatures can induce irreversible protein unfolding, resulting in protein aggregation and loss of biological activity (S. Bondos and Matthews 2021). The sensitivity of proteins to temperature variations underscores the importance of maintaining thermal homeostasis within cells to ensure proper protein folding and function.

In addition to temperature, pH fluctuations represent another crucial environmental factor influencing protein folding. Proteins exhibit optimal stability and activity within a specific pH range dictated by their amino acid composition and structural context. Deviations from this optimal pH range can disrupt electrostatic interactions, leading to changes in protein conformation and stability. For instance, at extremes of pH, ionizable amino acid residues may become protonated or deprotonated, altering their charge distribution and affecting intra- and intermolecular interactions (Konermann 2012; Platzer, Okon, and McIntosh 2014; M. S. Lee, Salsbury Jr., and Brooks III 2004). Consequently, pH fluctuations can perturb protein folding

pathways, leading to the formation of misfolded or aggregated protein species with compromised biological functions.

Furthermore, osmotic pressure, resulting from differences in solute concentrations across cellular membranes, represents another environmental factor that impacts protein folding and stability. Osmotic pressure influences the hydration state of proteins, affecting their solubility and conformational stability (Wennerström and Oliveberg 2022). High osmotic pressures, such as those encountered during osmotic stress or dehydration, can lead to protein denaturation and aggregation, as water molecules are depleted from the protein surface, exposing hydrophobic regions that promote protein-protein interactions. Conversely, low osmotic pressures may induce protein swelling and conformational changes, altering protein-protein interactions and cellular signaling pathways. The osmotic environment of cells plays a critical role in maintaining protein homeostasis and cellular function, as disruptions in osmotic balance can compromise protein folding, membrane integrity, and cell viability.

1.2.2 Cellular Factors Modulating Protein Folding

In addition to environment, other cellular factors influence protein folding and are essential for comprehending the nuances of cellular physiology. In the bustling and dynamic environment of the cell, numerous factors shape the folding kinetics and stability of proteins, ultimately defining their functional properties.

One significant factor of protein folding within cells is intracellular molecular crowding. The cell's interior is densely packed with macromolecules, creating limited space for molecular diffusion and interaction. This crowding imposes steric constraints on protein folding, affecting the accessibility of folding intermediates and altering folding and unfolding rates (Gomes and Faísca 2019; Kuznetsova, Turoverov, and Uversky 2014). Additionally, the stability of proteins in crowded environments can be influenced by specific interactions such as hydrophobic interactions and hydrogen bonding with crowding agents, resulting in a more complex effect on the folding of various proteins (Macdonald et al. 2016; Miklos et al. 2011).

Post-translational modifications (PTMs) also play a crucial role in regulating protein folding pathways. PTMs dynamically modify protein structure and function, influencing folding, stability, and activity. For example, phosphorylation can induce conformational changes (Liang

et al. 2006), while glycosylation affects protein solubility and recognition by chaperones (Zacchi et al. 2014; Broncel et al. 2010). The interplay between PTMs and protein folding is pivotal in cellular function.

Another critical factor affecting protein folding is the cellular redox state, which regulates disulfide bond formation. Disulfide bonds stabilize protein structures, and their formation is influenced by the balance between oxidizing and reducing agents in the cell. Oxidizing conditions promote disulfide bond formation, enhancing protein stability and correct folding, while reducing conditions facilitate disulfide bond reduction, leading to misfolding and aggregation (Eben and Imlay 2023; Rajapaksha, Pandeya, and Wei 2020; Ramírez-Palma et al. 2021). Disruption of cellular redox homeostasis is implicated in various diseases (Bhattacharyya et al. 2014; Lennicke et al. 2016), emphasizing its significance in protein folding and cellular function.

1.2.3 Protein Folding in Extremophile Species

As discussed previously, cellular environment exerts significant influence on protein folding processes. Factors such as temperature, pH, and salinity can alter the stability and conformation of proteins. Extremophiles, including thermophiles thriving in high temperatures, psychrophiles inhabiting cold environments, acidophiles surviving in low pH, as well as other groups of species, have evolved distinct protein folding patterns to cope with their extreme habitats. Understanding these adaptations is essential for unraveling the mechanisms underlying protein stability and function.

Extremophiles exhibit specific folding patterns tailored to their harsh environments. For instance, thermophilic proteins often possess increased thermostability through enhanced hydrophobic interactions and rigid structures (Ahmed et al. 2022; Gromiha et al. 2013; Rathi, Höffken, and Gohlke 2014). In contrast, psychrophilic proteins adapt to cold temperatures by employing different strategies such as enhanced glycosylation (L. Li et al. 2020), upregulation and isoform exchange of cold-shock proteins (Koh et al. 2017). These unique folding patterns not only enable extremophiles to thrive in extreme conditions but also offer valuable insights into the fundamental principles of protein folding.

Evolutionary adaptations of extremophile proteins have a great effect on their stability and denaturation behavior in non-native environments. Proteins from psychrophilic species, optimized for cold temperatures, are likely to experience denaturation not only at extremely high temperatures but also under mesophilic conditions (20-40°C). Conversely, thermophilic proteins, adapted to withstand high temperatures, may undergo cold denaturation when exposed to normal or lower temperatures. These temperature-dependent denaturation trends highlight the importance of considering the native cellular environment when studying protein structures from extremophiles. Expressing extremophile proteins in mesophilic model organisms or utilizing them in standard laboratory conditions could lead to misfolding or destabilization, potentially compromising structural and functional analyses. This logic extends to species adapted to other environmental factors such as pH, salinity, and pressure, as their adaptations can also influence protein folding and stability.

Extremophiles have garnered significant research interest due to their potential applications in various fields. Their specialized folding patterns hold promise for industrial biocatalysis, bioremediation, and pharmaceutical development. Moreover, studying extremophile proteins provides a window into evolutionary processes and the adaptive strategies employed by organisms facing extreme environmental challenges. By elucidating the specificities of protein folding in extremophiles, researchers can uncover novel mechanisms for protein stabilization and design.

1.3 Intrinsically Disordered Proteins

1.3.1 Structural and functional implications of intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) represent a fascinating class of proteins that lack well-defined tertiary structures under physiological conditions. Despite their lack of a stable structure, IDPs play pivotal roles in various cellular processes, showcasing unique structural and functional properties that distinguish them from their ordered counterparts.

One of the hallmark features of IDPs is their ability to adopt multiple conformations. Unlike globular proteins, which fold into well-defined three-dimensional structures, IDPs exist as dynamic ensembles of interconverting conformations (V. N. Uversky, Gillespie, and Fink 2000).

This conformational flexibility allows IDPs to interact with a diverse array of binding partners (Ferreon et al. 2009; Santofimia-Castaño et al. 2017) and adapt their structures to fulfill specific functional roles within the cell. By adopting different conformations, IDPs can exhibit versatile binding interfaces, enabling interactions with multiple interaction partners and facilitating the formation of transient protein complexes.

IDPs play crucial roles in cellular signaling and regulation pathways, where their structural plasticity allows for dynamic regulation of cellular processes. Many IDPs function as molecular switches, transitioning between different conformations in response to environmental cues or post-translational modifications (Berlow, Dyson, and Wright 2022). Through their interactions with signaling molecules, IDPs can modulate signal transduction cascades, regulate gene expression, and orchestrate catalysis of enzyme activity (DeForte and Uversky 2017; Bemporad et al. 2008). Moreover, IDPs often act as scaffolds or adapters, facilitating the assembly of multiprotein complexes involved in various cellular processes, including cell cycle regulation, DNA repair, and apoptosis (Bernardini et al. 2023; Yoon et al. 2012; Kai 2016; Xuejun Jiang et al. 2003).

IDPs serve as hubs in protein-protein interaction networks, where they act as central nodes connecting multiple signaling pathways and regulatory circuits. Due to their promiscuous binding properties, IDPs can interact with numerous protein partners simultaneously, forming dynamic interaction networks that integrate diverse cellular signals and coordinate complex cellular behaviors (Kurzbach et al. 2014; Cho et al. 2021). By serving as hubs in protein interaction networks, IDPs contribute to the robustness and adaptability of cellular signaling networks, allowing cells to respond dynamically to changing environmental conditions and physiological demands (Cho et al. 2021).

IDPs exhibit remarkable structural and functional properties that underlie their diverse roles in cellular physiology. Their conformational flexibility, involvement in signaling and regulation pathways, and centrality in protein-protein interaction networks highlight the importance of IDPs in orchestrating cellular processes and maintaining cellular homeostasis. Understanding the structural and functional implications of IDPs provides valuable insights into their roles in biological processes, as well as into their evolution.

1.3.2 Molecular mechanisms underlying the folding and regulation of intrinsically disordered proteins

Post-translational modifications play a crucial role in modulating the conformational dynamics of IDPs. PTMs such as phosphorylation, acetylation, methylation, and ubiquitination introduce chemical modifications to specific residues within IDPs, altering their structural and functional properties. For example, phosphorylation of serine, threonine, or tyrosine residues within IDPs can induce conformational changes by introducing electrostatic repulsion or promoting interactions with other proteins or ligands (Y. Shang et al. 2021; Pandey et al. 2023). Similarly, acetylation and methylation can affect the hydrophobicity or charge distribution of IDPs (Y. Luo et al. 2014; Migliori et al. 2010), influencing their folding propensity and interaction with binding partners. By regulating the conformational dynamics of IDPs, PTMs play a pivotal role in cellular signaling, transcriptional regulation, and protein-protein interactions.

The binding of intrinsically disordered proteins to partner molecules induces structural transitions that are essential for their functional regulation. IDPs often undergo disorder-to-order transitions upon binding to specific ligands, proteins, nucleic acids, or membranes, adopting a well-defined structure that is tailored for interaction with their binding partners. This induced folding process involves the formation of transient secondary or tertiary structures, such as α-helices, β-sheets, or coiled-coil motifs, which enable IDPs to recognize and bind to their targets with high specificity and affinity. Additionally, the binding-induced structural transitions in IDPs can modulate their enzymatic activity, subcellular localization, or protein-protein interaction networks, thereby regulating diverse cellular processes (Wetzler et al. 2018; Rangarajan, Kulkarni, and Hannenhalli 2015).

One way to assess protein folding dynamics is using free energy landscape slopes. As a protein transitions from an unfolded to a folded state, its stability increases and the conformational variability decreases. This change of folding potential can be quantified through free energy of the protein. As protein structure reaches its optimal state, free energy decreases and the protein reaches its lowest free energy state. In structured proteins this energy landscape is funnel-shaped, however, the steepness of energy change is lower in IDPs than in structured proteins (Chong and Ham 2019). This difference reflects the flexible and structural dynamicity of IDPs, indicating a larger number of accessible conformations. At the same time, specific interactions between

secondary structures of IDPs and binding partners can make the energy landscape much steeper, making IDPs folding dynamics more similar to those of structured proteins upon binding.

1.3.3 Role of intrinsically disordered proteins in health and disease

IDPs have emerged as key players in the pathogenesis of various diseases, with their dysregulation implicated in the development and progression of neurodegenerative disorders, cancer, and other debilitating conditions (Kai 2016; Kelaini et al. 2021; Cario et al. 2022).

Neurodegenerative disorders, such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis (Grossman 2019), are characterized by the progressive degeneration of neurons and the accumulation of misfolded proteins within the brain. Intriguingly, many of the proteins associated with neurodegenerative disorders, including amyloid-beta, tau, alpha-synuclein, and TDP-43, exhibit intrinsically disordered regions that play critical roles in their aggregation and toxicity (Ferreon et al. 2009; Gu et al. 2021; Cario et al. 2022). Dysregulation of IDPs in these disorders can lead to the formation of toxic protein aggregates and disrupt cellular homeostasis, ultimately contributing to neuronal dysfunction and cell death.

In cancer, intrinsically disordered proteins contribute to the pathogenesis of the disease through aberrant signaling pathways and dysregulated protein-protein interactions. IDPs such as p53, c-Myc, and HIF-1α play pivotal roles in regulating cell proliferation, apoptosis, and tumor angiogenesis, with their dysregulation implicated in oncogenesis and tumor progression (Darling and Uversky 2018; Russo et al. 2016; Santofimia-Castaño et al. 2020). Moreover, IDPs are frequently involved in protein-protein interactions with key signaling molecules and transcription factors, influencing the activation of oncogenic pathways and the maintenance of cancer cell phenotypes.

Targeting the conformational dynamics and interactions of IDPs represents a promising approach for developing therapeutics aimed at mitigating protein aggregation and ameliorating neurodegenerative pathology and cancers (Randolph, Parra, and Libich 2021). Understanding the intricate processes of IDP folding, assessing their abundance in biological systems, and exploring their evolutionary aspects could unlock valuable insights for advancing treatments in the future. Delving deeper into the study of IDP folding, abundance, and evolution may pave the way for innovative discoveries and therapeutic interventions in various disease contexts.

1.4 Objectives and Significance of the Study

The objectives of this thesis are outlined across three chapters, each addressing critical aspects of protein structure and cellular environment in diverse biological contexts. Through these chapters, we aim to shed light on the intricate relationships between protein folding, environmental adaptations, and species-specific variations in protein structure. Additionally, we seek to uncover the functional significance of intrinsically disordered proteins in bacterial physiology, offering insights into the intricate relationship between IDP abundance and bacterial optimal growth temperature.

In Chapter 2, we delve into investigating how the cellular environment of the thermophilic bacterium *Thermus thermophilus* influences protein folding dynamics, particularly focusing on secondary structure. Traditional protein structure studies predominantly utilize *Escherichia coli* and other model organisms for gene expression. However, protein folding is intricately influenced by factors within the cellular environment, including chaperone proteins, pH levels, temperature, and ionic concentrations. Given the differences in these factors, particularly temperature and chaperone activity, native proteins from extremophiles may encounter challenges in folding properly when expressed in mesophilic model organisms like *E. coli*. Through computational analyses, we aim to deepen our understanding of the dependency of protein folding on expression systems, providing valuable insights into protein folding dynamics in extreme environments.

Chapter 3 describes a comparative analysis of mammalian ACE2 proteins, employing protein secondary structure analysis to identify regions of high and low variability. ACE2 serves as the receptor for the SARS-CoV-2 virus, making it a focal point for understanding host-virus interactions and species-specific differences in susceptibility to viral infections. By examining variations in protein secondary structure across different mammalian species, we aim to identify potential species of interest for epidemiologists, as well as provide background for interpretation of structural determinants contributing to species-specific differences. This chapter will provide valuable insights into host-virus interactions and inform the development of therapeutic strategies targeting ACE2-mediated pathways.

In Chapter 4, our objective is to explore differences in the abundance of IDP-coding genes between thermophilic and mesophilic bacteria. IDPs lack stable tertiary structures under

physiological conditions and play diverse roles in cellular processes, including signaling and regulation. Through comparative proteomic analyses, we aim to assess the prevalence and functional implications of IDPs in bacterial proteomes with varying optimal growth temperatures. This chapter will contribute to our understanding of the intricate relationship between IDP abundance and OGT and provide information for potential IDP evolution studies, as well as for engineering proteins with tailored functional properties for biotechnological applications.

The chapters of this thesis address specific research objectives, collectively contributing to our understanding of protein structure and function in diverse biological contexts. By employing a combination of computational and bioinformatics approaches, we aim to uncover novel insights into protein folding dynamics, species-specific variations in protein structure, and the significance of intrinsically disordered protein abundance in different bacteria. Ultimately, the findings from this study have the potential to inform biotechnological innovations, therapeutic interventions, and our broader understanding of protein structure-function relationships in biology.

1.5 Bioinformatics Methods and Data Sources

Throughout this thesis, several bioinformatics tools and statistical methods have been employed to analyze protein folding dynamics, species-specific variations, and the abundance of intrinsically disordered proteins. This section provides a concise overview of these methods, while detailed information on their applications and nuances is presented within the respective chapters where they are utilized. Additionally, our criteria for data selection would be discussed here.

DAMBE (Xia 2018), a comprehensive bioinformatics program, played a pivotal role in our research. Thanks to its highly diverse range of applications, it was useful throughout all experiments described in the following chapters. Specifically, it played a role in our research by facilitating the processing of amino acid and secondary structure sequences, as well as computing and handling phylogenetic data. For example, we used DAMBE's implementation of UPGMA to construct phylogenetic trees used for quasi-contrast calculations described in Chapter 4. Furthemore, local BLAST (Johnson et al. 2008) databases have been created and BLAST algorithm was run across them through DAMBE, as described in detail in Chapter 2.

Python served as the primary programming language for our data analysis and data handling tasks, leveraging both bioinformatics packages like Biopython (Cock et al. 2009) and statistical libraries such as Scikit-learn and Scipy. Biopython has been especially important for Chapter 4 of this Thesis, as we used the package to handle and collect sequence data, produce alignments and compute distances, among other applications. Scikit-learn and Scipy were used to produce models, such as Ordinary Least Squares (OLS) and run statistical tests. Additional calculations, such as Jensen-Shannon divergence, have been calculated using custom functions to allow compatibility with our data structure.

Our research leveraged the power of Google Colab notebooks, cloud-based Jupyter notebooks, to maintain an interactive coding environment. Certain programs, such as the standalone CD-HIT (Fu et al. 2012) described in Chapter 4, were executed through Jupyter sessions. Python was utilized to generate inputs for CD-HIT and to continue the analysis using the program's outputs. The complete code is available in repositories at https://github.com/alibekk93/project-protein_folding_distances and https://github.com/alibekk93/IDP_analysis/tree/RAPID.

For intrinsically disordered protein (IDP) calculations, we employed two methods: RAPID (Yan et al. 2013) and fIDPnn (G. Hu et al. 2021). The indeed very fast RAPID algorithm was run through a web server at http://biomine.cs.vcu.edu/servers/RAPID, while the more computationally intensive fIDPnn was executed locally using a Docker container obtained from https://gitlab.com/sina.ghadermarzi/fldpnn_docker. This combination of tools and platforms facilitated efficient and scalable analysis of protein folding dynamics and intrinsically disordered regions across our datasets and this process is further described in Chapter 4.

Google Colab notebooks, the cloud-based Jupyter notebooks interpretation, have been utilised in order to maintain interactive use of the code and some programs have been ran through Jypyter sessions. For example, we used a standalone CD-HIT (Fu et al. 2012) program, as described in Chapter 4, in a Jupyter session, using python both to generate CD-HIT inputs, as well as to continue analysis using the outputs of the program. Our notebooks, containing all the code, are available on repositories at https://github.com/alibekk93/project-protein_folding_distances and https://github.com/alibekk93/IDP_analysis/tree/RAPID. Additionally, IDP calculations have been performed using RAPID (Yan et al. 2013) and fIDPnn (G. Hu et al. 2021). RAPID, indeed a very fast algorithm, has been run through a web server at

http://biomine.cs.vcu.edu/servers/RAPID. As for fIDPnn, it is a much slower model, therefore we used a Docker container, available at https://gitlab.com/sina.ghadermarzi/fldpnn_docker, to run the predictions locally.

This research drew upon multiple authoritative data sources to obtain comprehensive information on protein structures, sequences, and optimal growth temperatures. The Protein Data Bank (PDB) (Burley et al. 2017) served as the primary resource for acquiring structural data, including secondary and three-dimensional protein structures. UniProt (The UniProt Consortium 2021), a renowned protein sequence database, provided access to entire bacterial proteomes in FASTA format, containing amino acid sequences. Additionally, TEMPURA (Sato et al. 2020), a specialized database, furnished valuable data on the optimal growth temperatures for various bacterial species. These sources were selected for their reliability, ease of use, and the abundance of relevant information they offered, enabling a thorough investigation of the research objectives.

Chapter 2. Proteins from Thermophilic *Thermus thermophilus* Often Do Not Fold Correctly in a Mesophilic Expression System Such as *Escherichia coli*

Alibek Kruglikov[1], Yulong Wei[1], Xuhua Xia[1,2]

1. Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa,

Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486.

2. Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada K1H 8M5.

This chapter was originally published as: Kruglikov, A., Wei, Y. & Xia, X. (2022). Proteins from Thermophilic Thermus thermophilus Often Do Not Fold Correctly in a Mesophilic Expression System Such as Escherichia coli. ACS Omega, 7(42), 37797–37806. doi: 10.1021/acsomega.2c04786.

Author contributions: A.K. carried out the experiment and wrote the manuscript with support from Y.W. and X.X.; X.X. supervised the project. All authors reviewed the manuscript.

2.1 Abstract

Majority of protein structure studies use *E. coli* and other model organisms as expression systems for other species' genes. However, protein folding depends on cellular environment factors, such as chaperone proteins, cytoplasmic pH, temperature, and ionic concentrations. Because of differences in these factors, especially temperature and chaperones, native proteins in organisms such as extremophiles may fold improperly when they are expressed in mesophilic model organisms. Here we present a methodology of assessing the effects of using *E. coli* as the expression system on protein structures. We compare these effects between eight mesophilic bacteria and *T. thermophilus*, a thermophile, and found that differences are significantly larger for *T. thermophilus*. More specifically, helical secondary structures in *T. thermophilus* proteins are often replaced by coil structures in *E. coli*. Our results show unique directionality in misfolding when proteins in thermophiles are expressed in mesophiles. This indicates that

extremophiles, such as thermophiles, require unique protein expression systems in protein folding studies.

2.2 Introduction

Identification of protein structure is a major requirement in studies on protein functions (J. Yang et al. 2015; Whisstock and Lesk 2003) and interactions (Brady and Sharp 1997; Gohlke, Hendlich, and Klebe 2000; Q. C. Zhang et al. 2012) and in the design of novel enzymes (J. M. Wood et al. 2003; Koga et al. 2012; Kiss et al. 2013). As of July 2022, there were more than 190,000 records in Protein Data Bank (PDB) (Berman et al. 2000), the largest database of protein structures, (Burley et al. 2017) of which almost 180,000 were protein structure entries. The deposited structures can be used in various fields of research. For example, PDB data had been recently used in research related to COVID-19 (Wibmer et al. 2021; Khan et al. 2021; Coutard et al. 2020), protein evolution (Schüler and Bornberg-Bauer 2016; Konaté et al. 2019; Sharir-Ivry and Xia 2017), computational enzyme (Harrington et al. 2017), and drug (Durairaj and Shanmughavel 2019) design, as well as for training computational protein structure prediction algorithms (Jumper et al. 2021; Senior et al. 2020; Waterhouse et al. 2018; Krivov, Shapovalov, and Dunbrack 2009; J. Yang et al. 2020).

While PDB holds a relatively large variety of protein types and source species, diversity is much lower for expression systems used in structure determination experiments. Protein source species are the species that the protein-coding sequences were taken from, and protein expression systems are the species that these proteins were grown in (Figure 2.1). A majority of PDB experiments use *Escherichia coli* (*E. coli*) as the expression system. For example, out of over 1,800 PDB entries with *Bacillus subtilis* (*B. subtilis*) as the source organism, more than 1,700 were grown in *E. coli*. For most other species, the proportion of proteins that were grown in *E. coli* is even higher.

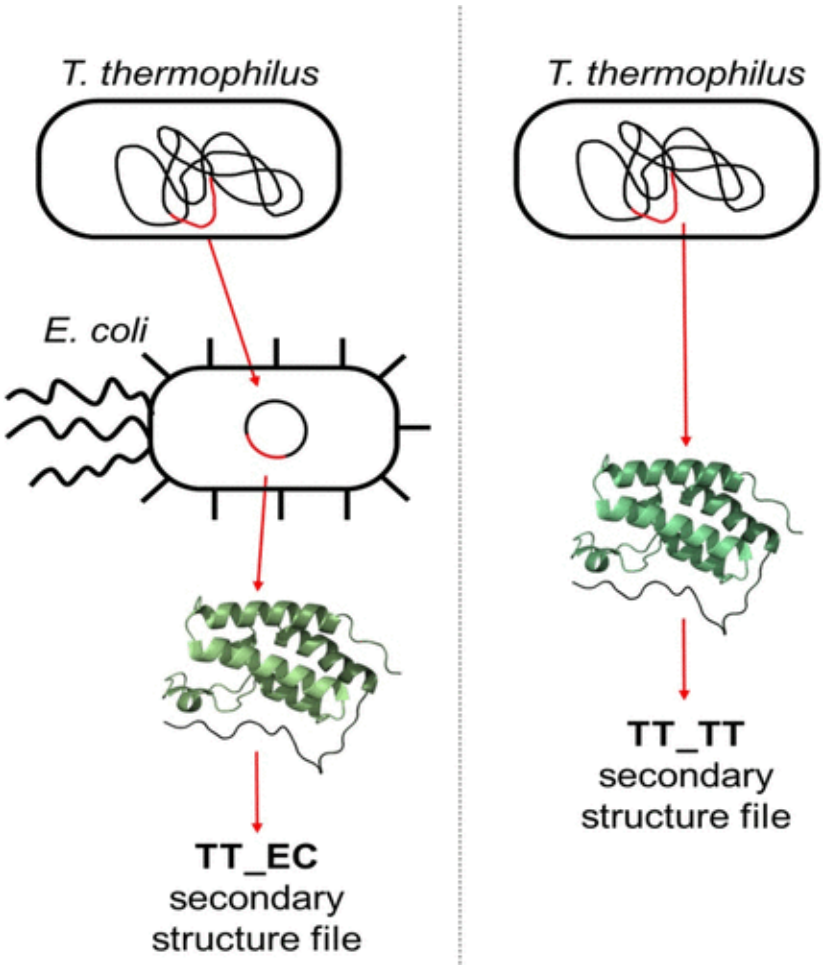

Figure 2.1. Example of protein structures in different expression systems. TT_EC file is formed using proteins that have *T. thermophilus* as source organism and *E. coli* as expression system. TT_TT is formed using proteins that have *T. thermophilus* as both source and expression system (sometimes referred to as "native" expression system in this research). Figure adapted from (Kruglikov, Wei, and Xia 2022).

While using recombinant model organisms ─ those with genetically recombined genes ─ for protein production is generally effective, lower protein activity and solubility can be observed in many cases, including the formation of inclusion bodies (Rosano and Ceccarelli 2014; Sørensen and Mortensen 2005; Kaur, Kumar, and Kaur 2018). Various protocols and methodologies have been developed in an effort to improve recombinant protein production quality; however, their effectiveness is not uniform across different protein source species. For many thermophilic species, production of recombinant protein in *E. coli* in active form is still a major challenge. For

example, multiple studies show that a large fraction of *Thermus thermophilus* (*T. thermophilus*) proteins are formed in insoluble and/or in inactive form when grown in recombinant *E. coli* (Hidalgo et al. 2004; López-López, Cerdán, and González-Siso 2015; Niehaus et al. 1999; Krefft et al. 2017), potentially because *T. thermophilus* has optimal growth temperatures around 70–80 °C, which are much higher than that of *E. coli* (37 °C). It has been previously reported that soybean Late Embryogenesis Abundant proteins became more hydrated upon heating (Soulages et al. 2002), suggesting that protein solubility is influenced by the cellular environment; therefore what is not soluble in mesophilic *E. coli* may well be soluble in *T. thermophilus*. In fact, induction temperature is one of the most critical growing conditions to produce soluble protein (Y. Kim et al. 2008).

These findings suggest that proteins may misfold in recombinant expression systems having dissimilar cellular environments. Nevertheless, expression system is not always considered important or even reported in structural studies. Moreover, protein structure prediction models, including the most advanced ones, such as AlphaFold2 (Jumper et al. 2021), do not use expression system and cellular environment as factors in training and validation. This disparity highlights the need for a methodology to assess the effect of different expression systems on protein structure determination.

Here we present a metric to evaluate protein secondary structure (SS) differences and analyze how changing the expression system from "native" to *E. coli* affects protein SS. We used PDB data to create AA/SS data sets for *T. thermophilus* (TT) and seven nonthermophilic bacteria species (where AA means amino acid and SS means secondary structure). For each species (say XX), there are two sets of protein structure data, one obtained with XX as both the protein source and the expression system and the other with XX as the source species but E. coli as the expression system. These two sets of protein structures are represented as XX_XX and XX_EC (Figure 2.1, where XX is TT). We then processed the data into probability matrices and calculated Jensen–Shannon divergence (JSD) between the XX_XX matrix and the XX_EC matrix. This JSD measured the difference in protein structure between XX_XX and XX_EC. We found that JSD was significantly higher between TT_TT and TT_EC than between XX_XX and XX_EC (where XX is a mesophilic bacterial species). This implies that *T. thermophilus* proteins in nonthermophilic species do not fold in the same way as in their "native" thermophilic

expression system, while for the other species the expression system did not affect protein folding.

2.3 Materials

For a fair comparison, we need the same protein from a source species but expressed in different expression systems, one being the source system (native) and the other being *E. coli*. An overview of our methodology generating such data is summarized in Figures 2.2 and 2.3.

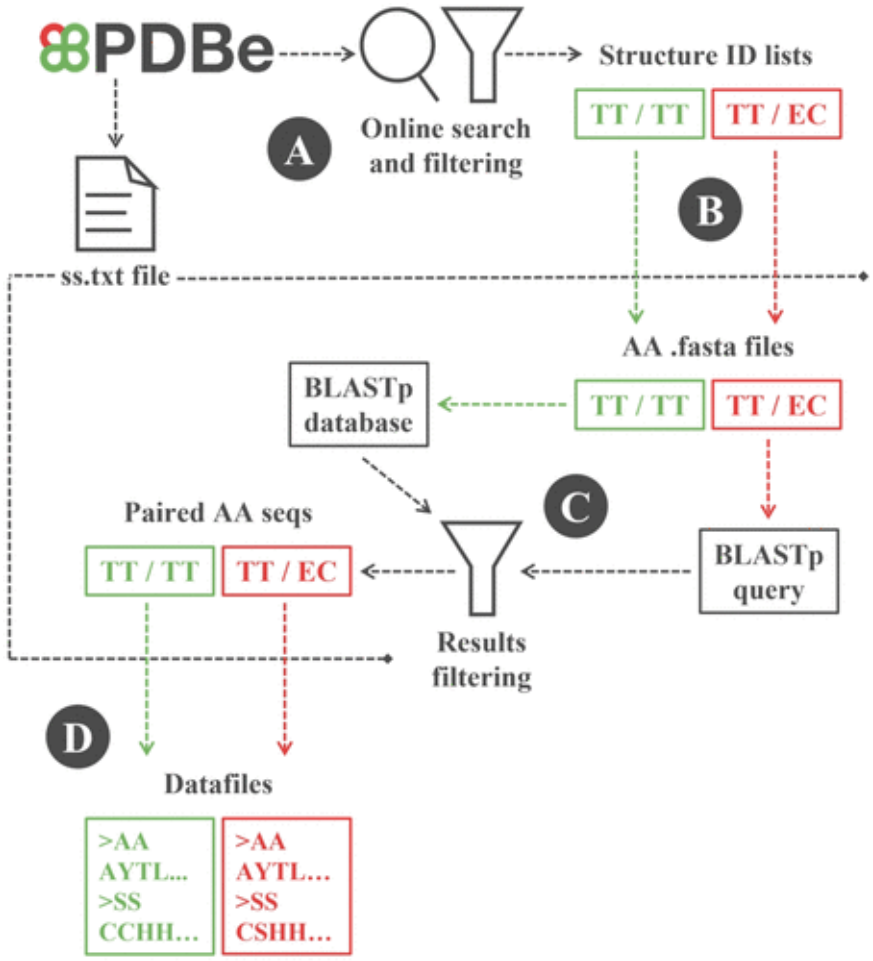

Figure 2.2. Overview of methodology (part 1). Relevant structure IDs were found through a search on PDBe (A) and processed into AA.fasta files (B). The found structures with the same protein origin species and different expression species were paired on the basis of AA sequences using BLASTp (C), and the paired AA/SS were saved as datafiles (D). Figure adapted from (Kruglikov, Wei, and Xia 2022).

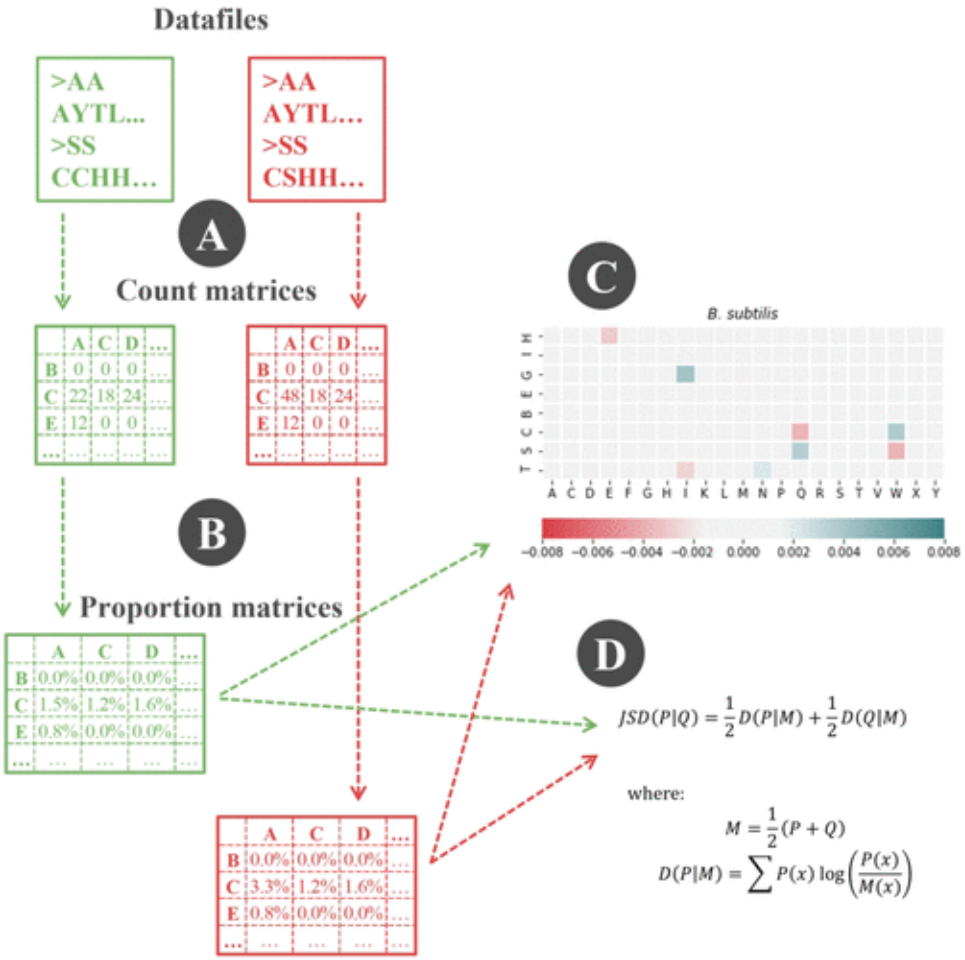

Figure 2.3. Overview of methodology (part 2). Datafiles were used to construct AA/SS count (A) and proportion matrices (B). For each studied species, JSD was calculated between the "native" proportion matrix and the recombinant *E. coli* proportion matrix (D). The differences between the matrices were also visualized using heat maps (C). Figure adapted from (Kruglikov, Wei, and Xia 2022).

2.3.1 Data Collection

Bacterial species from which we collected the data are listed in Table 2.1 and were selected because they have sufficient data available on PDB both as protein source and expression systems. The PDB online advanced searching tool was used to create separate ID lists for each source/expression pair. More specifically, PDB Europe (PDBe) was used because of its more advanced filtering interface; however, its data are the same as on PDB. Advanced search can be

accessed through this link:
https://www.ebi.ac.uk/pdbe/entry/search/index/?advancedSearch:true=.

Table 2.1. Species Used in the Analysis [a]

protein source species	expression system	structure IDs list	entries before BLASTp	entries after BLASTp
Bacillus subtilis	*B. subtilis*	BS_BS	8	4
Bacillus subtilis	*E. coli*	BS_EC	1145	17
Desulfovibrio vulgaris	*D. vulgaris*	DV_DV	26	4
Desulfovibrio vulgaris	*E. coli*	DV_EC	67	20
Lactococcus lactis	*L. lactis*	LL_LL	25	5
Lactococcus lactis	*E. coli*	LL_EC	166	6
Pseudomonas fluorescens	*P. fluorescens*	PF_PF	13	4
Pseudomonas fluorescens	*E. coli*	PF_EC	275	2
Pseudomonas putida	*P. putida*	PP_PP	3	1
Pseudomonas putida	*E. coli*	PP_EC	542	37
Salmonella enterica	*S. enterica*	SE_SE	34	29
Salmonella enterica	*E. coli*	SE_EC	860	17
Streptomyces rubiginosus	*St. rubiginosus*	SR_SR	17	17
Streptomyces rubiginosus	*E. coli*	SR_EC	11	11
Thermus thermophilus	*T. thermophilus*	TT_TT	19	7
Thermus thermophilus	*E. coli*	TT_EC	1091	3

[a] For each protein source species there are two expression systems used ─ that of the source species ("native" expression system) and *E. coli*. Lists of structure IDs were created for each source/expression pair. For example, BS_BS contains IDs of structures with *B. subtilis* as both

protein source organism and protein expression system, while BS_EC has IDs of structures with *B. subtilis* as protein source organism and *E. coli* as protein expression system.

We used several filters to obtain protein ID lists: (1) organism name, (2) expression host name, (3) resolution, and (4) molecule type. Protein source and expression system were set in accordance with Table 2.1, molecular type was set to "protein" and experimental method was set to X-ray diffraction only with resolution between 0 and 2.5 Å. Note that proteins are purified in their naturally folded state in the expression system, ideally in their functional forms, before crystallization and X-ray diffraction.

Structure IDs were taken for each protein origin/expression system pair found using the online search and used to obtain the corresponding SS and AA sequences. We used a PDB Secondary Structure file in FASTA format (latest version is accessible at https://cdn.rcsb.org/etl/kabschSander/ss.txt.gz ─ sometimes multiple refreshes of the page are required to obtain the file; alternatively, a copy of the file is available at https://github.com/alibekk93/project-protein_folding_distances/ss.txt.gz) to collect SS and AA. Each structure in this file is represented by AA and SS strings, where each single AA corresponds to a single SS at the same position. AA decoded as "X" is sometimes present and corresponds to an error in original experiment. The numbers of structure IDs identified for each ID list are shown on Table 2.1. Identical AA sequences from the same expression systems were allowed as their corresponding SS sequences could differ and provide relevant data.

2.3.2 Filtering the Data Using BLASTp

To identify proteins whose structure has been determined when they were expressed in both the source species (native environment) and in *E. coli*, we processed the AA sequences obtained from the previous step into BLASTp database files using DAMBE. (Xia 2018) For each pair of compared protein databases a BLASTp search was performed using AA sequences of the proteins, with the recombinant *E. coli* expression system database as query and the native database as BLASTp database.

We used ungapped BLASTp with three-letter words to match proteins from different databases. Minimal matching identity was set to 95% to allow for small variations of AA sequence due to point mutations and random errors in experiment without allowing different proteins to be

matched. Minimal matching length was set to 50 to remove short matching sequences, and the maximal E-value was set to 0.01. BLASTp parameters were set so that only proteins with very similar AA chains and those likely to be the same protein were kept in both databases and so that only matching parts could contribute to the analysis. For each pair obtained, sequence start and end were used to cut out the matching parts of the sequences and not include the nonmatching parts.

Filtering protein databases using BLASTp removed most of the proteins from the original Protein ID lists. This happened because most entries on PDB are only available with one expression system (grown in *E. coli*). On the other hand, those entries available in the native expression system are also not always available with the *E. coli* expression system. For example, the original BS_BS database contained eight entries and the original BS_EC database had 1,145 entries; however, after BLASTp only four entries from BS_BS matched with 17 entries from BS_EC. The full list of numbers of entries in the databases before and after BLASTp is shown in Table 2.1.

2.3.3 Construction of Probability Matrices

First, tabular BLASTp results were used to create data files in FASTA format for each protein source species/expression system combination, each containing the AA and SS sequences. Structure IDs were used to obtain AA and SS sequences from the PDB file. Query and database start and end positions were used to cut the matching parts from the obtained sequences. This way only matching parts of the proteins would be left.

In many cases a single-query structure would match more than one database structure and vice versa; therefore, it was necessary to multiply AA and SS sequences in those cases to make sure that matching parts are the same length. In this way we obtained chains from the BLAST database with identical AA sequences, but possibly different SS sequences, and had each variation of SS in correct proportions. AA and SS sequences were concatenated for each of the databases so that all AA sequences were in one line and all SS sequences were in another line of the resulting file.

Data files were processed into count matrices to count each AA/SS combination. We converted count matrices into probability matrices by simple division of each count value by count matrix

sum. Protein SS can be described with three SS types: H (helix), E (sheet), and C (coil) or with eight SS types: H (α-helix), I (π-helix), G (310-helix), E (β-sheet), B (β-bridge), C (coil), S (bend), and T (turn). We refer to the two classification systems as 3-SS (three types of SS) and 8-SS (eight types of SS) in this work. While many models and studies, especially the earlier ones, use 3-SS, using 8-SS can give more details. In order to work with both 8-SS and 3-SS, we transformed the original 8-SS matrices to 3-SS by adding up the matrix values.

2.3.4 Jensen–Shannon Divergence Calculation and Statistical Analysis

After the matrices were created, they were compared to each other, so that for each protein source species the two matrices compared were with *E. coli* and "native" expression systems. Jensen–Shannon divergence (JSD) was used to evaluate differences between matrices. A common measure of probability distribution differences is Kullback–Leibler divergence (D_{KL}). It is nonsymmetric, which means that D_{KL} of distribution P from distribution Q does not have to be equal to D_{KL} of distribution Q from distribution P. D_{KL} is defined as

$$D_{KL}(P||Q) = \sum P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

where $P(x)$ and $Q(x)$ are discrete probability distributions.

JSD is a metric similar to D_{KL}, and it could be called a symmetrized and smoothed version of it. It is defined as

$$\mathrm{JSD}(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)$$

where

$$M = \frac{1}{2}(P + Q)$$

and where $P(x)$ and $Q(x)$ are discrete probability distributions and D is D_{KL}.

In our case JSD is a more suitable measure than D_{KL} due to the large number of zeros in our data. D_{KL} requires a division of one probability by another, and when the denominator probability is zero (which is very often the case in our data), calculation results in infinity. Standard practice is to drop zeros completely, but in our case that would be dropping a very significant amount of our

data, because probability matrices contain a lot of zeros. Contrary to D_{KL}, JSD does not have this problem thanks to its symmetric nature. Probability distributions are compared not with each other but with their average distribution, which means that only positions where both probability matrices are zero need to be dropped ─ those positions are exact in any case and therefore dropping them is not an issue.

Larger JSD would mean that changing the protein expression system from "native" to *E. coli* had more effects and that a recombinant protein SS is a worse representative of native SS. Moreover, in order to deduce possible mechanisms of how the change of the expression system affects protein folding, we calculated difference between the matrices by subtracting recombinant proportion matrix values from native proportion matrix. Subtraction results would not be a good evaluation of matrix divergence but can show at which AA/SS positions the differences between matrices are large.

We estimated the statistical significance of calculated JSD using resampling techniques. Bootstrapping was used to calculate 95% confidence intervals of JSD ─ AA/SS positions were randomly resampled with replacement 1,000 times from the original data files and JSD were calculated for them. In addition, we used permutation to test the significance of differences between species' JSD values by combining all data files into a uniform distribution and resampling positions from it 10,000 times to form data of 1,000 residues in length.

2.4 Results

2.4.1 Magnitude of Expression System Effect

Calculated results are summarized on Table 2.2. In addition, we visualize bootstrapping results on box plots (Figures 2.4 and 2.5) and permutation results on histograms (Figure 2.6). The largest JSD was found between *T. thermophilus* matrices, and this was the case using both 8-SS and 3-SS. Moreover, bootstrapping results show that *T. thermophilus* JSD is the only one significantly higher than other species' JSD for both 8-SS and 3-SS.

Table 2.2. Mean JSD between Native and Recombinant AA/SS Matrices [a]

protein source species	JSD8	JSD8 *p*-value	JSD3	JSD3 *p*-value
Bacillus subtilis	0.010	0.082	0.001	0.697

protein source species	JSD8	JSD8 *p*-value	JSD3	JSD3 *p*-value
Desulfovibrio vulgaris	0.008	0.343	0.003	0.004
Lactococcus lactis	0.007	0.594	0.003	0.012
Pseudomonas fluorescens	0.009	0.270	0.002	0.101
Pseudomonas putida	0.002	1.000	0.000	1.000
Salmonella enterica	0.003	0.999	0.000	1.000
Streptomyces rubiginosus	0.004	0.995	0.001	0.992
Thermus thermophilus	0.033	0.000	0.014	0.000

[a] *p*-values are one-sided and are from bootstrapping analysis testing the null hypothesis that different expression systems have no effect on protein structure. JSD8 and JSD3 are the calculated JSD using 8 or 3 types of SS. Both JSD8 and JSD3 are the greatest for *T. thermophilus* matrices, and that is the only species where both metrics are significantly different from bootstrapped distributions. This indicates that switching the protein expression system from "native" to *E. coli* affects the folding of *T. thermophilus* proteins more than other species' proteins.

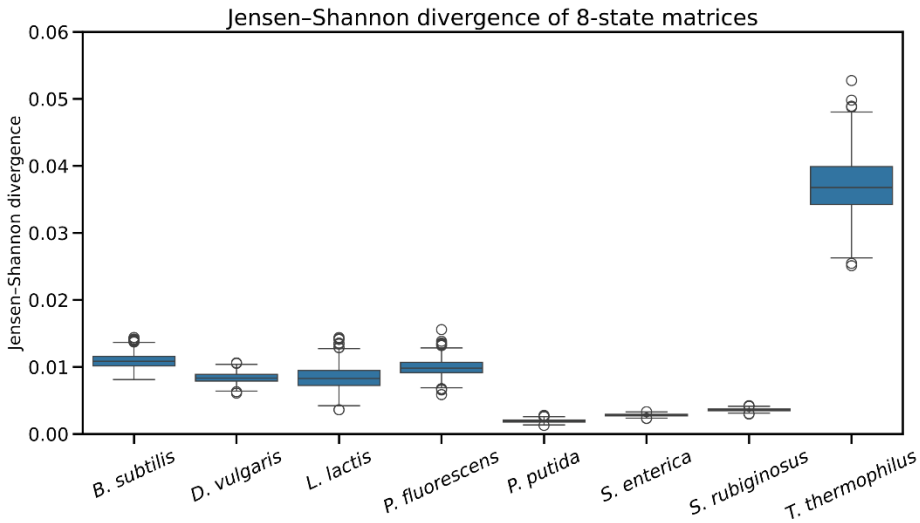

Figure 2.4. Box plots of bootstrapped JSD (8 SS types). High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria. Figure adapted from (Kruglikov, Wei, and Xia 2022).

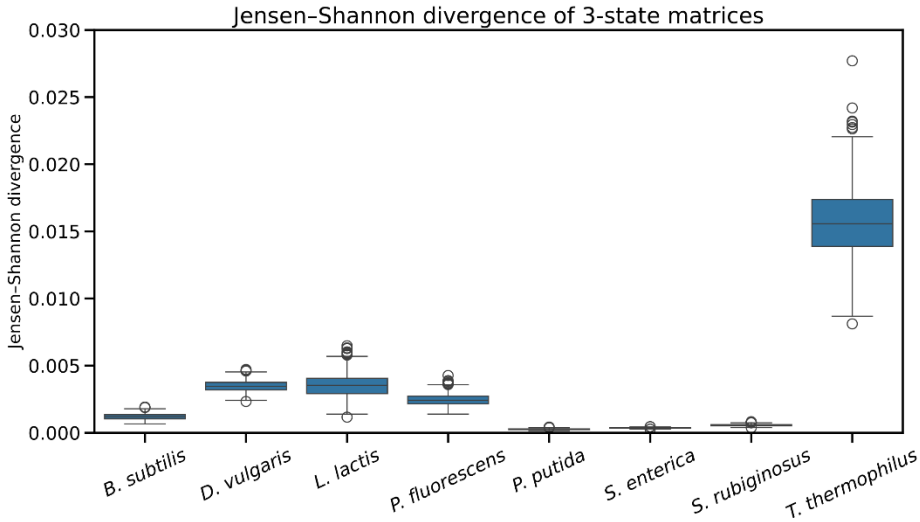

Figure 2.5. Box plots of bootstrapped JSD (3 SS types). High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria. Figure adapted from (Kruglikov, Wei, and Xia 2022).

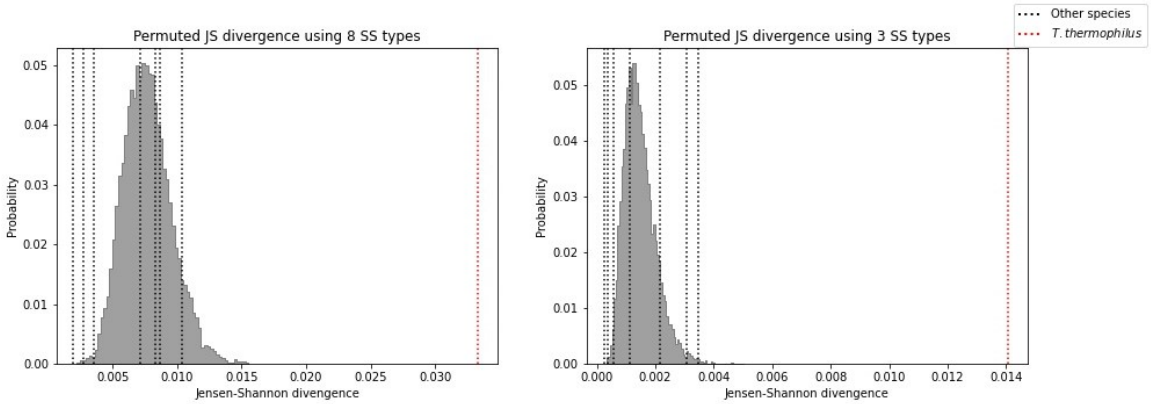

Figure 2.6. Distributions of permuted JSD results. *T. thermophilus* JSD (red line) is much larger than JSDs of the other species (gray lines) and the nonspecific JSD (gray histogram). Figure adapted from (Kruglikov, Wei, and Xia 2022).

2.4.2 Directionality of Expression System Effect

Heat maps in Figures 2.7 and 2.8 show differences between proportion matrices used in JSD calculations. These figures help in identifying which particular elements of matrices were most different and display potential directionality of the differences. In line with JSD results, *T. thermophilus* matrices show more differences than any other species' matrix pair. It can be seen that using *E. coli* as an expression system for thermophilic proteins leads to lower frequencies in helices and higher frequencies of coils. The 8-SS heat map shows that this effect is particularly strong on 310-helices (structure G). Other protein expression systems considered in this study show smaller differences from that of *E. coli* as there are no large JSD values detected for them.

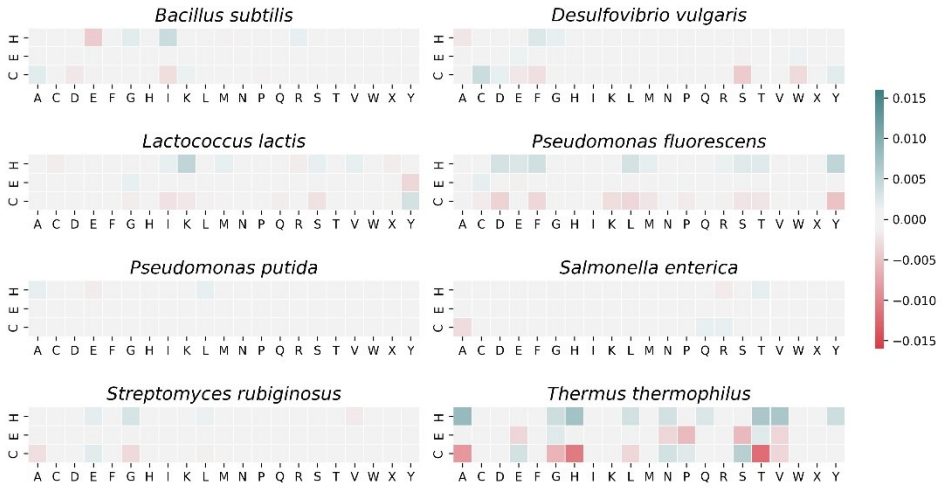

Figure 2.7. Heat maps of proportion matrices differences with 3 SS types showing directionality of the effect induced by using *E. coli* as the expression system. More negative values (red) indicate larger proportions in *E. coli* as the expression system; more positive values (green) indicate larger proportions in "native" expression systems. The effects were most visible in *T. thermophilus*, where helices (H) were observed more frequently when proteins were expressed in *T. thermophilus* and coils (C) were instead more abundant when proteins were expressed in *E. coli*. No such effect nor directionality of differences could be seen in other species. The three SS types are H (helix), E (sheet), and C (coil). Figure adapted from (Kruglikov, Wei, and Xia 2022).

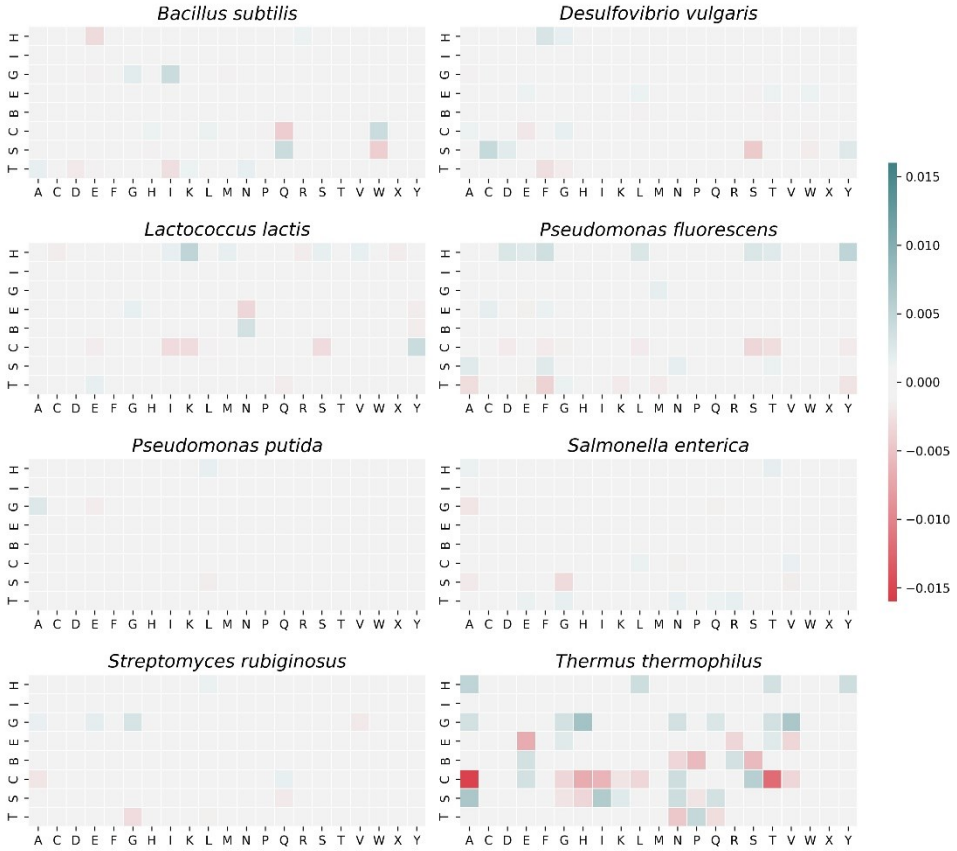

Figure 2.8. Heat maps of proportion matrices differences with 8 SS types showing directionality of effect induced by using *E. coli* as the expression system. More negative values (red) indicate larger proportions in *E. coli* as the expression system; more positive values (green) indicate larger proportions in "native" expression systems. Strong effects could be observed in *T. thermophilus*, where α-helices (H) and 310-helices (G) were observed more frequently when proteins were expressed in *T. thermophilus* and coils (C) were instead more abundant when proteins were expressed in *E. coli*. No such effect nor directionality of differences can be seen in other species. The eight SS types are H (α-helix), I (π-helix), G (310-helix), E (β-sheet), B (β-bridge), C (coil), S (bend), and T (turn). Figure adapted from (Kruglikov, Wei, and Xia 2022).

No strong patterns have been discovered in terms of variability of secondary structures between different amino acids (Figures 2.7 and 2.8). While distances between matrices of *T. thermophilus* proteins seem to be high with hydrophobic alanine, valine, and glycine, that is also the case for histidine (charged) and threonine (polar and uncharged).

2.5 Discussion

2.5.1 Lack of Required Chaperones

There are several possible explanations for variations in JSD for different species. Because *T. thermophilus* is a thermophile, it is adapted to protein denaturation, partially through chaperone-dependent protein folding. It is possible that when *E. coli* is used as an expression system, certain helices are not formed or repaired due to a lack of these chaperones and coils are formed instead. For example, DnaK chaperone expression requires less ATPase activity in *T. thermophilus* than in *E. coli* and it participates in protein folding mediation (Schlee and Reinstein 2002). Trigger factor proteins also show differences in structure and activity between the species (Godin-Roulling et al. 2015). It is possible that such effects are different in intrinsically unstructured proteins that require chaperone activity for correct folding and structural stability (Gsponer et al. 2008). While a common way around this problem is to co-express required folding chaperones together with the studied protein, it is not always clear whether that was done on PDB because not all structures there have publications and, even if they do, it is not always clearly explained whether co-expression of chaperones was performed. Moreover, co-expression of chaperones does not always provide the desired effects as differences in other cell-specific factors may lead to chaperones losing their activity or even becoming toxic for the host (Sahdev, Khattar, and Saini 2008).

2.5.2 Suboptimal Cellular Environment

Due to thermophilic adaptations of *T. thermophilus*, it is possible that the tendency toward coil structures instead of helical structures in *E. coli* as an expression system is a result of differences in cellular conditions of the expression systems, namely, nonoptimal folding temperatures. This would explain why this effect is more apparent for 310-helices than α-helices, as the latter are more stable (Rohl and Doig 1996). Regardless, future research directions may prompt researchers to study varying environmental conditions of expression systems and their effect on protein folding using data with varying temperature, pH, and salinity.

Differences in protein solubility due to temperatures could have affected protein crystallization and thus structure identification (Mikol and Giegé 1989; Chayen et al. 1988; McPherson 1985). To assess this possibility, an analysis of structures with different crystallization techniques performed might be necessary. That kind of analysis would help to determine whether effects

observed in this experiment are due to differences in cellular environments and chaperones or due to experimental design.

2.5.3 Codon Optimization

Protein folding can be affected by the rate of protein synthesis, which can be controlled by codon usage (Plotkin and Kudla 2011; Zylicz-Stachula et al. 2014). Assuming equal translation initiation rates, nonoptimal codon usage in the recombinant expression system will lead to slower rates of protein production, which in turn may lead to protein misfolding and aggregation (Nedialkova and Leidel 2015). Unfortunately, the PDB itself has no information about codon optimization in experiments and nucleotide sequences are not provided. Moreover, not all records have corresponding publications with full description of experimental and even the ones that had been published often do not have information on whether codons were optimal. This means that lack of codon optimization is a potential factor that caused high differences between *E. coli* and *T. thermophilus* expression systems in our analysis.

2.5.4 Protein Crowding

Macromolecular crowding is another potential explanation of our results. As concentrations of proteins and other macromolecules inside the cells increase, the volume available for new proteins being produced falls (Ellis 2001; Kinjo and Takada 2002; Miklos et al. 2011). Crowding leads to an increase in protein thermodynamic activity, which affects folding, among other processes (J. S. Kim and Yethiraj 2011; Samiotakis and Cheung 2011; Kuznetsova, Turoverov, and Uversky 2014). In prokaryotes this effect is more profound than in eukaryotes (Ellis 2001).

Naturally, protein crowding may occur in both *E. coli* and *T. thermophilus* expression systems, as well as other systems, but recombinant expression systems are more likely to have this effect (Westphal et al. 2017; Ninh et al. 2015). *E. coli* natural cellular concentration could be unsuitable for *T. thermophilus* protein folding and may lead to increased crowding and inclusion body formation (Sørensen and Mortensen 2005). In addition, chaperone-assisted misassemble prevention mechanisms may be compromised in recombinants as they would lack the required chaperones (Westphal et al. 2017; Hartl, Bracher, and Hayer-Hartl 2011).

2.5.5 Significance

The connection between protein structure and protein function is established, as structure directly dictates function (Berg, Tymoczko, and Stryer 2002). Improperly folded proteins often lose their initial functions and can even gain novel toxic functions in their place. For example, many misfolded proteins related to Parkinson's and Alzheimer's diseases have neurotoxic functions (Winklhofer, Tatzelt, and Haass 2008). It is possible that thermophilic proteins grown in *E. coli* lose their functions entirely or partially. Previous studies identified that *T. thermophilus* enzyme activity is reduced when using mesophilic recombinant hosts, such as *E. coli* (Hidalgo et al. 2004; Krefft et al. 2017; Fujino et al. 2020; Goda et al. 2005). Additionally, *E. coli* had been previously shown to be an inadequate expression system for thermophilic proteins in functional metagenomic (Angelov et al. 2009; 2011; Leis et al. 2015) and directed evolution studies (Chautard et al. 2007; Mate et al. 2020; Bosch et al. 2021). Our results are in line with the previous findings; we also expand on them, showing how a mesophilic expression system can affect secondary structures of thermophile proteins and that the change has directionality toward less helices and more coils.

Additionally, helical structures have been shown to be more common in thermostable proteins and are associated with thermostability (Miotto et al. 2019; Vogt and Argos 1997). The higher tendency for helix formation for *T. thermophilus* proteins when grown in "native" expression system could be a mechanism of protein stabilization under higher temperatures. Using *E. coli* as an expression system led to higher proportions of coils and therefore might have reduced thermostability adaptation.

In some cases, thermophile protein misfolding can be removed by subunit rearrangement caused by heating of the protein (Goda et al. 2005). However, that is not very common and more often the problem of misfolding can be solved by using thermophiles as expression systems for thermophilic proteins can be a solution to the problem of misfolding (Hidalgo et al. 2004; Fujino et al. 2020). These facts suggest that the protein expression system cellular environments need to be matched with those of protein source organisms in order to facilitate correct folding.

2.5.6 Study Limitations

Our study has multiple limitations which we should address here as well. First, it is evident that the number of protein structures which remain after all filtering procedures is very small for many of the species, including *T. thermophilus*. While we attempt to lower the impact of the low number of structures with resampling, it is still possible that the differences that we observe are related to specific protein structures or even by some errors in PDB experiments. In addition to the main results, described previously, we calculated JSD between matrices without BLASTp filtering. The rest of the procedures were kept the same as before. This way we could greatly increase the data size; however, the drawback is that proteomes now consisted of very different proteins and therefore these results cannot be fully conclusive either. Nevertheless, we found that JSD between *T. thermophilus* matrices is much higher than for all other species, the same as with the main results. We provide these additional results in Figures 2.9 and 2.10.

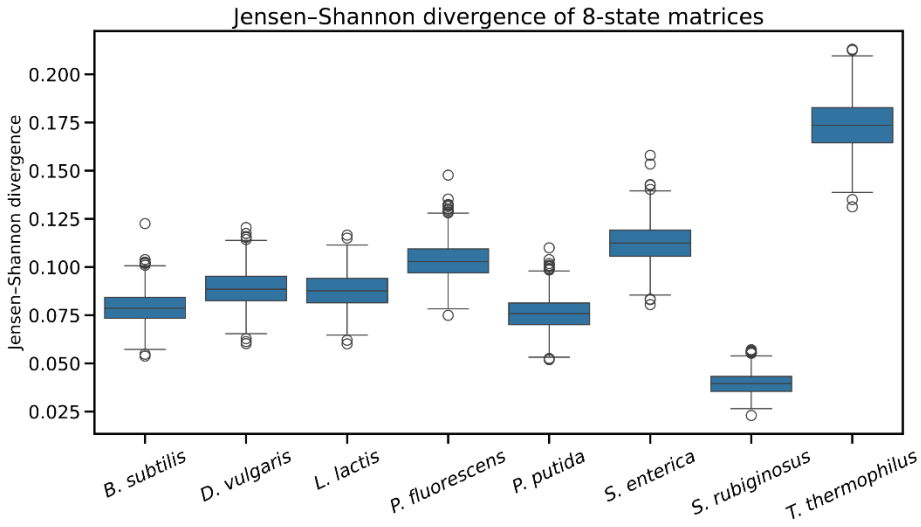

Figure 2.9. Boxplots of bootstrapped JSD (8SS types) — data with no BLASTp filtering. High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria. Figure adapted from (Kruglikov, Wei, and Xia 2022).

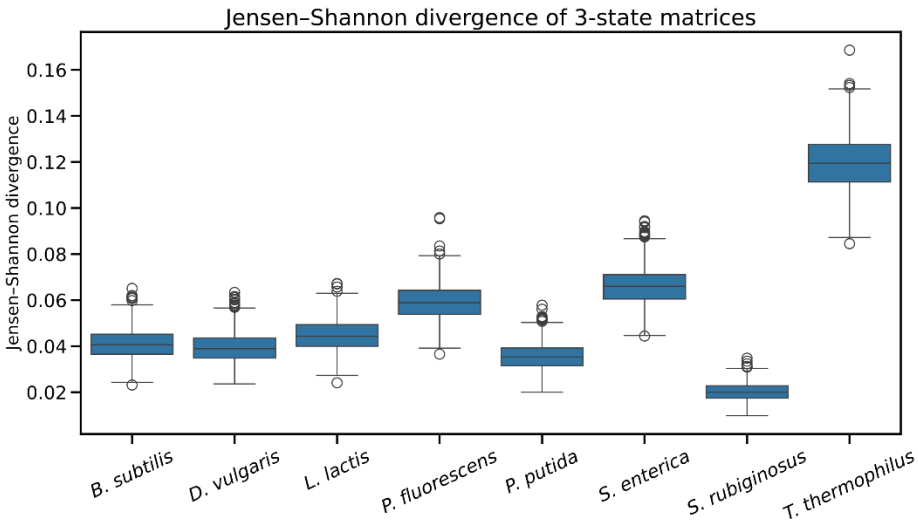

Figure 2.10. Boxplots of bootstrapped JSD (3 SS types) — data with no BLASTp filtering. High JSD indicates larger differences between "native" and *E. coli* expression systems. *T. thermophilus* JSD are much higher than those of other bacteria. Figure adapted from (Kruglikov, Wei, and Xia 2022).

Second, our approach of using matrices in calculating JSD is double-edged. On one hand, using matrices allows us to compare entire proteomes rather than single proteins in a simple and computationally efficient way. This way we can compare proteomes which consist of very different proteins. On the other hand, only a pairwise comparison would show what effect protein type has on differences in folding. Ideally, all proteomes in our study should consist of the same proteins and in that case a pairwise comparison would be highly advantageous. Additionally, for the sake of simplicity and easier interpretation, our matrices were computed using one-to-one AA/SS pairing. This approach neglects potential effects that neighbor AA has on SS. It may be beneficial to use windows of several AA/SS to compute matrices in future research.

As we allowed AA sequences to differ by 5% during our BLASTp filtering step, the datafile AA sequences had some level of variation and this could have an effect on SS and JSD8/JSD3. We looked at how AA similarity affected JSD8 and JSD3, and there seems to be no relationship (Figures 2.11 and 2.12). *T. thermophilu*s large JSD8 and JSD3 results are highly unlikely to be

explained by AA differences; however, this is still possible due to the complex nature of the AA/SS relationship and this possibility should not be ignored completely.

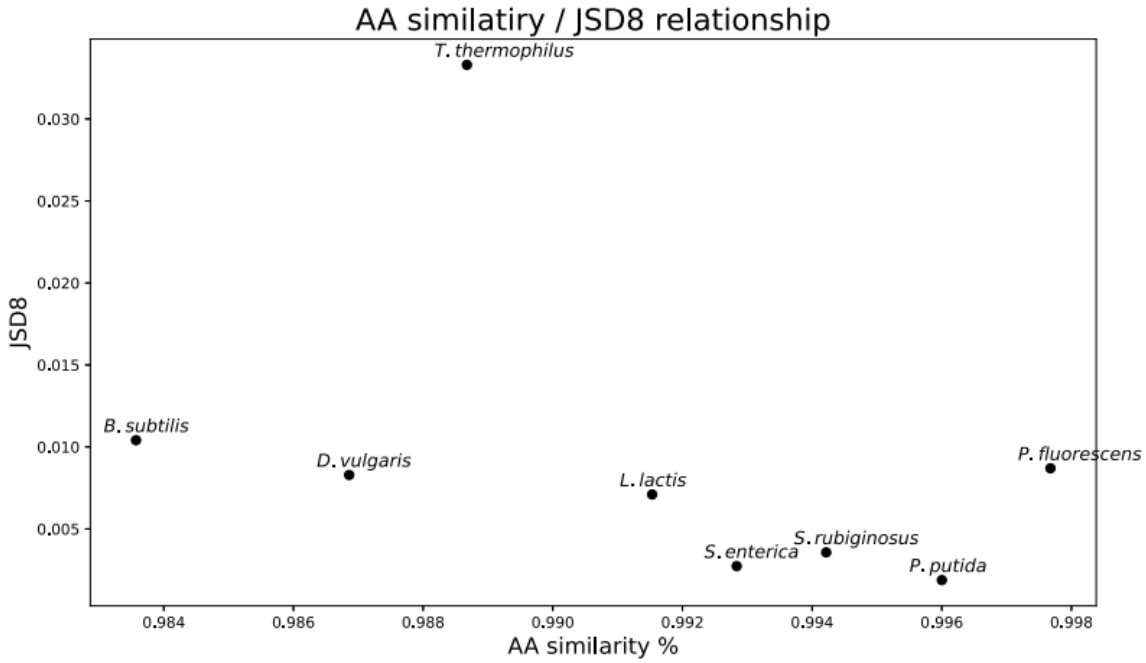

Figure 2.11. Scatterplots of overall datafile AA similarity / JSD (8SS types) relationship. AA similarity variation seems to have no significant effect on JSD8. Figure adapted from (Kruglikov, Wei, and Xia 2022).

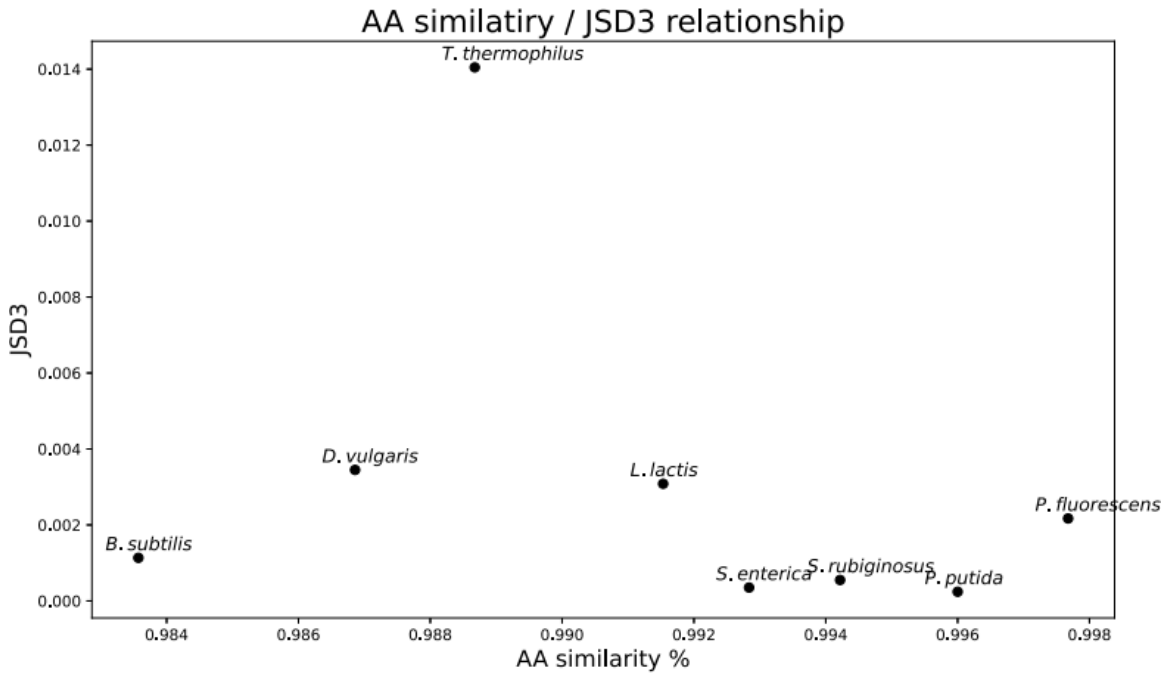

Figure 2.12. Scatterplots of overall datafile AA similarity / JSD (3SS types) relationship. AA similarity variation seems to have no significant effect on JSD3. Figure adapted from (Kruglikov, Wei, and Xia 2022).

We believe that obtaining more structural data would be essential in order to design a study which would not have the limitations that we discussed. This is especially the case with data of expression systems other than *E. coli*. Often predicted structures could be used when PDB does not have sufficient data; however, to our knowledge, no protein structure prediction model has an expression system as a feature.

2.6 Conclusion

In conclusion, our results show that thermophilic protein folding in mesophilic *E. coli* introduces significant changes on the structure level. Misfolding of thermophilic proteins grown using mesophilic hosts can lead to loss or change of protein functions which will harm both research and industrial applications. While there can be many possible explanations for the reasons of misfolding, it is important to study *T. thermophilus* and other extremophiles protein expression with protein source species as protein expression systems in order to minimize expression system effects. It is also evident that a much higher diversity of expression systems on PDB is essential for more thorough understanding of expression system effects on protein folding.

Chapter 3. Applications of Protein Secondary Structure Algorithms in SARS-CoV-2 Research

Alibek Kruglikov[1], Mohan Rakesh[1], Yulong Wei[1], Xuhua Xia[1,2]

1. Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa,

Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486.

2. Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada K1H 8M5.

This chapter was originally published as: Kruglikov, A., Rakesh, M., Wei, Y. & Xia, X. (2021). Applications of Protein Secondary Structure Algorithms in SARS-CoV-2 Research. Journal of Proteome Research, 20(3), 1457–1463. doi: 10.1021/acs.jproteome.0c00734.

Author contributions: A.K. and M.R. wrote the manuscript with support from Y.W. and X.X. M.R. produced secondary structure predictions; A.K. designed and performed original research described in the study. X.X. supervised the project. All authors reviewed the manuscript.

3.1 Abstract

Since the outset of COVID-19, the pandemic has prompted immediate global efforts to sequence SARS-CoV-2, and over 450 000 complete genomes have been publicly deposited over the course of 12 months. Despite this, comparative nucleotide and amino acid sequence analyses often fall short in answering key questions in vaccine design. For example, the binding affinity between different ACE2 receptors and SARS-COV-2 spike protein cannot be fully explained by amino acid similarity at ACE2 contact sites because protein structure similarities are not fully reflected by amino acid sequence similarities. To comprehensively compare protein homology, secondary structure (SS) analysis is required. While protein structure is slow and difficult to obtain, SS predictions can be made rapidly, and a well-predicted SS structure may serve as a viable proxy to gain biological insight. Here we use predicted SS to compare ACE2 proteins and to evaluate the zoonotic origins of viruses. As computational tools are much faster than wet-lab experiments,

these applications can be important for research especially in times when quickly obtained biological insights can help in speeding up response to pandemics.

3.2 Introduction

Since the outbreak of COVID-19 in late December of 2019, more than 450,000 full genomes of SARS-CoV-2 have been sequenced and deposited in GISAD database (https://www.gisaid.org/, last accessed February 1, 2021). Both SARS-CoV-2 (Zhou et al. 2020) and SARS-CoV (G. Lu, Wang, and Gao 2015; Hulswit, de Haan, and Bosch 2016; Hoffmann, Hofmann-Winkler, and Pöhlmann 2018) encode a Spike (S) protein, hereafter respectively referred to as SARS-2-S and SARS-S. The S1 receptor binding domain (RBD) binds to host Angiotensin-converting enzyme 2 (ACE2) receptor to mediate cell entry. The efficacy of this interaction determines host specificity and severity of infection (Coutard et al. 2020; Hoffmann, Hofmann-Winkler, and Pöhlmann 2018; Andersen et al. 2020). Given a mammalian species, a high similarity between human ACE2 (hACE2) and mammalian ACE2 at S protein contact sites implies high susceptibility, and one can expect to determine species susceptibility to SARS-CoV or SARS-CoV-2 infections by comparative amino acid sequence analyses at contact sites at the ACE2 receptors.

3.2.1 Secondary Structure Studies are Required to Understand Host Susceptibility to SARS-CoV-2

The above expectation, while largely correct, is not completely accurate. For example, of the 18 amino acid (AA) sites in contact between hACE2 and the RBD of SARS-S, nine AA sites differ between ferret ACE2 and hACE2, but both ferret ACE2 and hACE2 are effective as receptors for binding to RBD and mediating viral entry into host cells. In contrast, ACE2 from mouse and rat also differ from hACE2 by nine AA sites, but they cannot support viral RBD binding and viral entry (G. Lu, Wang, and Gao 2015). This discrepancy invokes two simple explanations. First, AA sites beyond the 18 contact sites may also contribute to structural interactions and those sites might be more similar between hACE2 and ferret ACE2 than between hACE2 and mouse and rat ACE2. Second, structural similarity is not fully reflected in sequence similarity; i.e., structural similarity between hACE2 and ferret ACE2 may be greater than that between hACE2 and the mouse and rat ACE2. Only through structural studies can we hope to gain mechanistic insights into the differences in mammalian susceptibility to SARS-CoV-2.

Nevertheless, protein structure is difficult to obtain, and well-predicted protein secondary structure (SS) may serve as the next best answer. The Protein Data Bank (PDB) is the main depository of experimentally determined 3D protein structures, and around 160 thousand protein structures are deposited (Burley et al. 2017). In comparison, over 216 million AA sequences can be found in the NCBI GenBank database as of May 2020 (Clark et al. 2016). This inequality arises because experimental determination of structures is an expensive and lengthy process (Terwilliger, Stuart, and Yokoyama 2009; Mardis 2006).

In silico structure prediction techniques are faster and cheaper, and they have been useful in many research areas. For example, SS predictions have been used in enzyme structure similarity calculations (Rehman et al. 2020), ribosomal protein comparison (Anger et al. 2013), protein activity mechanisms (Y. I. Wu et al. 2009), COVID-19 proteomics (Jumper, J. et al., n.d.), and many other areas. In this work we review examples of protein secondary structure predictions (PSSP) algorithms, and their practical uses in pandemics research. We also describe an example of our own PSSP analysis on S protein-ACE2 binding to study species' susceptibility to SARS-CoV and SARS-2-CoV. The examples described highlight how PSSP can be a useful tool in pandemics research.

3.2.2 An Evaluation of Current PSSP Algorithms

In protein structure models, AA sequences are used to predict secondary and tertiary protein structures. SS are often classified in either three states or eight states of structures. Early PSSP models predict three secondary structure types: helix (H), strand (E), and coil (C), whereas in recent years, PSSP models have shifted to predict structures in eight states. Figure 3.1 summarizes PSSP programs developed over the years.

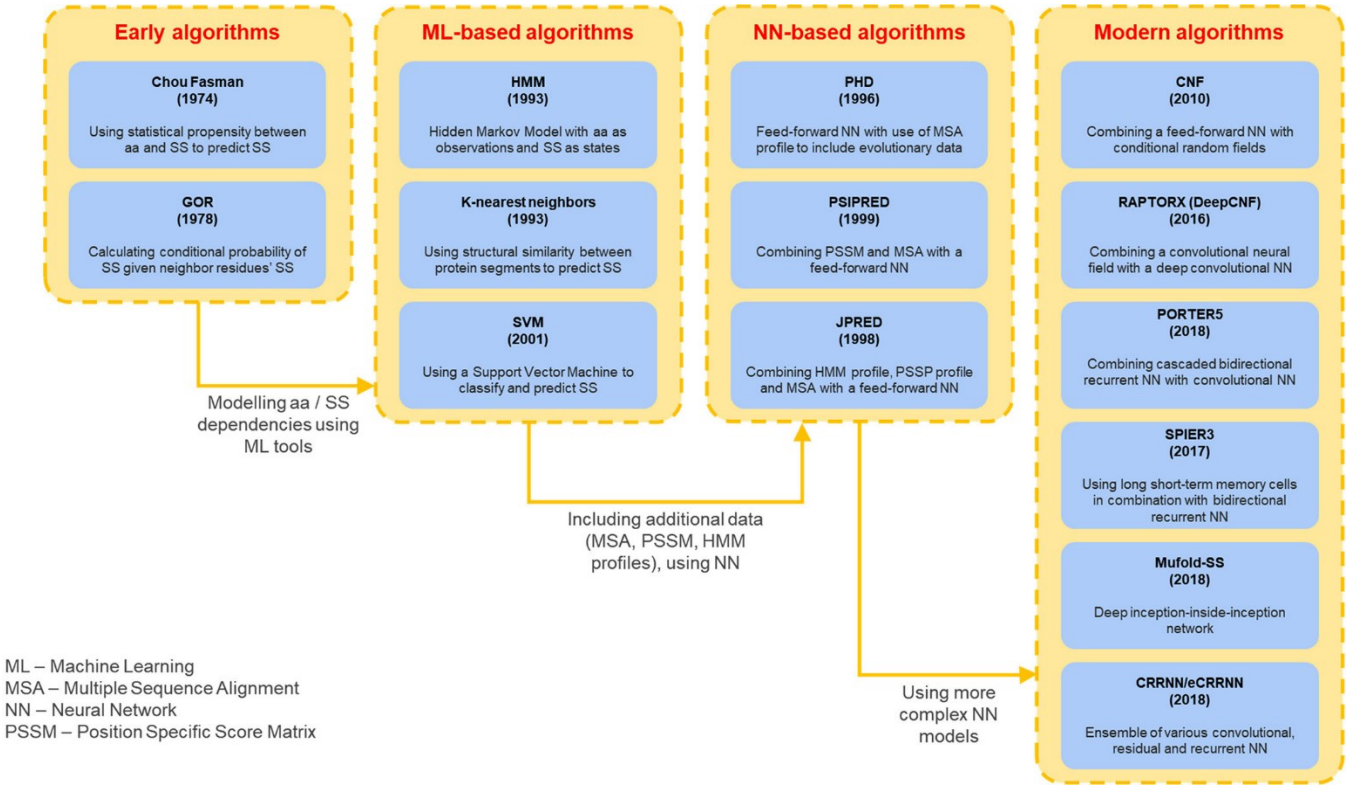

Figure 3.1. An overview of PSSP programs and implemented computational algorithms developed over the past 50 years (Chou and Fasman 1974; Kloczkowski et al. 2002; Asai, Hayamizu, and Handa 1993; Yi and Lander 1993; Hua and Sun 2001; Rost, Sander, and Schneider 1994; McGuffin, Bryson, and Jones 2000; Drozdetskiy et al. 2015; Z. Wang et al. 2010; S. Wang et al. 2016; Torrisi, Kaleel, and Pollastri 2018; Heffernan et al. 2017; B. Zhang, Li, and Lü 2018)**.** Figure adapted from (Kruglikov et al. 2021).

In addition to PSSP, protein structures can be modeled at the 2D level as contact maps (Yuan and Bystroff 2007) and at the 3D level as tertiary structures (Sarkar et al. 2015; Kwon et al. 2016). While modeling in 2D or 3D are appealing, there are several reasons why PSSP can be practical. First, unlike 2D or 3D structures, PSSP is reported as a sequence and can be used together with AA chains in multiple sequence alignments. This makes PSSP modeling useful in determining proteins that might be more similar in structures than in nucleotide or AA sequence. Second, the sequential nature allows alignment of SS elements with known or exploratory protein hotspots. Lastly, PSSP is faster and less computation-heavy than 3D predictions.

Typically, three metrics are used to evaluate accuracy of PSSP programs: Q3, Q8, and Segment Overlap (SOV) scores. Q3 and Q8 represent the percentages of SS sequence positions correctly predicted by the models using three or eight structure states, respectively. SOV is a more complex measure that represents the percentage of segment overlap between predicted and correct sequences. Different protein databases can be used for the evaluation, and the best practice is to use multiple data sets. Tables 3.1 and 3.2 show a collection of different PSSP models' accuracies calculated using various protein datasets (S. Wang et al. 2016; Torrisi, Kaleel, and Pollastri 2018; Heffernan et al. 2017; Fang, Shang, and Xu 2018; B. Zhang, Li, and Lü 2018; Y. Yang et al. 2018; Smolarczyk, Roterman-Konieczna, and Stapor, n.d.). Note that models are continually retrained with new protein structures, so there are discrepancies in reported accuracy values. Also, depending on data sets and metrics used, results of PSSP programs comparisons vary.

Table 3.1. A Comparison of PSSP Programs by Q3 Accuracy Assessments [a]

Program	TS115 (%)	CASP10 (%)	CASP11 (%)	CASP12 (%)	TS2019 (%)	CB513 (%)
JPRED4	77.1	81.6	80.4	78.8	76.6	81.7
PSIPRED v4.0	80.2	81.2	80.7	80.5	82.3	79.2
CNF	–	78.9	79.1	–	–	78.3
RAPTORX (DeepCNF)	82.3	84.4	84.7	82.1	–	82.3
SPIDER3	83.9	82.6	81.5	79.9	84.4	–
PORTER5	–	–	–	–	84.5	–
MUFOLD-SS	–	86.5	85.2	83.4	85.9	82.7
CRRNN	–	86.1	84.2	82.6	–	87.3
eCRRNN	–	87.8	85.9	83.7	–	87.8

[a] Accuracy scores (in percentage) are obtained from the programs' publication papers and from Yang et al. (Y. Yang et al. 2018) and Smolarczuk et al. (Smolarczyk, Roterman-Konieczna, and Stapor, n.d.).

Table 3.2. A Comparison of PSSP Programs by Q8 Accuracy Assessments [a]

program	CASP10 (%)	CASP11 (%)	CASP12 (%)	TS2019 (%)	CB513 (%)
CNF	64.8	65.1	–	–	64.9
RAPTORX (DeepCNF)	71.8	72.3	69.8	–	68.3
PORTER5	–	–	–	73.6	–
MUFOLD-SS	76.5	74.5	72.1	74.9	70.6
CRRNN	73.8	71.6	68.7	–	71.4
eCRRNN	76.3	73.9	70.7	–	74.0

[a] Accuracy scores (in percentage) are obtained from the programs' publication papers and from Yang et al. (Y. Yang et al. 2018) and Smolarczuk et al. (Smolarczyk, Roterman-Konieczna, and Stapor, n.d.).

In addition to prediction accuracy, it is important to consider the programs' usability and their limitations. While some programs are readily available through web servers, predictions through server are often limited by sequence length or number. For example, Mufold-SS only allows sequences of up to 700 AA long and Jpred4 only allows sequences of up to 800 AA long. In addition, most web servers only allow prediction of one protein sequence at a time, which is often impractical when working with a large number of sequences. Standalone versions of the programs do not have the restrictions of the web servers.

3.3 PSSP Methods have been Used Widely in Pandemics Research

3.3.1 Structural Conformation at SARS-CoV nsp5 Protein

Lu et al. (J.-H. Lu et al. 2005) explored the structure of the SARS-CoV nsp5 gene. With reference to SARS-CoV strain GD, comparative sequence analyses with 110 strains at *nsp5* showed that five *nsp5* had mutations. Secondary structure predictions were performed at the five mutated strains using PSIPRED and the analysis showed that all five mutated strains had identical predicted secondary structure, which implies that *nsp5* encoded proteins retain a conserved structure and may be a better therapeutic target than more rapidly evolving genes.

3.3.2 Rapid Evolution of Pandemic Norovirus Genogroups

Bull et al. (Bull et al. 2010) examined RNA polymerase and capsid protein similarities in five norovirus genogroups, of which the GII.4genogroup was associated with acute gastroenteritis global outbreaks. To evaluate whether this highly pathogenic genogroup had a greater epidemiological fitness than the other four genogroups, rate of mutation at RNA polymerase and capsid secondary structures were modeled using the CPH models Server. (Lund, O. et al. 2002) The PSSP model revealed that the 15 varying amino acid residues on capsid were located on the exposed loops in GII.4. Moreover, more pathogenic genogroups had more similarities with GII.4 in structure than less pathogenic ones.

3.3.3 Identification of a Potential Inhibitor of H1N1 Neuraminidase

Seniya et al. (Seniya et al. 2014) studied the potential effect of the *Boesenbergia pandurata* metabolite 4-hydroxy panduratin A to inhibit spread of Influenza A H1N1 (swine flu) infection. Influenza has two major surface proteins, neuraminidase (NA) and hemagglutinin (HA), to facilitate viral breach into host cell. To evaluate the potential of 4-hydroxy panduratin A to dock into active binding pockets of H1N1 NA, a homology-based protein structure prediction program, Modeler 9.10 (Sali and Blundell 1993) was used. In addition, I-TASSER (J. Yang et al. 2015) prediction was also used in combination with ab initio methods of modeling. These steps required secondary structure templates which were predicted using the PSIPRED server and rated using Z scores in LOMETS (S. Wu and Zhang 2007). The combination of PSSP and I-TASSER enabled the downstream analysis of protein interactions between the viral NA and the plant metabolite.

3.3.4 Determining Conserved Segments of H7N9 Hemagglutinin

Sarkar et al. (Sarkar et al. 2015) examined the Avian Influenza A (H7N9) hemagglutinin (HA) protein to determine conserved HA regions that could serve as potential peptide vaccines. As aforementioned, HA is one of the two major surface proteins that facilitate viral entry into host cells. In addition, HA can also elicit an antibody response during infection. The PSSP server, SABLE (Adamczak, Porollo, and Meller 2005), was used to predict accessible surface area (ASA) in 120 HA sequences from H7N9 strains, and Jpred (Cuff et al. 1998) and HHpred (Söding, Biegert, and Lupas 2005) were used to verify results. ASA, like secondary structure, is a 1D prediction; the aa sequence is converted to a sequence of numerical values, between 0 and

100, that describes aa sites accessibility in the solvent. Eight highly accessible regions were predicted by ASA and through epitope prediction, four regions were found with promising immuno-genic potential.

3.3.5 Computationally Designed Peptides to Block Binding between SARS-2-S and Host ACE2

Good binding between SARS-2-S and host ACE2 receptor is crucial for viral entry into host cells. This interaction has been extensively explored by experimental research as a COVID-19 vaccine target and by computational research aiming to design competitive binding peptides (Huang, Pearce, and Zhang 2020a) to bring forth new avenues to COVID-19 treatment. Using computational tools EvoEF2 (Huang, Pearce, and Zhang 2020b) and EvoDesign, (Pearce et al. 2019) Huang et al. (Huang, Pearce, and Zhang 2020a) designed peptide sequences that potentially bind competitively to SARS-2-S to limit viral entry. On the basis of a hACE2 structure template, they explored thousands of peptide designs through 3D modeling and selected best candidates by SARS-2-S binding affinity scored by PSSP performed in EvoDesign. The computational nature of this study allowed results to be obtained rapidly; currently, the computationally designed peptides are being evaluated experimentally (Huang, Pearce, and Zhang 2020a).

3.4 Using PSSP Models to Gain Biological Insight into Sars-Cov-2 and SARS-CoV Infectivity

3.4.1 Materials and Methods

Focusing on SARS-CoV-2, we tested the ability of several PSSP programs to predict SS of hACE2 and SARS-2-S S1 domain. We used experimentally derived SS from ACE2 structures available on PDB (1r42:A, 6m0j:A, 6m18:B, 6m1d:B, and 6m17:B; S1: 6vxx:A, 6vyb:A, 6m0j:E, and 6m17:E) to compare with SS predictions – these scores are shown in Table 3.3. Relatively low scores may be explained by the fact that membrane protein structures are hard to predict. Another possible reason is that the training data used for the PSSP programs were not specific enough to predict ACE2 and S1 proteins more accurately. The Q8 results for PSIPRED and JPRED4, which only predict three structure states, were expected to be lower than that of PORTER5 and MUFOLD-SS, which predicted eight structure states. However, Q8 results were

similar for all four programs (Table 3.3), possibly because extra types of secondary structures are rare in the studied proteins.

Table 3.3. Average PSSP Program Accuracies as Measured Using ACE2 and Spike Protein Data from PDB [a]

protein set	metric	PORTER5 (Torrisi, Kaleel, and Pollastri 2018, 5) (%)	MUFOLD-SS (Fang, Shang, and Xu 2018) (%)	PSIPRED (McGuffin, Bryson, and Jones 2000) (%)	JPRED4 (Drozdetskiy et al. 2015, 4) (%)
totals (other 2 sets combined)	Q3	75.2	77.1	77.7	76.5
	Q8	62.8	64.0	61.0	60.9
	SOV	57.6	57.8	60.3	58.3
hACE2 (1r42:A, 6m0j:A, 6m18:B, 6m1d:B, 6m17:B)	Q3	81.2	82.0	82.0	80.5
	Q8	69.9	70.8	65.2	65.1
	SOV	71.2	67.5	72.3	69.7
SARS-2-S S1 (6vxx:A, 6vyb:A, 6m0j:E, 6m17:E)	Q3	67.8	71.0	72.4	71.4
	Q8	54.0	55.5	55.7	55.6
	SOV	40.6	45.8	45.4	44.0

[a] PDB IDs are shown below the set names. Q3 and Q8 represent prediction accuracy for 3 or 8 SS types; SOV represents Segment Overlap score.

As previously mentioned, mammalian susceptibility to SARS-CoV cannot always be accurately predicted by differences in ACE2 AA sequences. This problem can be viewed as a mismatch between empirical and theoretical results. Using ACE2 PSSP instead of AA sequences, we attempt to explain this mismatch. To showcase that PSSP can circumvent this mismatch, Table 3.4 shows the $P_{distance}$, a measurement of differences in predicted SS between hACE2 and other species' ACE2. Here, we choose to use Mufold-SS to predict ACE2 SS (Table 3.3). $P_{distance}$ is based on Q3 and Q8 scores:

$$P_{distance} = 1 - \frac{M}{L}$$

where M is the number of residues that are the same in both windows and L is sequence length (analogous to Q3/Q8 evaluations). Mufold-SS can be robust with three states but not with eight states, as it assumes equal weight for all SS differences.

3.4.2 Results and Discussion

Calculated $P_{distance}$ scores are shown below in Table 3.4.

Table 3.4. $P_{distance}$ scores between hACE2 SS and Mammalian ACE2 SS [a]

SS sequence	$P_{distance}$
NM_001135696_Macaca_mulatta (Macaque)	0.0286
XM_008988993_Callithrix_jacchus (Marmoset)	0.0298
GQ999936_Rhinolophus_sinicus (Chinese horseshoe bat)	0.0335
EF569964_Rhinolophus_pearsonii (Pearson's horseshoe bat)	0.0410
AY996037_Cercopithecus_aethiops (African green monkey)	0.0435
NM_001130513_Mus_musculus (Mouse)	0.0472
AY881174_Paguma_larvata (Civet)	0.0472
XM_005074209_Mesocricetus_auratus (Hamster)	0.0497
NM_001012006_Rattus_norvegicus (Rat)	0.0509
AB211998_Procyon_lotor (Raccoon)	0.0547
NM_001310190_Mustela_putorius_furo (Ferret)	0.0584
EU024940_Nyctereutes_procyonoides (Raccoon dog)	0.0622
NM_001039456_Felis_catus (Cat)	0.0634

[a] ACE2 SS are predicted by Mufold-SS. (Fang, Shang, and Xu 2018)

The $P_{distance}$ shows that SS variations better explain patterns of SARS-CoV infectivity than hotspot AA differences. First, unlike differences in ACE2 AA, differences in ACE2 SS corroborate the finding that rats (Holmes 2005) are less susceptible to SARS-CoV than palm civets (Guan et al. 2003) and mice (W. Li et al. 2004), with $P_{distance}$ of 0.0509 (rats) vs 0.0472 (palm civets and mice). Second, ACE2 SS explains why Chinese horseshoe bats ($P_{distance}$ = 0.0335) are more susceptible to SARS-CoV than Pearson's horseshoe bats ($P_{distance}$ = 0.0410)

(Hou et al. 2010). Nonetheless, our findings cannot be generalized further, as not all patterns of infectivity are explained through $P_{distance}$. For example, $P_{distance}$ cannot explain why palm civets (0.0472) are more susceptible to SARS-CoV than Pearson's horseshoe bat (0.0410) (Guan et al. 2003; Hou et al. 2010).

To further examine the ACE2 of species shown in Table 3.4, we calculated AA sequence similarities using the Lake94 (Lake 1994) phylogenetic distance with hACE2 as reference. Indeed, with respect to hACE2, AA sequence similarities as measured by Lake94 poorly reflect similarities at SS as measured by $P_{distance}$ in many species (Figure 3.2: R2 = 0.179, P = 0.150), an example is *Rhinolophus sinicus*.

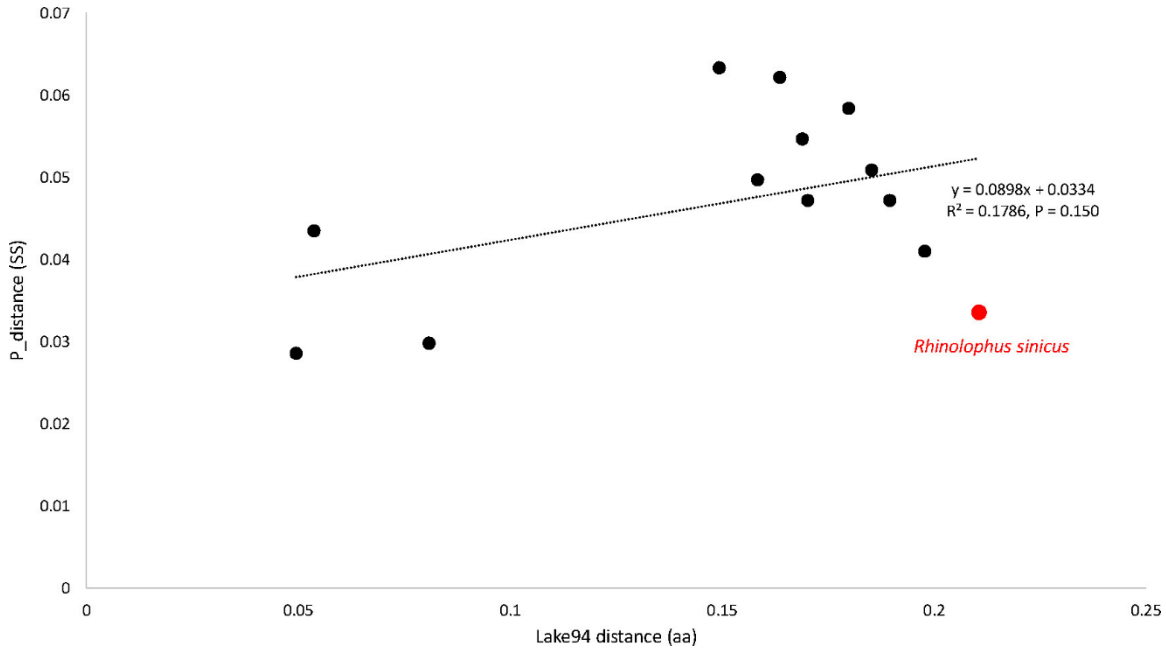

Figure 3.2. Lake94 distances measured at ACE2 AA sequences poorly correlate $P_{distance}$ measured at ACE2 SS. Sequence distances in mammalian ACE2 are calculated with respect to hACE2, and the 13 species considered are those listed in Table 3.4. Figure adapted from (Kruglikov et al. 2021).

We next performed multiple sequence alignment (MSA) using MAFFT (Katoh and Standley 2013) on ACE2 AA sequence and on predicted ACE2 SS sequence for *Rhinolophus sinicus* highlighted in red in Figure 3.2. Hotspot sites were highlighted in the alignment, representing hACE2 sites S19, Q24, D30, K31, H34, E35, E37, D38, Y41, Q42, L79, M82, Y83, K353, and

R393 that form contact with SARS-2-S at sites K417, G446, Y449, L455, F456, A475, F486, N487, Y489, Q498, T500, N501, G502, and Y505, as previously identified through X-ray crystallography experiments (Lan et al. 2020, 2; J. Shang et al. 2020).

Rhinolophus sinicus ACE2 seems to be more conserved at hotspot locations (boxed in light blue) than other regions at the SS level (Figure 3.3). Furthermore, lack of SS differences at some AA substitution sites can be explained by the nature of AA substitutions: some AA substitutions are considered conservative as they have similar physicochemical properties (Yampolsky and Stoltzfus 2005). Indeed, conservative D ↔ E, D ↔ N, E ↔ N, E ↔ Q, and K ↔ R are present at the regions boxed in yellow (Figure 3.3); these amino acids have similar properties and reduced substitution effects on predicted SS folding. On the other hand, some regions have many SS differences but relatively conserved AA (Figure 3.3: boxed in light red), one explanation for this discrepancy is that AA substitutions may influence SS at distant loci rather than closer ones due to complexities of hydrogen bond formation. Moreover, Lysine has been reported as preferred amino acids at C-terminus of proteins for α-helix formation (Forood, Feliciano, and Nambiar 1993), and reduced helix stabilization in the light red region could be caused by the K → N substitution.

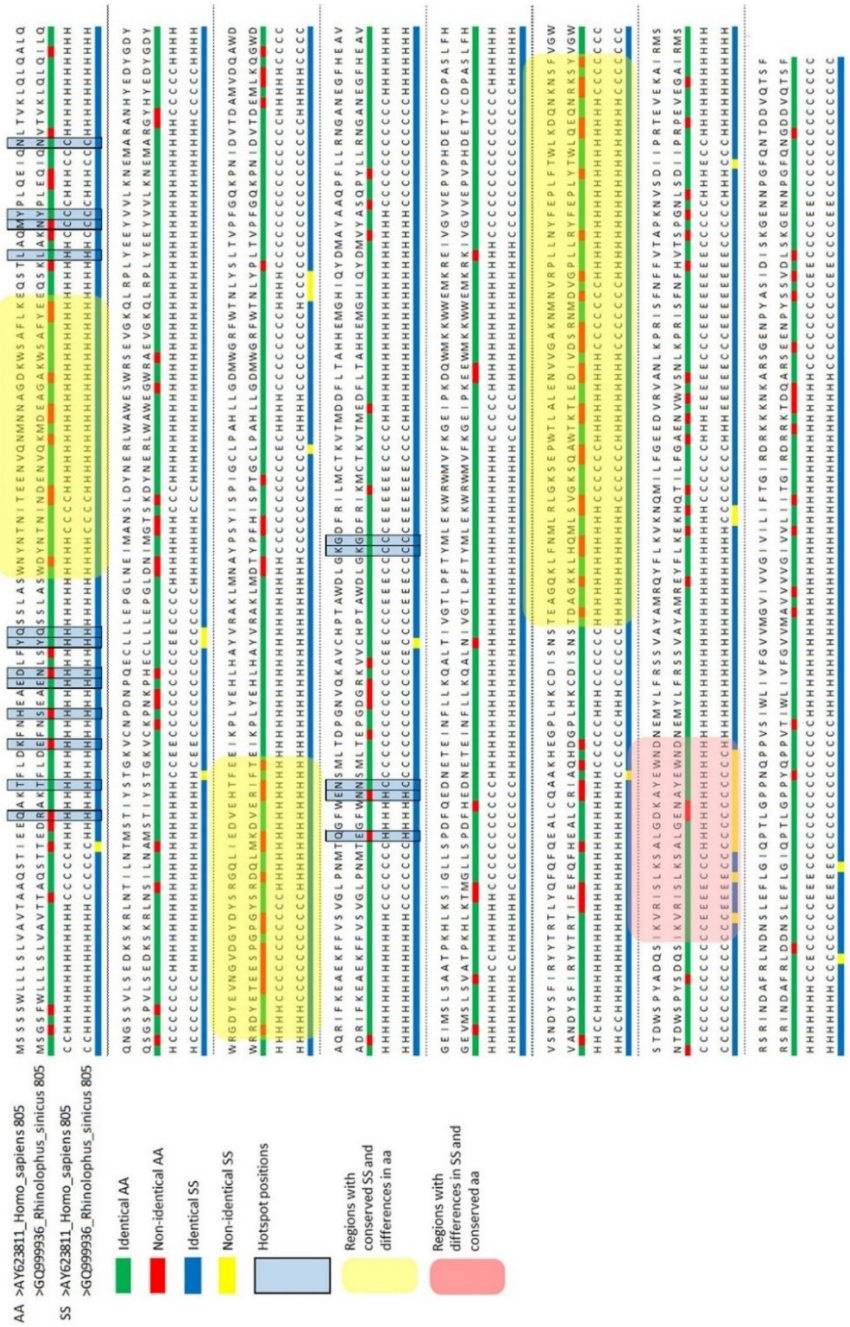

Figure 3.3. SS and AA alignments between *Rhinolophus sinicus* ACE2 and hACE2. Match and mismatch sites are respectively indicated by green and red for AA alignment and by blue and yellow for SS alignment. Notable regions where conservation levels differ between AA and SS alignments are boxed in light red and yellow. Hotspot positions boxed in light blue represent SARS-2-S contacting sites at hACE2 (Lan et al. 2020; J. Shang et al. 2020). Figure adapted from (Kruglikov et al. 2021).

3.5 Conclusion

PSSP programs can be applied to gain biological insights in rapid ways. These fast methods can be helpful to obtain important answers as an immediate response in pandemics research. Because some mutations, especially substitutions, might not induce structural changes, analysis on SS expands upon analysis of AA. In this study we evaluated some of the current PSSP programs and offered an example of PSSP analysis with a focus on SARS-CoV-2. Because coronavirus infection is achieved through binding between the viral Spike protein and the host ACE2 receptor, mammals with similar ACE2 structures could be potentially susceptible to these viruses. To identify ACE2 similarities between mammals and humans, comparisons were made at AA and SS levels. We showed that variations between predicted SS is not always consistent with variations in corresponding AA sequences. Specifically, differences at AA rarely led to different SS at ACE2 hotspot locations in *Rhinolophus sinicus*. The example above highlights potential application of PSSP algorithms in pandemics research.

Chapter 4. Comparative Analysis of Intrinsically Disordered Proteins in Mesophilic and Thermophilic Bacteria: Implications for Growth Temperature Adaptations

Alibek Kruglikov[1], Xuhua Xia[1,2]

1. Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa,

Ontario, Canada, K1N 6N5. Tel: (613) 562-5800 ext. 6886, Fax: (613) 562-5486.

2. Ottawa Institute of Systems Biology, Ottawa, Ontario, Canada K1H 8M5.

This chapter was originally published as: Kruglikov, A. & Xia, X. (2024). Mesophiles vs. Thermophiles: Untangling the Hot Mess of Intrinsically Disordered Proteins and Growth Temperature of Bacteria. International Journal of Molecular Sciences, 25(4), 2000. doi: 10.3390/ijms25042000.

Author contributions: A.K. designed the study, collected the data, performed computational analysis, and wrote the original draft of the manuscript. X.X. reviewed, edited the manuscript, supervised, and coordinated the study. All authors have read and agreed to the published version of the manuscript.

4.1 Abstract

The dynamic structures and varying functions of intrinsically disordered proteins (IDPs) have made them fascinating subjects in molecular biology. Investigating IDP abundance in different bacterial species is crucial for understanding adaptive strategies in diverse environments. Notably, thermophilic bacteria have lower IDP abundance than mesophiles, and a negative correlation with optimal growth temperature (OGT) has been observed. However, the factors driving these trends are yet to be fully understood. We examined the types of IDPs present in both mesophiles and thermophiles alongside those unique to just mesophiles. The shared group of IDPs exhibits similar disorder levels in the two groups of species, suggesting that certain IDPs unique to mesophiles may contribute to the observed decrease in IDP abundance as OGT

increases. Subsequently, we used quasi-independent contrasts to explore the relationship between OGT and IDP abundance evolution. Interestingly, we found no significant relationship between OGT and IDP abundance contrasts, suggesting that the evolution of lower IDP abundance in thermophiles may not be solely linked to OGT. This study provides a foundation for future research into the intricate relationship between IDP evolution and environmental adaptation. Our findings support further research on the adaptive significance of intrinsic disorder in bacterial species.

4.2 Introduction

4.2.1 Intrinsically Disordered Proteins and Their Abundance

For a long time, a defined protein structure was thought to be essential for protein functionality; however, this notion has been challenged as the concept of intrinsically disordered proteins (IDPs) has been established (Vladimir N. Uversky 2011; Dunker et al. 2002). IDPs, sometimes referred to as inherently unstructured proteins or nonfolding proteins, are proteins that lack a stable tertiary structure. Renowned for their flexibility, IDPs can adopt diverse conformations, setting them apart from structured proteins. This structural dynamism allows IDPs to engage in a wide range of biochemical functions, underscoring their versatility in cellular regulation (Wright and Dyson 2015), signaling cascades, and intricate molecular interactions. Moreover, the structural disorder has been associated with an increase in both the number and variety of functions based on Swiss-Prot function tags (Xie et al. 2007).

The tendency towards intrinsic disorder in proteins can be predicted using protein amino acid (AA) composition. Disordered proteins have a higher proportion of hydrophilic and uncompensated positively or negatively charged AAs than ordered ones; therefore, physiochemical properties such as absolute mean charge and mean hydrophobicity can be used to classify proteins as ordered or disordered [3,4]. The charge-hydrophobicity phase space could be plotted, and such plots have been proven to be reliable predictors of protein disorder (V. N. Uversky, Gillespie, and Fink 2000). This concept is the core of the advanced computational tools predicting IDP/IDR, such as PONDR (Xue et al. 2010), IUPred (Erdős, Pajkos, and Dosztányi 2021), fIDPnn (G. Hu et al. 2021), and many more. These tools use protein AA composition and often incorporate a window-based analysis to predict intrinsic disorder in proteins. Specifically, disordered proteins tend to exhibit a higher proportion of hydrophilic and uncompensated

positively or negatively charged AAs compared to ordered ones. The sliding window of AA along the sequence allows the assessment of local patterns and variations in physiochemical properties.

IDPs are widespread across all life domains (Peng et al. 2015) including viruses (Anjum et al. 2022). Their abundance is influenced by various factors, including, for example, organism complexity, with larger genomes generally displaying higher levels of IDPs (DeForte and Uversky 2017). Eukaryotes generally show both a higher frequency (Apic, Gough, and Teichmann 2001a; Ekman, Björklund, and Elofsson 2007) and longer lengths of IDPs (Apic, Gough, and Teichmann 2001b; Ekman et al. 2005; Liu and Rost 2004) compared to prokaryotes. Notably, within prokaryotes, IDP abundance is influenced by optimal growth temperatures (OGT) being significantly larger in mesophiles than in extremophiles adapted to higher temperatures (Burra, Kalmar, and Tompa 2010; Pancsa, Kovacs, and Tompa 2019). These results challenge the conventional understanding of the advantageous role of IDPs in extreme conditions, as they play an important role in detecting changes in the environment (S. E. Bondos, Dunker, and Uversky 2022).

Studying IDP abundance is essential for understanding cellular functioning, regulatory systems, and adaptive responses to the environment, as well as providing insights into their evolution across proteomes. IDPs play diverse roles in many cellular processes, and understanding factors influencing IDP abundance provides a key to unraveling the dynamic and flexible nature of these proteins, shedding light on their functional significance in cellular systems. While existing findings do describe a general pattern, specific causal elements underlying the association between OGT and IDP abundance remain unknown, offering an intriguing knowledge gap. The question at hand is determining if mesophiles have a larger number of IDPs or instead possess analogous proteins at greater disorder levels. In addition, the apparent connection between OGT and IDP abundance could be a phylogenetic consequence rather than a direct result of the OGT effect on IDP abundance. Our research strives to go beyond existing boundaries to answer these questions. To accomplish this, we assessed differences in abundance of IDP genes between the IDP groups. Additionally, we conducted a quasi-independent contrast calculation to gauge the impact of phylogeny on the relationship between OGT and IDP abundance.

4.2.2 Identification of IDP Groups

IDPs can be classified in many ways, including based on their molecular functions, functional features, sequence conservation, expression patterns, and biophysical properties (R. van der Lee et al. 2014). As an extension of using AA physiochemical properties to calculate protein absolute mean charge and mean hydrophobicity, more complex parameters, such as the fraction of charged residues (FCR) and net charge per residue (NCPR), can be used to separate IDPs into strong polyelectrolytes, strong polyampholytes, boundary, and weak IDPs, as represented on Figure 4.1 (Das and Pappu 2013):

Weak polyampholytes/polyelectrolytes (region 1): contain a small number of both positively and negatively charged AAs, as well as an approximately neutral overall charge. These proteins are often globules and tadpoles.

Boundary proteins (region 2, also known as Janus sequences): proteins that resemble both region 1 and region 3 properties. Specific properties of these proteins are largely context-dependent.

Strong polyampholytes (region 3): contain a significant number of both positively and negatively charged AAs, as well as an approximately neutral overall charge. These proteins are often flexible and form distinctly nonglobular coil-like, hairpin-like, or chimeric conformations.

Negative strong polyelectrolytes (region 4) and positive strong polyelectrolytes (region 5): contain a large number of either positively or negatively charged AAs, which results in either a strongly positive or a strongly negative overall charge. These proteins are often very flexible and form swollen coil-like conformations.

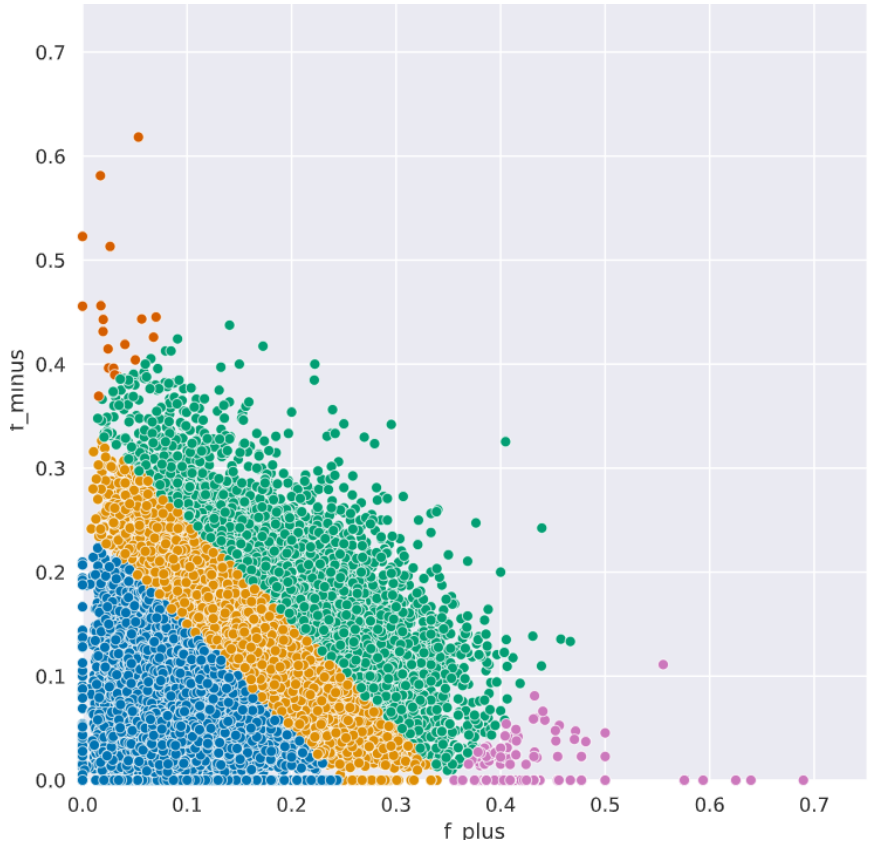

Figure 4.1. Scatter plot of proportions of positively-charged AA (f plus) and negatively-charged AA (f minus), representing the five IDP regions. Blue color shows region 1, orange shows region 2, green is for region 3, red is for region 4 and purple is for region 5.

This type of classification also allows the separation of globules from swollen coils (Mao et al. 2010). In addition to the above classification, we grouped proteins based on their AA similarity and reported molecular functions where they were available. Finally, we identified clusters of similar proteins across thermophiles and mesophiles to detect any potential differences in disorder levels between the two species groups. We visualized aligned disorder values for the most divergent clusters in order to assess whether there are any patterns leading to that divergence. Adapted from (Das and Pappu 2013).

4.2.3 Quasi-Independent Contrasts

Analysis of quasi-independent contrasts is an important method that helps us to unravel the relationships between OGT, IDP abundance, and phylogenetic relationships between the bacterial species used in the analysis. By employing quasi-independent contrasts, we can control

for shared ancestry among species, ensuring a more accurate assessment of the direct impact of OGT on IDP abundance. The need for phylogeny-based comparative methods becomes evident when examining relationships between genes, phenotypes, and environmental factors among related species. Traditional statistical methods may be inadequate for quantifying these relationships due to the inherent co-ancestry among data points.

Independent and quasi-independent contrast comparison offers a more sophisticated means of addressing this challenge. As described by Xia (Xia 2020), the method involves the minimization of the residual sum of squares by inferring ancestral states, accounting for phylogenetic influences through weighting factors. By applying quasi-independent contrasts, we can assess the relationship between OGT and IDP abundance while accounting for phylogenetic relationships, thus providing a more precise evaluation of how OGT directly influences IDP abundance. The contrasts between the two variables can be fitted into a linear model with an intercept fixed at the origin, and that model can then be interpreted to provide additional insights.

4.3 Materials and Methods

4.3.1 Data Sources and Availability

We used UniProt (The UniProt Consortium 2021) as a source for the AA sequence data and TEMPURA (Sato et al. 2020) for bacteria OGT data. TEMPURA contains bacteria and archaea OGT with ribosomal 16s gene used for reference. We downloaded the database entirely and then filtered it to only contain data for bacteria with recorded 16s accession numbers.

For each of the remaining species, we searched for a reference proteome on UniProt and downloaded them if they were available and had at least 1,000 proteins. As a result, our dataset consisted of 1,132,382 proteins from 304 species.

4.3.2 Protein Clustering

All proteins in our dataset have been clustered using CD-HIT (Fu et al. 2012) with a setting for a minimal global similarity score of 0.7. Clusters have been filtered to follow these conditions:

- Proteins from at least 10 different species per cluster;
- At least one candidate IDP (identification described in the disorder calculation subsection);

This way, we generated 616 clusters of interest. Additionally, we obtained UniProt molecular function tags for each protein from these clusters.

4.3.3 Disorder Calculations

Two rounds of disorder prediction have been performed:

First, we calculated disorders for each protein of the dataset using RAPID (Yan et al. 2013). While this model has the disadvantage of only predicting a single overall disorder metric for a given protein, it is in, fact, rapid and, given the large number of proteins in our dataset, is a suitable model for initial filtering. Additionally, we calculated the FCR and NCPR for each protein to group them into five classes, as described by Das and Pappu (Das and Pappu 2013).

Based on the RAPID results and FCR/NCPR, we identified IDP candidates as those that satisfy at least one of the following conditions:

- RAPID disorder score ≥ 0.5;
- Total number of residues $\times$ RAPID disorder score ≥ 100;
- IDP type 3, 4, or 5 (strong polyampholytes or positive/negative strong polyelectrolytes);

This way, we significantly narrowed down the list of proteins for further analysis as well as defined a binary ordered/disordered separation for our dataset.

The second round of disorder calculation was performed on clusters of interest (see the relevant section about clustering for more information) using fIDPnn (G. Hu et al. 2021), a more advanced but also much more time-consuming method than RAPID. This model has shown to be a very effective one (Kurgan et al. 2023; Necci, Piovesan, and Tosatto 2021), as well as able to output a disorder score for each residue of the protein, which allowed us to compare IDPs at the residue level. By using the two-round approach, we were able to evaluate entire proteomes using a faster model, identify potential IDPs, and then obtain more detailed results for these candidates using a more complex but slower model.

4.3.4 Cluster Disorder Alignment

The AA sequences of each protein in a cluster were aligned using ClustalW (Thompson, Higgins, and Gibson 1994) with a gap insertion penalty of 1 and a gap extension penalty of 0.5.

The disorder scores were smoothed for the plots with a moving average. The sliding window was equal to protein length divided by 30.

4.3.5 Quasi-Independent Contrast Calculation

We referred to Xia's least squares method of quasi-independent contrast calculation (Xia 2020). First, we obtained 16s nucleotide sequences based on TEMPURA accession numbers for all bacteria, with an addition of one sequence from archaea species, NG_046384.1 of *Pyrobaculum ferrireducens*, which was used as an outgroup. Four of the sequences were removed using DAMBE (Xia 2018), as they were for entire genomes rather than 16 s. We aligned the resulting sequences using MAFFT (Katoh and Standley 2013) with default parameters except for specification for nucleotide sequences and calculated the distances based on the aligned sequence identity. The distances then have been used to build a phylogenetic tree using UPGMA and NG_046384.1 as outgroups.

For each leaf of the resulting tree, which represented one of the species from our dataset, we computed overall fraction of disorder (FOD), as predicted by RAPID and OGT, as recorded in TEMPURA. For each internal node, average FOD and OGT were used as initial guesses and were later optimized using RSS minimization. Finally, contrasts have been calculated between each offspring pair sharing the same ancestor.

4.4 Results and Discussion

4.4.1 Overall IDP Abundance in Different Proteomes

We identified a weak negative relationship between OGT and overall fraction of disorder (FOD) predicted by RAPID (Figure 4.2). This result partially supports previous findings that thermophiles should have a lower IDP abundance. The relationship is significant, but the effect size seems to be very small ($R^2 = 0.016$, slope = −0.0003, and p-value = 0.030). Additionally, when separating species into thermophilic (OGT of at least 40 °C) and mesophilic (all other species) and comparing their FOD as predicted by RAPID (Figure 4.3) using a two-sided t-test, a significant difference was observed with mesophiles having more disorder (p-value = 6.898×10^{-43}). However, the effect size is very small: the thermophilic average FOD = 0.1301 ± 0.0004 and the mesophilic average FOD = 0.1364 ± 0.0001. The high significance is very likely the result of the large sample size in this case and not the strength of the relationship.

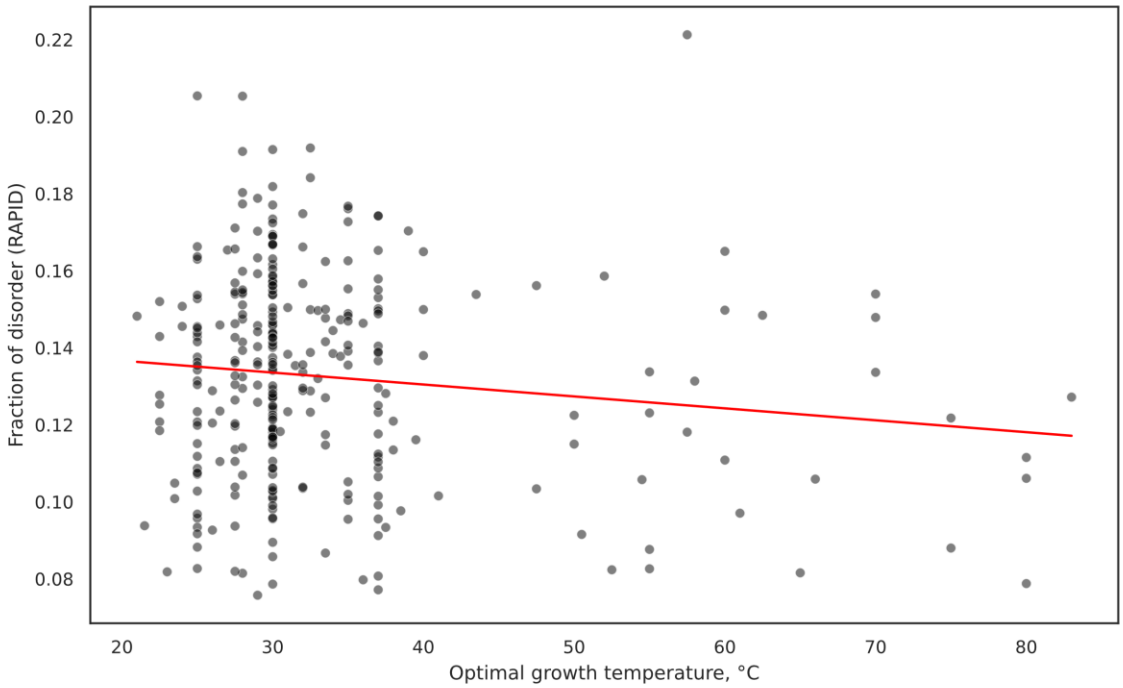

Figure 4.2. Scatter plot of OGT and FOD, with line showing Ordinary Least Squares (OLS) model. High OGT is associated with lower FOD. Effect size seems to be small but statistically significant ($R^2 = 0.016$, slope = −0.0003, and p-value = 0.030) for the linear model. Majority of organisms are mesophiles, which could potentially skew the results of modeling. Figure adapted from (Kruglikov and Xia 2024).

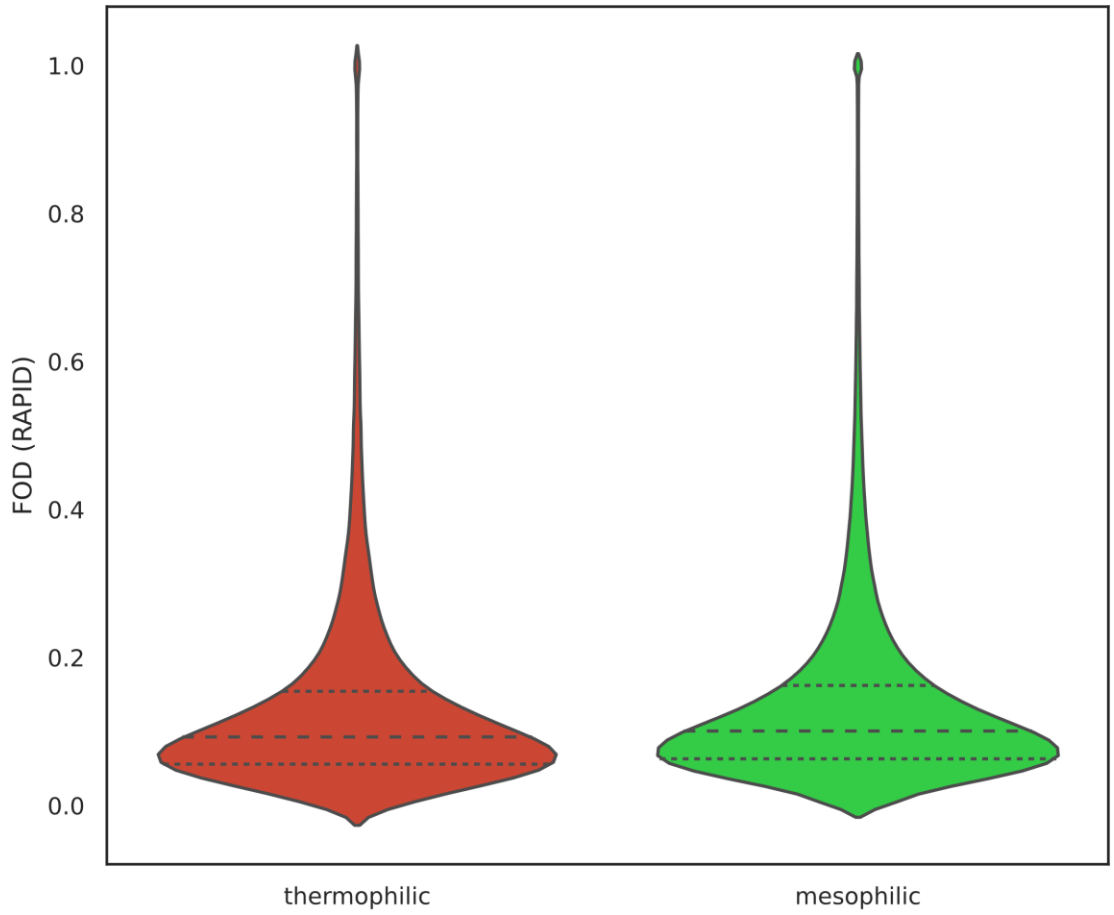

Figure 4.3. FOD distributions for thermophilic and mesophilic proteins. The two distributions seem to be very similar even though a statistically significant difference has been observed between the mean values (*t*-test *p*-value = 6.898×10^{-43}). This significance is likely to be the result of the large sample size. Thermophilic average FOD = 0.1301 ± 0.0004 and mesophilic average FOD = 0.1364 ± 0.0001. All proteins from the dataset have been assessed, and FOD has been predicted using RAPID. Figure adapted from (Kruglikov and Xia 2024).

The effect of OGT on FOD can also be seen in the decreased variation in FOD as ODT increases (Figure 4.2). With low OGT, FOD can be low or high. However, high OGT might seem to be selected against high FOD, pushing the variation in FOD to a lower range.

4.4.2 Overall IDP Abundance in Orthologs

For each cluster identified using CD-HIT as described in Materials and Methods, we recalculated disorder predictions using fIDPnn, a more accurate but also a much slower model. FOD

calculations were found to be highly correlated between RAPID and fIDPnn, so the use of RAPID as a fast initial filter model has been justified (Figure 4.4).

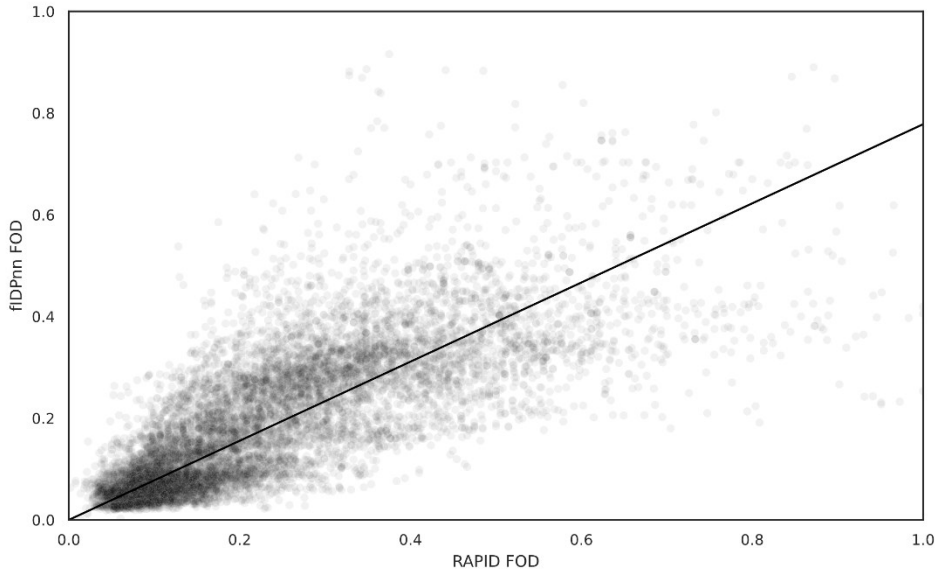

Figure 4.4. Scatter plot of FOD contrast / OGT contrast relationship. Figure adapted from (Kruglikov and Xia 2024).

The already weak negative relationship between OGT and FOD that we observed for the overall dataset (Figure 4.2) has not been seen for cluster data when using all clustered proteins as a subset. Surprisingly, an unexpected positive relationship emerges for both RAPID and fIDPnn-predicted FOD (Figure 4.5). This intriguing finding challenges the notion that orthologous proteins shared between thermophiles and mesophiles have greater disorder levels in mesophiles. Comparison of mean FOD values across the two datasets also produced opposite results from those of overall proteomes (Figure 4.6).

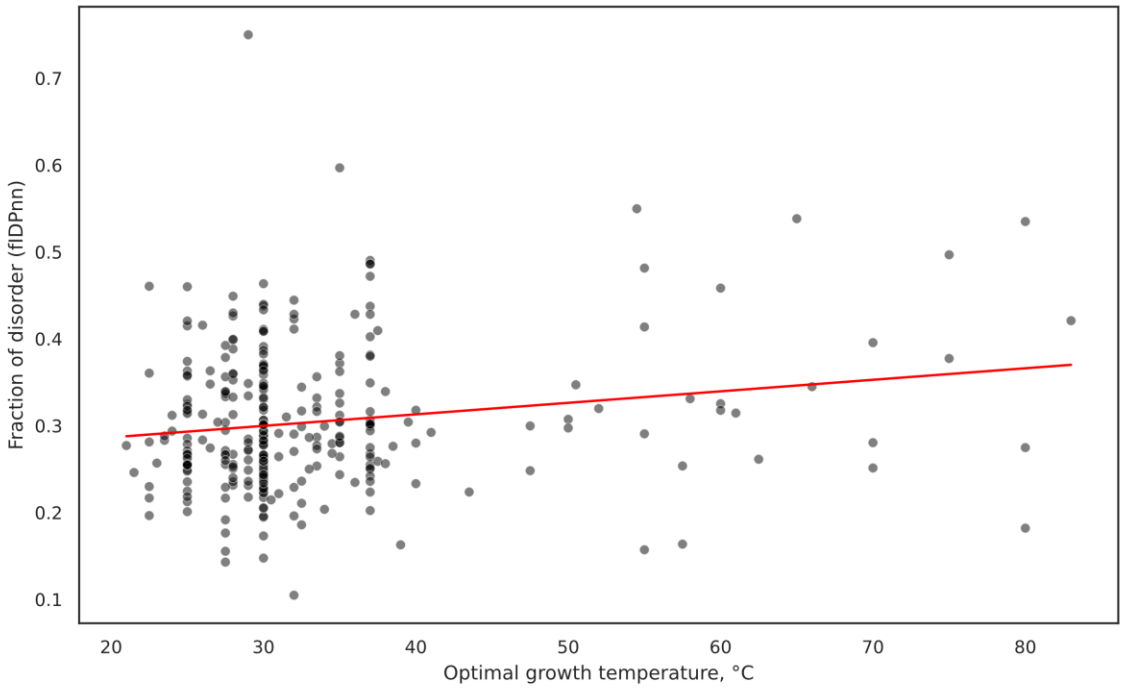

Figure 4.5. Scatter plot of OGT and FOD, with line showing Ordinary Least Squares (OLS) model. Positive relationship between OGT and FOD is observed for clustered data. R^2 = 0.052, slope = 0.0017, and *p*-value = 5.69×10^{-5} for the linear model. FOD has been calculated using flDPnn. Linear model with RAPID FOD showed similar results (R^2 = 0.031, slope = 0.0013, and *p*-value = 0.002). The positive relationship is opposite to the one for overall data (Figure 4.2). Figure adapted from (Kruglikov and Xia 2024).

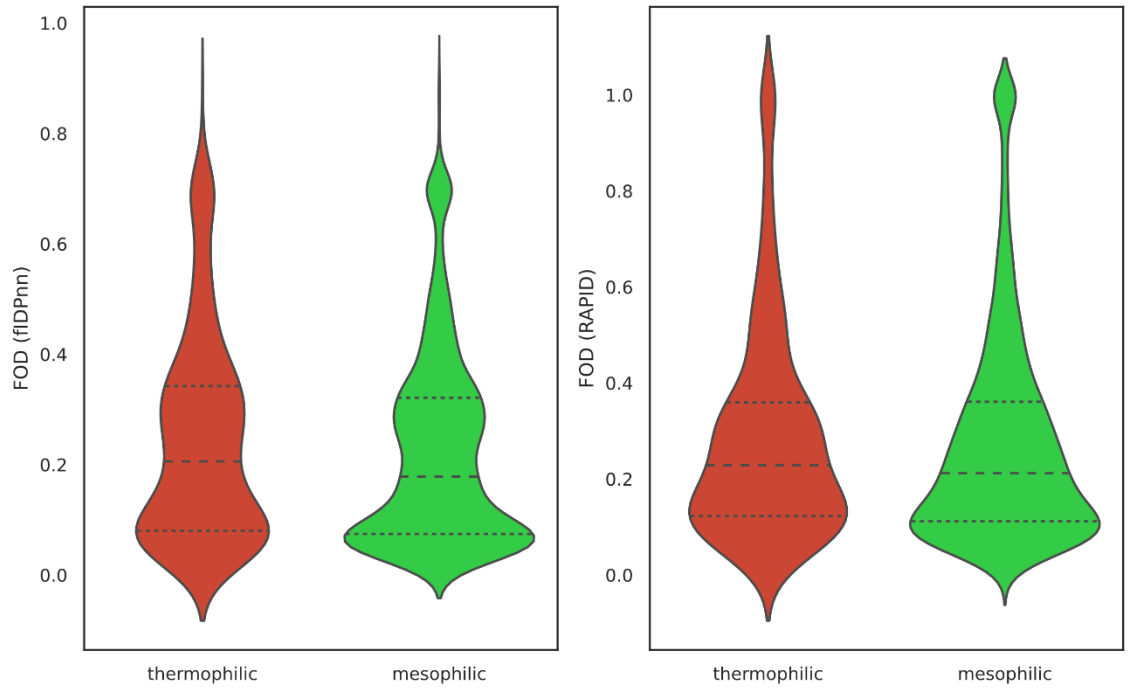

Figure 4.6. FOD distributions for thermophilic and mesophilic orthologs. Left violin plot shows distributions of FOD calculated by fIDPnn, and right violin plot shows distributions of FOD calculated by RAPID. The pairs of distributions seem to be very similar even though a statistically significant difference has been observed between the mean values of fIDPnn FOD (*t*-test *p*-value = 0.0025 for fIDPnn and 0.167 for RAPID). Using fIDPnn, thermophilic average FOD = 0.2425 ± 0.007 and mesophilic average FOD = 0.2232 ± 0.002. Using RAPID, thermophilic average FOD = 0.2887 ± 0.009 and mesophilic average FOD = 0.2760 ± 0.003. Interestingly, the differences are in opposite directions from the overall data (Figure 4.3). Figure adapted from (Kruglikov and Xia 2024).

4.4.3 Abundance in Different IDP Classes and Proteins with Different Molecular Functions

Utilizing the classification based on the fraction of charged residues (FCR) and net charge per residue (NCPR) of amino acid sequences, clustered orthologs were categorized into five distinct classes: weak polyampholytes/polyelectrolytes, boundary proteins, strong polyampholytes, negative strong polyelectrolytes, and positive strong polyelectrolytes. Interestingly, none of the orthologs were classified as negative strong polyelectrolytes, while all other classes were

represented by some clustered proteins. The corresponding fraction of disorder (FOD) values for each class, as predicted by the fIDPnn model, are presented in Table 4.1.

Table 4.1. FOD for different classes of ortholog IDPs in mesophilic and thermophilic bacteria.

IDP Class	Thermophilic FOD	Mesophilic FOD
Weak polyampholytes/polyelectrolytes	0.303 ± 0.054; *n* = 21	0.192 ± 0.007; *n* = 609
Boundary proteins	0.181 ± 0.007; *n* = 422	0.176 ± 0.002; *n* = 7830
Strong polyampholytes	0.358 ± 0.014; *n* = 180	0.357 ± 0.004; *n* = 2145
Negative strong polyelectrolytes	-	-
Positive strong polyelectrolytes	0.682 ± 0.016; *n* = 13	0.706 ± 0.005; *n* = 256

For each class, the table provides the mean FOD value along with its standard error, denoted as ±, and the sample size (n) representing the number of proteins within each category. Comparable levels of disorder between thermophiles and mesophiles were found for boundary proteins and strong polyampholytes, but some differences could be observed between the two species groups for weak polyampholytes/polyelectrolytes and positive strong polyelectrolytes. Thermophilic weak polyampholytes/polyelectrolytes had more disorder than mesophilic ones (FOD of 0.303 for thermophiles and 0.192 for mesophiles, *t*-test *p*-value = 0.004). Conversely, positive string polyelectrolytes were found to be more disordered in mesophiles than in thermophiles, although this effect was not found to be statistically significant (FOD of 0.682 for thermophiles and 0.706 for mesophiles, *t*-test *p*-value = 0.320). This finding may be a possible explanation for the negative correlation observed in Figure 4.2 and could be explained by the higher compactness of IDPs in higher temperatures (Thole, Waudby, and Pielak 2023).

Similarly to the above, we calculated the average FOD for identified molecular functions, as tagged on UniProt (Table 4.2). We found that IDP orthologs tagged as activator and nuclease were unique to only mesophiles, although they were not found at large levels there either–only

33 activators and 18 nucleases. At the same time, activator proteins had a relatively high FOD of 0.339, as predicted by fIDPnn. Moreover, we did not observe differences in FOD levels across any of the molecular functions that had been identified for orthologs in both mesophiles and thermophiles, especially among the more disordered ones.

Table 4.2. FOD for different function tags of ortholog IDPs in mesophilic and thermophilic bacteria.

IDP Function Tag	Thermophilic FOD	Mesophilic FOD
Activator	-	0.339 ± 0.010; n = 33
Nuclease	-	0.110 ± 0.003; n = 18
Chaperone	0.090 ± 0.010; n = 18	0.114 ± 0.003; n = 516
DNA-binding	0.276 ± 0.032; n = 35	0.264 ± 0.008; n = 477
Elongation factor	0.125 ± 0.034; n = 33	0.070 ± 0.005; n = 367
Excision nuclease	0.053 ± 0.005; n = 7	0.054 ± 0.001; n = 192
Hydrolase	0.064 ± 0.005; n = 14	0.108 ± 0.009; n = 209
Initiation factor	0.217 ± 0.026; n = 3	0.209 ± 0.004; n = 56
Isomerase	0.090 ± 0.024; n = 3	0.076 ± 0.001; n = 91
Ligase	0.060 ± 0.008; n = 12	0.058 ± 0.002; n = 336
Lyase	0.077 ± 0.005; n = 11	0.076 ± 0.001; n = 230
Multifunctional enzyme	0.055 ± 0.000; n = 1	0.052 ± 0.001; n = 13
Oxidoreductase	0.102 ± 0.018; n = 13	0.073 ± 0.002; n = 245
Peroxidase	0.082 ± 0.007; n = 3	0.097 ± 0.004; n = 59
Protease	0.082 ± 0.008; n = 7	0.080 ± 0.001; n = 284
RNA-binding	0.165 ± 0.033; n = 17	0.126 ± 0.005; n = 388
Receptor	0.095 ± 0.000; n = 1	0.112 ± 0.016; n = 10
Repressor	0.331 ± 0.086; n = 4	0.263 ± 0.017; n = 51
Ribosomal protein	0.467 ± 0.018; n = 110	0.437 ± 0.004; n = 1859
Rotamase	0.191 ± 0.030; n = 4	0.209 ± 0.010; n = 36
Serine protease	0.062 ± 0.000; n = 1	0.063 ± 0.001; n = 49
Sigma factor	0.142 ± 0.011; n = 18	0.149 ± 0.003; n = 154
Topoisomerase	0.075 ± 0.007; n = 6	0.087 ± 0.001; n = 151
Transferase	0.087 ± 0.012; n = 27	0.073 ± 0.003; n = 713
Translocase	0.049 ± 0.004; n = 7	0.076 ± 0.004; n = 127
rRNA-binding	0.294 ± 0.010; n = 96	0.284 ± 0.002; n = 1700
tRNA-binding	0.338 ± 0.015; n = 55	0.297 ± 0.005; n = 686

The abovementioned results indicate that thermophiles are more likely to lack some IDPs that are present in mesophiles than to have less disordered orthologs in most cases. At the same time, a slight increase in the average FOD has been seen for mesophilic coil-like proteins, which are often involved in signaling through binding to various partners (Shao et al. 2021; Kolonko et al. 2020). On the other hand, weak polyampholytes and polyelectrolytes might be more disordered in thermophiles because these IDPs may be more involved with adaptations to high temperatures. Interestingly, we were able to find examples of disorder differences between thermophilic and mesophilic weak polyampholytes/polyelectrolytes going in both directions. Large ribosomal subunit protein uL11 orthologs (UniProt IDs A0A7V5PNC3, A0A291PC16, A0A1B4VGG8, A0A250KZM7, and A0A5C1EBZ5, among others) were generally more disordered in thermophiles. Conversely, small acid-soluble spore protein sspB orthologs (UniProt IDs A0A0D8BNT0, A0A0D8BRQ7, A0A2K9J164, A0A0U4FDZ6, and A0A221MG13, among others) were found to be more disordered in mesophiles.

Among the IDPs from clusters that turned out to be unique to mesophiles, the majority were tagged as either ribosomal or rRNA-binding proteins (714 and 607 IDs out of 3469 proteins with tagged molecular functions). Additionally, some were tagged as transferase (278), chaperone (220), tRNA-binding (196), and DNA-binding (190). Apart from transferases and, partially, chaperones, all these groups are generally short proteins with significant disorder levels. We can also see that proteins with the same molecular functions are abundant in thermophiles, and it is possible that the large number of clusters being unique to mesophiles is due to the large number of variants of these proteins in general across all the species.

4.4.4 Analysis of Aligned Ortholog Clusters

Given the results of cluster analysis from the previous section, it seems that the relationship between OGT and IDP abundance is a complex one, and a look into the nature of the aligned clusters may reveal some patterns between IDP AAs and levels of disorder. We identified 10 ortholog clusters with the largest absolute differences between mesophilic and thermophilic FOD (Table 4.3). Among these, six had thermophilic IDPs that were more disordered than their mesophilic orthologs, and four were more disordered in mesophiles. The larger FOD within each cluster is underlined. The majority of these proteins turned out to be ribosomal proteins, although

we also identified a rubredoxin, an acyl carrier, a spore protein, and a cupin protein among the clusters.

Table 4.3. Most divergent FOD between thermophilic and mesophilic orthologs.

Protein Name	FOD (Thermophilic)	FOD (Mesophilic)	Absolute FOD Difference	Cluster Members
Rubredoxin	0.085	0.389	0.304	A0A291P6P4, A0A410H536, A0A1B2LXP3…
Acyl carrier	0.534	0.392	0.142	A0A291P5S3, A0A410H1W4, A0A386X534…
Spore protein	0.647	0.786	0.139	A0A0D8BNT0, A0A0D8BRQ7, M5R4X2…
LRSP * bL19	0.357	0.239	0.118	A0A1U9K6D3, A0A1B9NF78, A0A1B0ZK26…
SRSP bS21	0.596	0.484	0.112	A0A0D5YVA4, A0A1Z4BT12, A0A1L3J4J5…
SRSP uS14	0.476	0.370	0.106	A0A0P0DDQ1, A0A0D5YRD0, A0A0S2I2L9…
LRSP bL28	0.512	0.610	0.098	A0A0D8BU85, M5QWZ7, A0A1D7QW46…
Cupin	0.385	0.287	0.097	A0A0K2SHK7, A0A0D5NPB9, A0A4P6K4Z4…
LRSP uL24	0.323	0.414	0.090	A0A0D8BQ30, M5QVZ2, A0A1D7QZW9…
SRSP uS14	0.501	0.417	0.083	A0A291PBX7, A0A7C9NQP7, A0A3T1DHB4…

* LRSP = large ribosomal subunit protein; SRSP = small ribosomal subunit protein.

The identified clusters have been aligned and visualized in order to investigate any potential patterns and regions that contribute most to the observed differences in disorder levels (Figures 4.7–4.16). Additionally, WebLogo (Crooks et al. 2004) diagrams have been created for these alignments for the assessment of AA consensus sequences. Hydrophobic and acidic AAs seem to be prevalent in regions where thermophilic IDP has a higher level of disorder, possibly indicating some temperature sensitivity of these residues. Conversely, polar AAs seem to be more frequent in IDPs that show larger disorder in mesophiles, although these AAs are generally

common in IDPs. Combined with the other results, these findings suggest that neither the functional background of IDPs nor their AA composition have simple relationships with the levels of disorder in mesophiles and thermophiles. Instead, the relationship is a highly complex one, and further research into these factors' contribution to IDP formation would be beneficial.

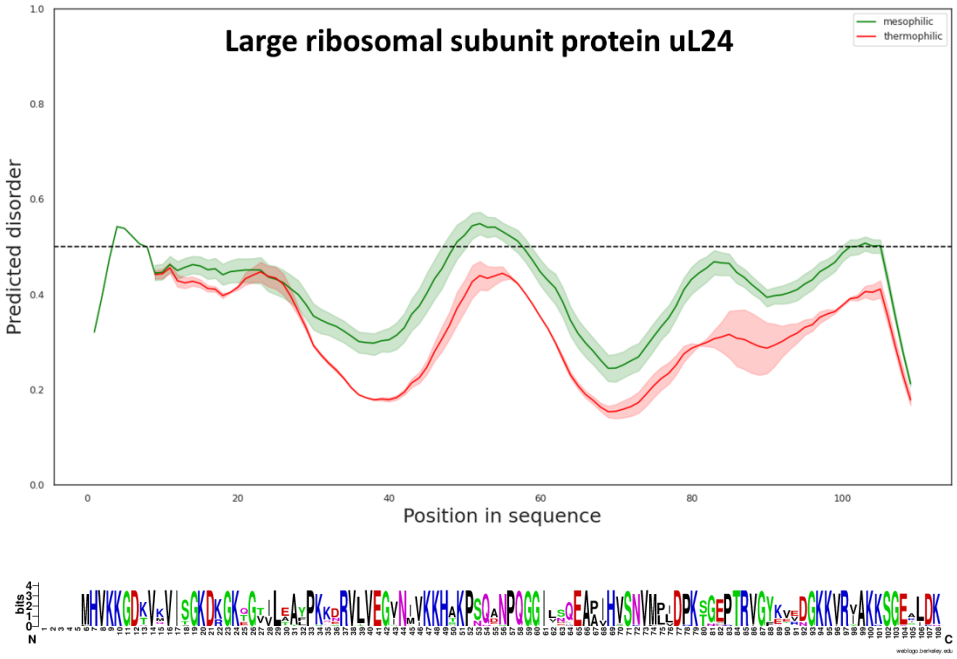

Figure 4.7. Aligned disorder (uL24) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

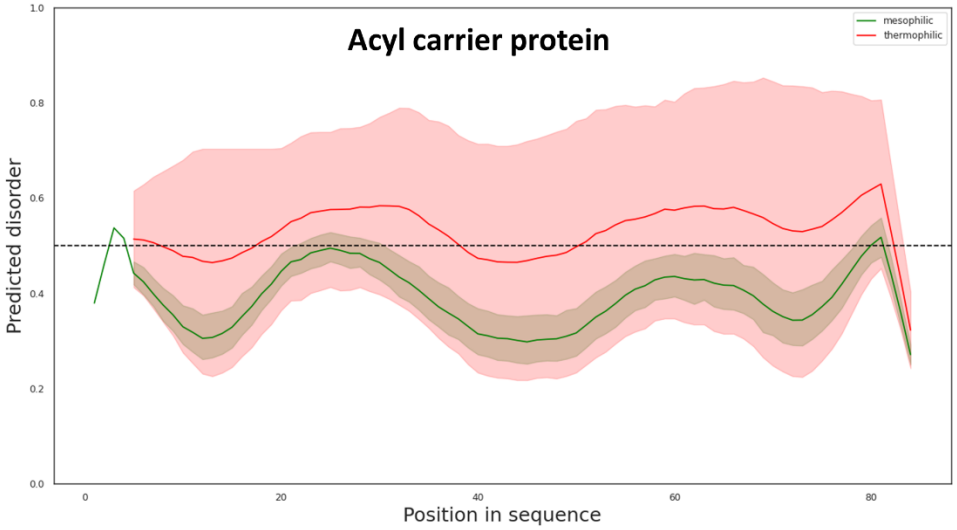

Figure 4.8. Aligned disorder (Acyl carrier protein) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

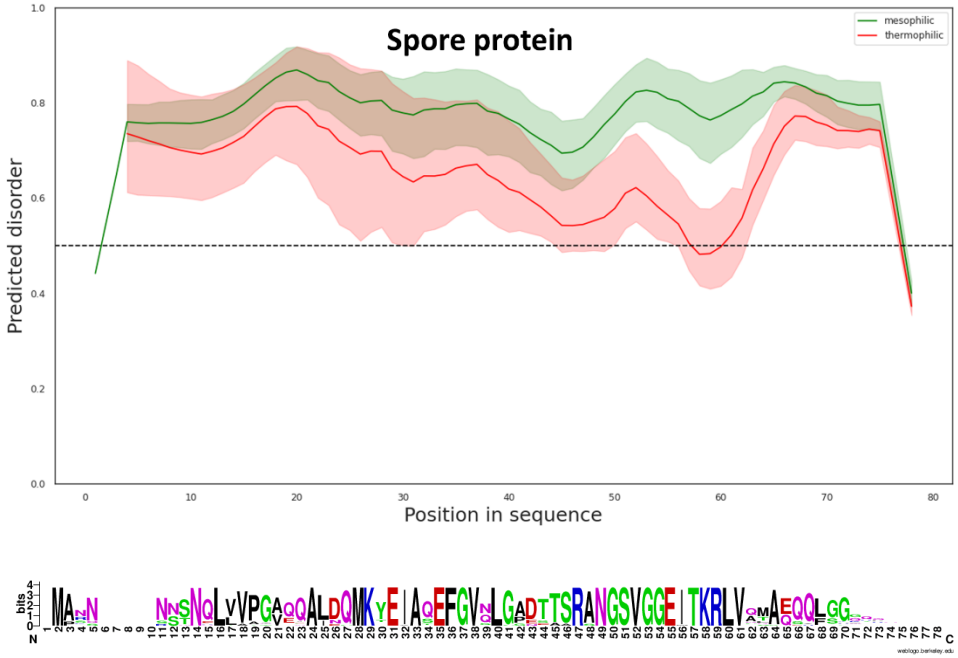

Figure 4.9. Aligned disorder (Spore protein) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

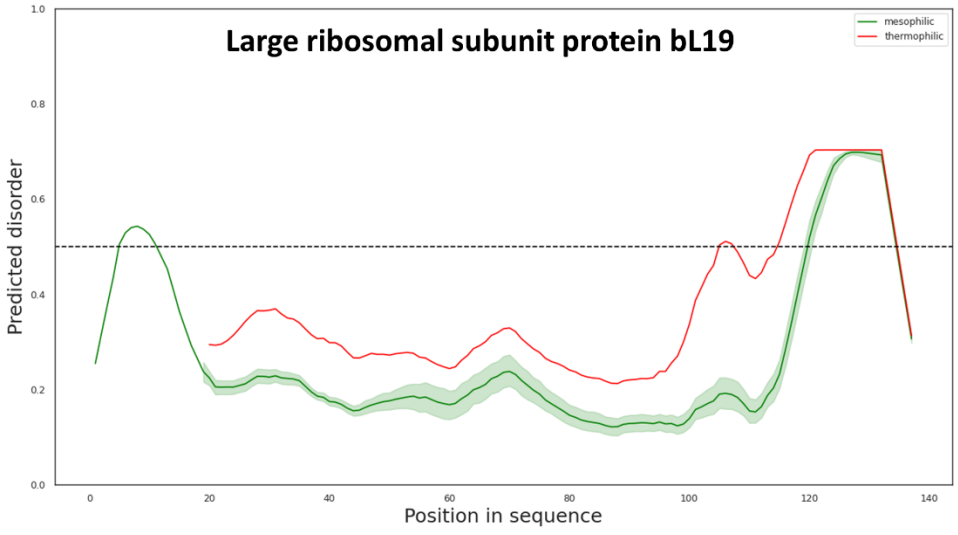

Figure 4.10. Aligned disorder (bL19) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

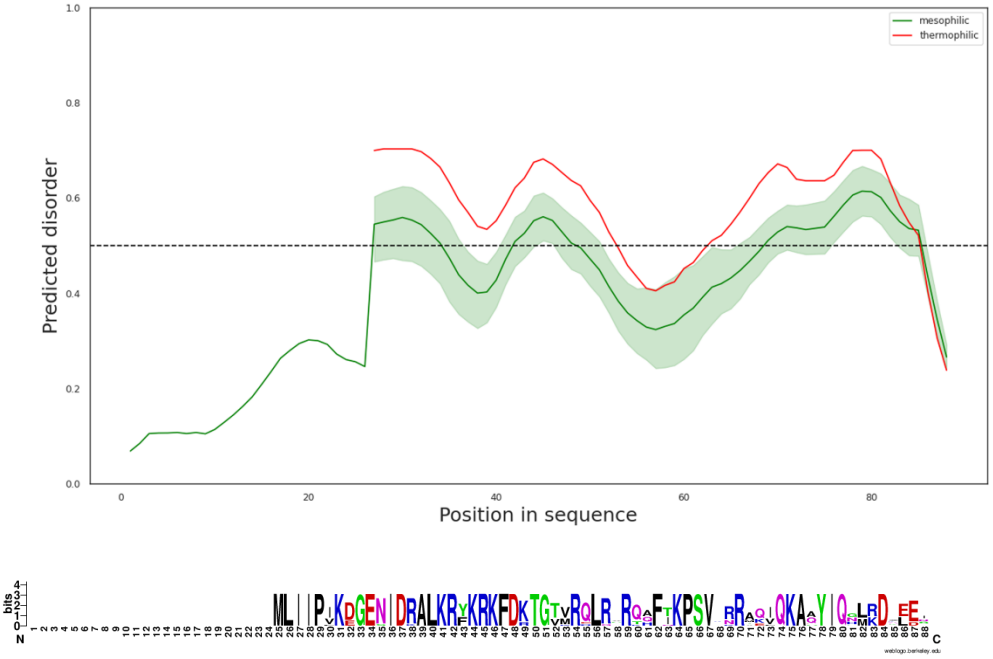

Figure 4.11. Aligned disorder (bS21) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

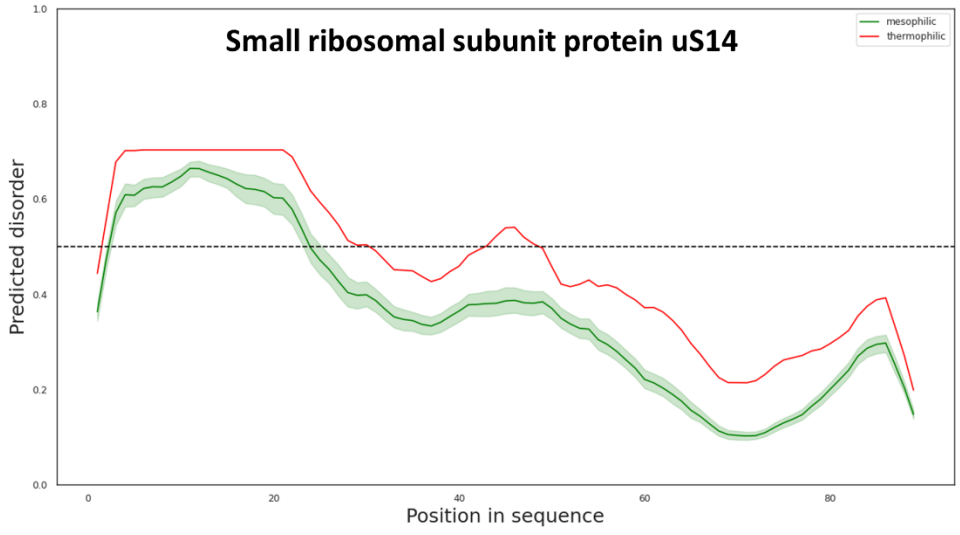

Figure 4.12. Aligned disorder (uS14) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

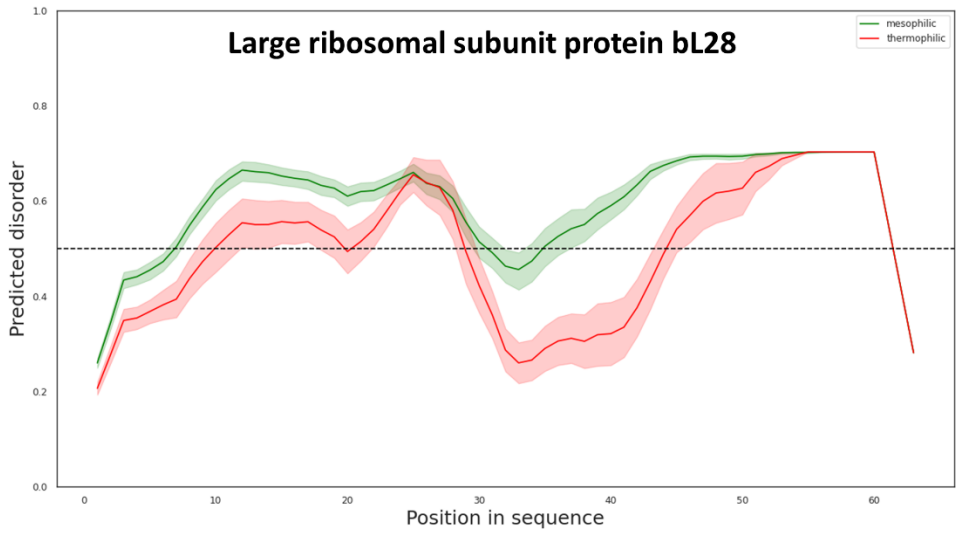

Figure 4. 13. Aligned disorder (bL28) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

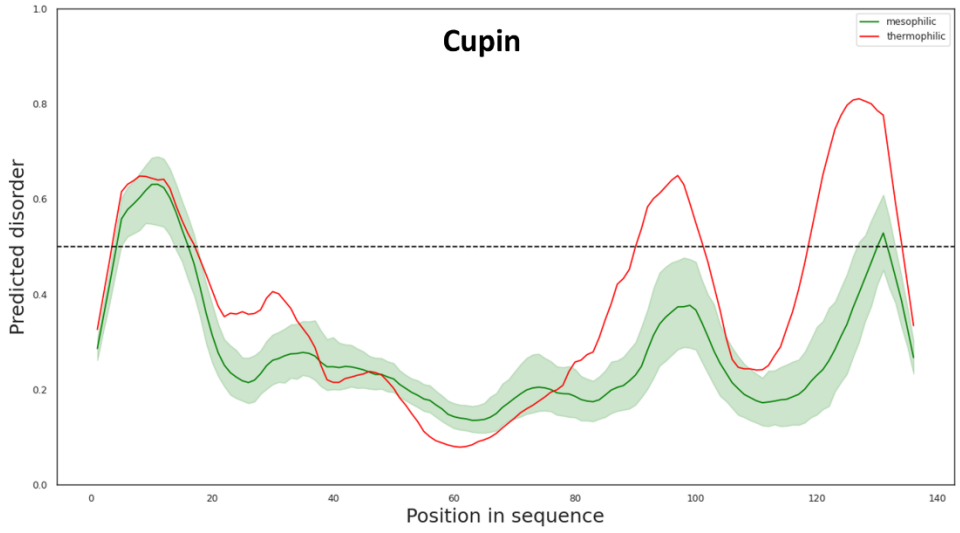

Figure 4.14. Aligned disorder (Cupin) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

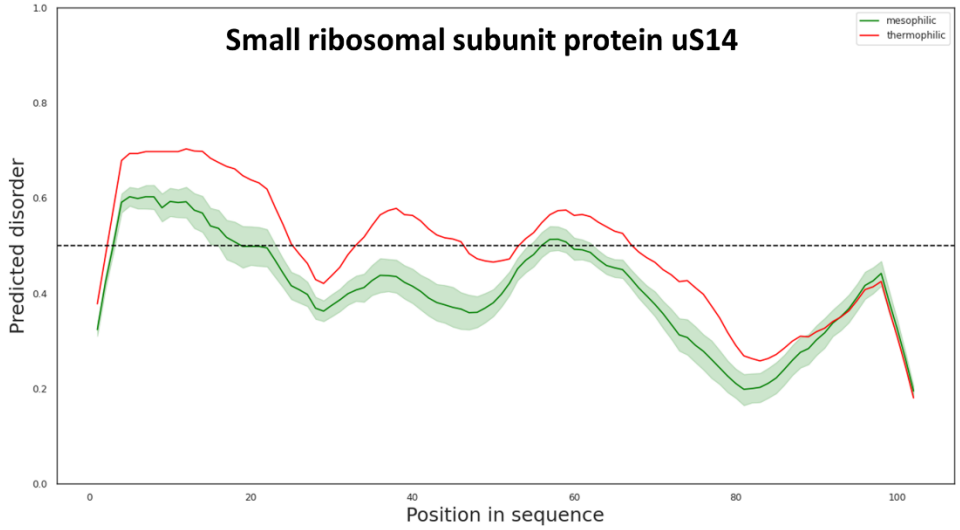

Figure 4.15. Aligned disorder (uS14) and corresponding WebLogo. Figure adapted from (Kruglikov and Xia 2024).

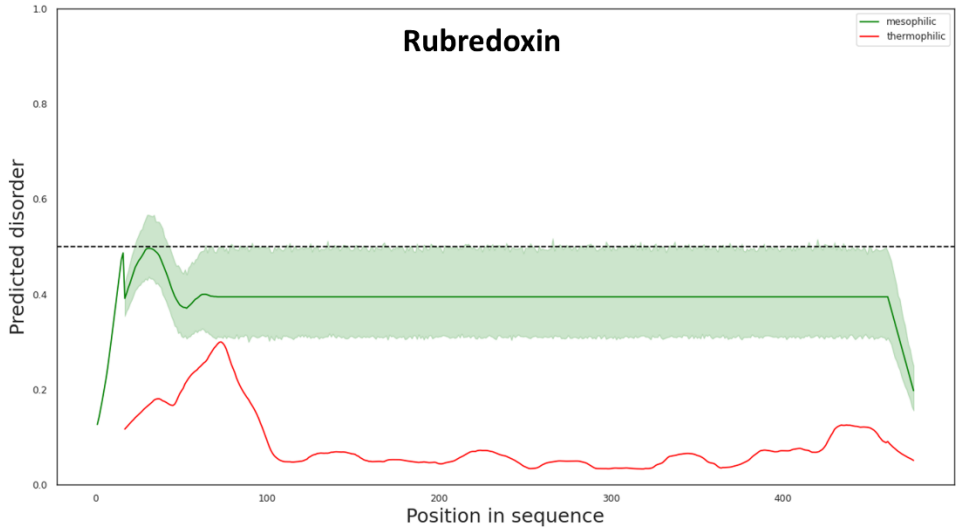

Figure 4.16. Aligned disorder (Rubredoxin) and corresponding WebLogo. The WebLogo has been limited to 100 first positions because of truncation. Figure adapted from (Kruglikov and Xia 2024).

4.4.5 Phylogeny Impact on FOD/OGT Relationship

In our study, we tried to assess whether the weak negative correlation between OGT and FOD, as observed in the overall proteome comparison, persists when accounting for phylogeny using quasi-independent contrasts. Surprisingly, our findings indicate that phylogeny exerts a more substantial influence on the relationship than OGT in bacterial species. Contrary to the initial observation in the overall dataset, the weak negative relationship has not been observed for contrast data (R2 = 0.002, slope = 8.552×10^{-5}, and *p*-value = 0.491). The scatter plot of the contrasts, illustrated in Figure 4.17, suggests an absence of any noticeable relationship, implying that OGT may not be a decisive factor influencing IDP abundance. Instead, it appears to be a characteristic carried along with the relative taxa, adding a nuanced layer to our understanding of factors affecting IDP abundance in bacteria.

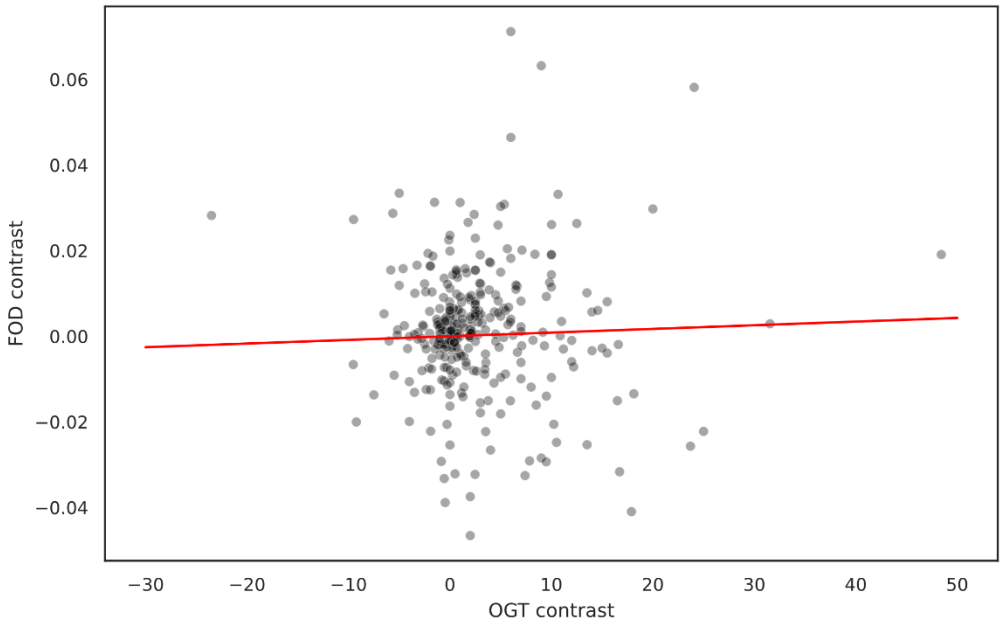

Figure 4. 17. Scatter plot of FOD contrast/OGT contrast relationship. Contrast controls for phylogeny impact, and we observe no relationship between IDP abundance and temperature differences; therefore, phylogeny seems to be a more important factor than OGT. Figure adapted from (Kruglikov and Xia 2024).

Several possible factors may be driving the observed phylogenetic effects. First, variables related to genomic characteristics may contribute to the observed differences in IDP abundance. For example, genome size and GC content are known to be correlated with IDP abundance (Pavlović-Lažetić et al. 2011), and other genomic features associated with phylogeny could also influence the evolution of IDPs independent of OGT. Second, the co-evolution of protein networks within specific phylogenetic groups may play a role in shaping IDP evolution. Interactions between proteins and their partners, influenced by shared evolutionary history, could contribute to the observed patterns in disorder abundance. Lastly, shared ancestry and evolutionary relationships may contribute to the observed patterns in IDP abundance, with closely related bacterial species inheriting similar traits, including features related to IDPs, such as amino acid composition, structural motifs, or functional roles in cellular processes. These traits may be unrelated to the OGT of the species but have an effect on IDP abundance.

These conclusions are particularly intriguing since they add context to previous findings showing thermophiles had lower disorder abundance than mesophiles. The lack of a clear functional justification for these findings suggests that factors other than OGT should be considered. Our findings call for a reconsideration of the relationship between IDP abundance and environmental conditions, emphasizing the importance of phylogeny and potentially other variables such as genome size. Future research should take into account this complex network of elements in order to improve the accuracy and comprehensiveness of investigations in unraveling the complexity of IDP evolution in bacterial species.

Chapter 5. Conclusion

5.1 Importance of Cellular Environment in Protein Folding

The results of this study underscore the critical role of cellular environment factors in influencing protein folding dynamics. Chapter 2 highlighted the challenges faced when thermophilic proteins from organisms like *Thermus thermophilus* are expressed in mesophilic systems like *Escherichia coli*. The misfolding observed in these scenarios can lead to alterations in protein structure and function, emphasizing the importance of utilizing appropriate expression systems to minimize such effects. In Chapter 3, the application of protein secondary structure algorithms in SARS-CoV-2 research showcased the significance of understanding protein structure beyond amino acid sequences. By analyzing predicted secondary structures of ACE2 proteins across different mammalian species, we gained insights into host-virus interactions and potential susceptibility variations. As cellular environment impacts protein folding even at secondary structure level, this approach demonstrated the value of secondary structure analysis in elucidating structural determinants relevant to species-specific differences in viral infections. Furthermore, in Chapter 4 we delved into the abundance of intrinsically disordered proteins (IDPs) in mesophiles and thermophiles, revealing intriguing patterns related to organism growth temperature (OGT). The results presented in this chapter highlight the complexity of the relationship between OGT and IDP abundance, showing that it is perhaps much more nuanced than previously reported. This complexity also exemplifies how cellular environment can be a factor not only to folding of globular proteins, but also of IDPs. Collectively, our findings emphasize the intricate relationship between cellular environment factors and protein folding dynamics across diverse biological contexts.

The presented data underscore the critical significance of incorporating cellular environments into protein structure studies and functional evaluations. Factors like chaperone proteins, pH levels, temperature, and ionic concentrations within the cellular milieu exert substantial influence on protein folding dynamics. These environmental factors can serve as valuable features in applications such as protein structure prediction. Notably, current state-of-the-art prediction models like AlphaFold and AlphaFold2, along with models for predicting intrinsic disorder, do not yet integrate cellular environment data into their predictions. By acknowledging and accommodating these environmental factors, researchers can enhance the accuracy of protein

structure and function assessments. This approach not only deepens our comprehension of biological processes but also holds the potential to unveil novel insights into protein behavior across diverse biological contexts.

5.2 Challenges of Data Availability in Studying Diverse Cellular Environments

During our research, we encountered challenges related to data limitations for diverse cellular environments, particularly evident in Chapter 2 where we investigated the impact of expression systems on protein folding dynamics. While *T. thermophilus* stands out as a well-studied extremophile species with abundant protein structures in the Protein Data Bank (PDB), the available data remains insufficient for effectively capturing differences in cellular environments for other thermophiles. Moreover, even the existing data predominantly originates from the *E. coli* expression system, leading to a significant underrepresentation of native thermophilic expression systems in structural studies. Moving forward, we advocate for future structural experiments to incorporate a broader array of expression systems and encompass a diverse range of cellular environment factors, including varying temperatures, pH levels, salinity, and other relevant environment descriptors. This approach will not only enhance our understanding of protein folding dynamics but also provide a more comprehensive insight into the influence of cellular environments on protein structure and function.

Data availability posed a challenge across various chapters of this research. Primarily, the dataset predominantly consisted of mesophilic species in terms of optimal growth temperature (OGT), with thermophiles being adequately represented. However, the representation of psychrophiles and hyperthermophiles was insufficient, prompting a decision to restrict the study to better reflect the available data landscape. Additionally, thermophilic species were significantly less researched compared to mesophiles in terms of UniProt functional tags and protein names, with a considerable portion of proteins from thermophiles remaining uncharacterized. The constraints we encountered hindered our ability to fully unravel the intricate connection between OGT and IDP abundance, underscoring the potential value of additional information on protein functions in enhancing our understanding.

As more information becomes available and research delves deeper into diverse cellular environments, researchers are poised to make significant strides in investigating the intricate relationships between protein structure, environmental adaptations, and species-specific variations. The evolving landscape of data availability holds promise for shedding light on complex biological phenomena and unraveling the mysteries surrounding protein folding dynamics in extreme environments. By expanding our understanding of how cellular factors influence protein structure and function, we can pave the way for innovative discoveries and transformative insights into the adaptive strategies of organisms across varying growth temperatures. With a growing wealth of knowledge and a commitment to exploring diverse biological contexts, researchers are primed to embark on exciting journeys of discovery that will shape the future of molecular biology and bioinformatics.

5.3 Future directions

As we look towards the future of research in protein structure and cellular environments, there is a compelling need to explore innovative avenues that can deepen our understanding of the intricate interplay between biological systems and environmental factors. Specifically, there are three major areas in which future research may be fruitful: development of predictive models that are tailored to specific cellular environments, integration of additional multi-omics data into analysis of cellular environment effects on folding and further exploration of protein functions and mechanisms behind environmental adaptations.

As more data become available, incorporation of cellular environment factors into protein structure prediction software offers significant potential in enhancing our ability to predict folding dynamics and accurately annotate functional properties. These models, designed to consider the nuances of varied cellular conditions, will be able to provide insights into how environmental factors like temperature, pH levels, and ionic concentrations influence protein folding. Tailored for specific cellular contexts, these predictive models can help identify unique structural features and functional properties of proteins, deepening our understanding of protein behavior across diverse biological settings. In the realm of IDPs, customized predictive models can elucidate the relationship between IDP abundance and growth temperature, offering valuable insights into how organisms adapt to different environmental conditions.

Furthermore, the evolving landscape of multi-omics data offers a wealth of opportunities to unravel complex relationships between cellular environments, protein structures, and functional outcomes. Particularly noteworthy is the potential utilization of transcriptomics data to assess protein expression levels, complementing the insights provided in the chapters of this study. For instance, transcriptomics data could be utilized to analyze gene expression and pinpoint genes associated with responses to temperature fluctuations in extremophiles (Xinglin Jiang et al. 2013; Teoh et al. 2023). This approach can help identify distinct folding patterns unique to these genes and facilitate the identification of IDPs within this protein subset. By integrating diverse omics datasets, researchers can gain comprehensive insights into how molecular interactions shape cellular processes and organismal adaptations.

When exploring the relationship between protein functions, folding mechanisms, and environmental adaptations, significant potential for research emerges. Investigating how organisms adapt to extreme environments through unique protein folding mechanisms and structural adaptations unveils novel strategies employed by living systems to thrive in challenging conditions. Furthermore, considering that a substantial number of proteins analyzed in our IDP study lack functional annotations, utilizing tools to predict the functions of these proteins becomes essential. For instance, software like DEPICTER2 (Basu, Gsponer, and Kurgan 2023), capable of predicting intrinsic disorder and disorder function, can be instrumental in this regard. Additionally, tools such as SAP (Urhan et al. 2023) and bacLIFE (Guerrero-Egido et al. 2024), while not IDP-specific but tailored for bacterial analysis, offer valuable resources for exploring protein functions and adaptations in bacterial systems.

These forward-looking pathways set the stage for revolutionary discoveries that will not only propel our understanding of protein biology forward but also illuminate the extraordinary resilience and adaptability of organisms across varied ecological landscapes. It is our aspiration that forthcoming research endeavors will yield valuable insights building upon the findings presented here and in other studies.

References

Adamczak, Rafał, Aleksey Porollo, and Jarosław Meller. 2005. “Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins.” *Proteins* 59 (3): 467–75. https://doi.org/10.1002/prot.20441.

Ahmed, Zahoor, Hasan Zulfiqar, Lixia Tang, and Hao Lin. 2022. “A Statistical Analysis of the Sequence and Structure of Thermophilic and Non-Thermophilic Proteins.” *International Journal of Molecular Sciences* 23 (17): 10116. https://doi.org/10.3390/ijms231710116.

Andersen, Kristian G., Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes, and Robert F. Garry. 2020. “The Proximal Origin of SARS-CoV-2.” *Nature Medicine* 26 (4): 450–52. https://doi.org/10.1038/s41591-020-0820-9.

Anfinsen, Christian B. 1973. “Principles That Govern the Folding of Protein Chains.” *Science* 181 (4096): 223–30.

Angelov, Angel, Christoph Loderer, Susanne Pompei, and Wolfgang Liebl. 2011. “Novel Family of Carbohydrate-Binding Modules Revealed by the Genome Sequence of Spirochaeta Thermophila DSM 6192.” *Applied and Environmental Microbiology* 77 (15): 5483–89. https://doi.org/10.1128/AEM.00523-11.

Angelov, Angel, Markus Mientus, Susanne Liebl, and Wolfgang Liebl. 2009. “A Two-Host Fosmid System for Functional Screening of (Meta)Genomic Libraries from Extreme Thermophiles.” *Systematic and Applied Microbiology* 32 (3): 177–85. https://doi.org/10.1016/j.syapm.2008.01.003.

Anger, Andreas M., Jean-Paul Armache, Otto Berninghausen, Michael Habeck, Marion Subklewe, Daniel N. Wilson, and Roland Beckmann. 2013. “Structures of the Human and Drosophila 80S Ribosome.” *Nature* 497 (7447): 80–85. https://doi.org/10.1038/nature12104.

Anjum, Farah, Taj Mohammad, Purva Asrani, Alaa Shafie, Shailza Singh, Dharmendra Kumar Yadav, Vladimir N. Uversky, and Md Imtaiyaz Hassan. 2022. “Identification of Intrinsically Disorder Regions in Non-Structural Proteins of SARS-CoV-2: New Insights into Drug and Vaccine Resistance.” *Molecular and Cellular Biochemistry* 477 (5): 1607–19. https://doi.org/10.1007/s11010-022-04393-5.

Apic, G., J. Gough, and S. A. Teichmann. 2001a. “An Insight into Domain Combinations.” *Bioinformatics (Oxford, England)* 17 Suppl 1:S83-89. https://doi.org/10.1093/bioinformatics/17.suppl_1.s83.

———. 2001b. “Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes.” *Journal of Molecular Biology* 310 (2): 311–25. https://doi.org/10.1006/jmbi.2001.4776.

Asai, Kiyoshi, Satoru Hayamizu, and Ken’ichi Handa. 1993. “Prediction of Protein Secondary Structure by the Hidden Markov Model.” *Bioinformatics* 9 (2): 141–46.

Balchin, David, Manajit Hayer-Hartl, and F. Ulrich Hartl. 2020. “Recent Advances in Understanding Catalysis of Protein Folding by Molecular Chaperones.” *FEBS Letters* 594 (17): 2770–81. https://doi.org/10.1002/1873-3468.13844.

Bancet, Alexandre, Claire Raingeval, Thierry Lomberget, Marc Le Borgne, Jean-François Guichou, and Isabelle Krimm. 2020. “Fragment Linking Strategies for Structure-Based Drug Design.” *Journal of Medicinal Chemistry* 63 (20): 11420–35. https://doi.org/10.1021/acs.jmedchem.0c00242.

Basu, Sushmita, Jörg Gsponer, and Lukasz Kurgan. 2023. “DEPICTER2: A Comprehensive Webserver for Intrinsic Disorder and Disorder Function Prediction.” *Nucleic Acids Research* 51 (W1): W141–47. https://doi.org/10.1093/nar/gkad330.

Bemporad, Francesco, Joerg Gsponer, Harri I Hopearuoho, Georgia Plakoutsi, Gianmarco Stati, Massimo Stefani, Niccolò Taddei, Michele Vendruscolo, and Fabrizio Chiti. 2008. “Biological Function in a Non-native Partially Folded State of a Protein.” *The EMBO Journal* 27 (10): 1525–35. https://doi.org/10.1038/emboj.2008.82.

Berg, Jeremy M., John L. Tymoczko, and Lubert Stryer. 2002. “Protein Structure and Function.” *Biochemistry. 5th Edition*. https://www.ncbi.nlm.nih.gov/books/NBK21177/.

Berlow, Rebecca B., H. Jane Dyson, and Peter E. Wright. 2022. “Multivalency Enables Unidirectional Switch-like Competition between Intrinsically Disordered Proteins.” *Proceedings of the National Academy of Sciences* 119 (3): e2117338119. https://doi.org/10.1073/pnas.2117338119.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42. https://doi.org/10.1093/nar/28.1.235.

Bernardini, Andrea, Pooja Mukherjee, Elisabeth Scheer, Ivanka Kamenova, Simona Antonova, Paulina Karen Mendoza Sanchez, Gizem Yayli, Bastien Morlet, H.T. Marc Timmers, and László Tora. 2023. “Hierarchical TAF1-Dependent Co-Translational Assembly of the Basal Transcription Factor TFIID.” *bioRxiv*, April, 2023.04.05.535704. https://doi.org/10.1101/2023.04.05.535704.

Bhattacharyya, Asima, Ranajoy Chattopadhyay, Sankar Mitra, and Sheila E. Crowe. 2014. “Oxidative Stress: An Essential Factor in the Pathogenesis of Gastrointestinal Mucosal Diseases.” *Physiological Reviews* 94 (2): 329–54. https://doi.org/10.1152/physrev.00040.2012.

Bondos, Sarah E., A. Keith Dunker, and Vladimir N. Uversky. 2022. “Intrinsically Disordered Proteins Play Diverse Roles in Cell Signaling.” *Cell Communication and Signaling* 20 (1): 20. https://doi.org/10.1186/s12964-022-00821-7.

Bondos, Sarah, and Kathleen Matthews. 2021. “Protein Folding.” https://doi.org/10.1036/1097-8542.801070.

Bosch, Sandra, Esther Sanchez-Freire, María Luisa del Pozo, Morana Česnik, Jaime Quesada, Diana M. Mate, Karel Hernández, et al. 2021. “Thermostability Engineering of a Class II Pyruvate Aldolase from Escherichia Coli by in Vivo Folding Interference.” *ACS Sustainable Chemistry & Engineering* 9 (15): 5430–36. https://doi.org/10.1021/acssuschemeng.1c00699.

Brady, G Patrick, and Kim A Sharp. 1997. “Entropy in Protein Folding and in Protein—Protein Interactions.” *Current Opinion in Structural Biology* 7 (2): 215–21.

Broncel, Malgorzata, Jessica A. Falenski, Sara C. Wagner, Christian P. R. Hackenberger, and Beate Koksch. 2010. “How Post-Translational Modifications Influence Amyloid Formation: A Systematic Study of Phosphorylation and Glycosylation in Model Peptides.” *Chemistry – A European Journal* 16 (26): 7881–88. https://doi.org/10.1002/chem.200902452.

Bull, Rowena A., John-Sebastian Eden, William D. Rawlinson, and Peter A. White. 2010. “Rapid Evolution of Pandemic Noroviruses of the GII.4 Lineage.” *PLoS Pathogens* 6 (3): e1000831. https://doi.org/10.1371/journal.ppat.1000831.

Burley, Stephen K, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. 2017. “Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.” In *Protein Crystallography*, 627–41. Springer.

Burra, Prasad V., Lajos Kalmar, and Peter Tompa. 2010. “Reduction in Structural Disorder and Functional Complexity in the Thermal Adaptation of Prokaryotes.” *PLOS ONE* 5 (8): e12069. https://doi.org/10.1371/journal.pone.0012069.

Cario, Alisa, Adriana Savastano, Neil B. Wood, Zhu Liu, Michael J. Previs, Adam G. Hendricks, Markus Zweckstetter, and Christopher L. Berger. 2022. “The Pathogenic R5L Mutation Disrupts Formation of Tau Complexes on the Microtubule by Altering Local N-Terminal Structure.” *Proceedings of the National Academy of Sciences of the United States of America* 119 (7): e2114215119. https://doi.org/10.1073/pnas.2114215119.

Chautard, Helene, Emilio Blas-Galindo, Thierry Menguy, Laure Grand&amp, apos, Moursel, Felipe Cava, Jose Berenguer, and Marc Delcourt. 2007. “An Activity-Independent Selection System of Thermostable Protein Variants.” *Nature Methods* 4 (11): 919–22.

Chayen, N., J. Akins, S. Campbell-Smith, and D. M. Blow. 1988. “Solubility of Glucose Isomerase in Ammonium Sulphate Solutions.” *Journal of Crystal Growth* 90 (1–3): 112–16.

Chen, Qionghua, Yuelin Shen, and Jingyang Zheng. 2021. “A Review of Cystic Fibrosis: Basic and Clinical Aspects.” *Animal Models and Experimental Medicine* 4 (3): 220–32. https://doi.org/10.1002/ame2.12180.

Chen, Yun, Michael R. Strickland, Andrea Soranno, and David M. Holtzman. 2021. “Apolipoprotein E: Structural Insights and Links to Alzheimer Disease Pathogenesis.” *Neuron* 109 (2): 205–21. https://doi.org/10.1016/j.neuron.2020.10.008.

Cho, ByeongJin, Jaejun Choi, RyeongHyeon Kim, Jean Nyoung Yun, Yuri Choi, Hyung Ho Lee, and Junseock Koh. 2021. “Thermodynamic Models for Assembly of Intrinsically Disordered Protein Hubs with Multiple Interaction Partners.” *Journal of the American Chemical Society* 143 (32): 12509–23. https://doi.org/10.1021/jacs.1c00811.

Chong, Song-Ho, and Sihyun Ham. 2019. “Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins.” *Scientific Reports* 9 (October):14927. https://doi.org/10.1038/s41598-019-50825-6.

Chou, P. Y., and G. D. Fasman. 1974. “Prediction of Protein Conformation.” *Biochemistry* 13 (2): 222–45. https://doi.org/10.1021/bi00699a002.

Clark, Karen, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2016. “GenBank.” *Nucleic Acids Research* 44 (D1): D67-72. https://doi.org/10.1093/nar/gkv1276.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23. https://doi.org/10.1093/bioinformatics/btp163.

Coutard, B., C. Valle, X. de Lamballerie, B. Canard, N. G. Seidah, and E. Decroly. 2020. “The Spike Glycoprotein of the New Coronavirus 2019-nCoV Contains a Furin-like Cleavage Site Absent in CoV of the Same Clade.” *Antiviral Research* 176 (April):104742. https://doi.org/10.1016/j.antiviral.2020.104742.

Crooks, Gavin E., Gary Hon, John-Marc Chandonia, and Steven E. Brenner. 2004. “WebLogo: A Sequence Logo Generator.” *Genome Research* 14 (6): 1188–90. https://doi.org/10.1101/gr.849004.

Cuff, J. A., M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton. 1998. “JPred: A Consensus Secondary Structure Prediction Server.” *Bioinformatics (Oxford, England)* 14 (10): 892–93. https://doi.org/10.1093/bioinformatics/14.10.892.

Darling, April L., and Vladimir N. Uversky. 2018. “Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter.” *Frontiers in Genetics* 9 (May):158. https://doi.org/10.3389/fgene.2018.00158.

Das, Rahul K., and Rohit V. Pappu. 2013. “Conformations of Intrinsically Disordered Proteins Are Influenced by Linear Sequence Distributions of Oppositely Charged Residues.” *Proceedings of the National Academy of Sciences* 110 (33): 13392–97. https://doi.org/10.1073/pnas.1304749110.

DeForte, Shelly, and Vladimir N. Uversky. 2017. “Not an Exception to the Rule: The Functional Significance of Intrinsically Disordered Protein Regions in Enzymes.” *Molecular BioSystems* 13 (3): 463–69. https://doi.org/10.1039/C6MB00741D.

Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. 2015. “JPred4: A Protein Secondary Structure Prediction Server.” *Nucleic Acids Research* 43 (Web Server issue): W389–94. https://doi.org/10.1093/nar/gkv332.

Dunker, A. Keith, Celeste J. Brown, J. David Lawson, Lilia M. Iakoucheva, and Zoran Obradović. 2002. “Intrinsic Disorder and Protein Function.” *Biochemistry* 41 (21): 6573–82. https://doi.org/10.1021/bi012159+.

Durairaj, D. Ruban, and P. Shanmughavel. 2019. “In Silico Drug Design of Thiolactomycin Derivatives Against Mtb-KasA Enzyme to Inhibit Multidrug Resistance ofMycobacterium Tuberculosis.” *Interdisciplinary Sciences: Computational Life Sciences* 11 (2): 215–25. https://doi.org/10.1007/s12539-017-0257-0.

Eben, Stefanie S., and James A. Imlay. 2023. “Excess Copper Catalyzes Protein Disulfide Bond Formation in the Bacterial Periplasm but Not in the Cytoplasm.” *Molecular Microbiology* 119 (4): 423–38. https://doi.org/10.1111/mmi.15032.

Ekman, Diana, Asa K. Björklund, and Arne Elofsson. 2007. “Quantification of the Elevated Rate of Domain Rearrangements in Metazoa.” *Journal of Molecular Biology* 372 (5): 1337–48. https://doi.org/10.1016/j.jmb.2007.06.022.

Ekman, Diana, Asa K. Björklund, Johannes Frey-Skött, and Arne Elofsson. 2005. “Multi-Domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions.” *Journal of Molecular Biology* 348 (1): 231–43. https://doi.org/10.1016/j.jmb.2005.02.007.

Ellis, R. John. 2001. “Macromolecular Crowding: An Important but Neglected Aspect of the Intracellular Environment.” *Current Opinion in Structural Biology* 11 (4): 500. https://doi.org/10.1016/S0959-440X(00)00239-6.

Erdős, Gábor, Mátyás Pajkos, and Zsuzsanna Dosztányi. 2021. “IUPred3: Prediction of Protein Disorder Enhanced with Unambiguous Experimental Annotation and Visualization of Evolutionary Conservation.” *Nucleic Acids Research* 49 (W1): W297–303. https://doi.org/10.1093/nar/gkab408.

Fang, Chao, Yi Shang, and Dong Xu. 2018. “MUFOLD-SS: New Deep Inception-inside-Inception Networks for Protein Secondary Structure Prediction.” *Proteins* 86 (5): 592–98. https://doi.org/10.1002/prot.25487.

Fanning, Saranna, Dennis Selkoe, and Ulf Dettmer. 2020. “Parkinson’s Disease: Proteinopathy or Lipidopathy?” *Npj Parkinson’s Disease* 6 (1): 1–9. https://doi.org/10.1038/s41531-019-0103-7.

Farinha, Carlos M., and Isabelle Callebaut. 2022. “Molecular Mechanisms of Cystic Fibrosis – How Mutations Lead to Misfunction and Guide Therapy.” *Bioscience Reports* 42 (7): BSR20212006. https://doi.org/10.1042/BSR20212006.

Ferreon, Allan Chris M., Yann Gambin, Edward A. Lemke, and Ashok A. Deniz. 2009. “Interplay of Alpha-Synuclein Binding and Conformational Switching Probed by Single-Molecule Fluorescence.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (14): 5645–50. https://doi.org/10.1073/pnas.0809232106.

Forood, B., E. J. Feliciano, and K. P. Nambiar. 1993. “Stabilization of Alpha-Helical Structures in Short Peptides via End Capping.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (3): 838–42. https://doi.org/10.1073/pnas.90.3.838.

Francoeur, Paul G., Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. 2020. “Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design.” *Journal of Chemical Information and Modeling* 60 (9): 4200–4215. https://doi.org/10.1021/acs.jcim.0c00411.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics (Oxford, England)* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

Fujino, Yasuhiro, Shuichiro Goda, Yuri Suematsu, and Katsumi Doi. 2020. “Development of a New Gene Expression Vector for Thermus Thermophilus Using a Silica-Inducible Promoter.” *Microbial Cell Factories* 19 (1): 126. https://doi.org/10.1186/s12934-020-01385-2.

Giffard, Rona G., Lijun Xu, Heng Zhao, Whitney Carrico, Yibing Ouyang, Yanli Qiao, Robert Sapolsky, Gary Steinberg, Bingren Hu, and Midori A. Yenari. 2004. “Chaperones, Protein Aggregation, and Brain Protection from Hypoxic/Ischemic Injury.” *Journal of Experimental Biology* 207 (18): 3213–20. https://doi.org/10.1242/jeb.01034.

Goda, Shuichiro, Masaki Kojima, Yoshimi Nishikawa, Chizu Kujo, Ryushi Kawakami, Seiki Kuramitsu, Haruhiko Sakuraba, Yuzuru Hiragi, and Toshihisa Ohshima. 2005. “Intersubunit Interaction Induced by Subunit Rearrangement Is Essential for the Catalytic Activity of the Hyperthermophilic Glutamate Dehydrogenase from Pyrobaculum Islandicum.” *Biochemistry* 44 (46): 15304–13. https://doi.org/10.1021/bi050478l.

Godin-Roulling, Amandine, Philipp AM Schmidpeter, Franz X. Schmid, and Georges Feller. 2015. “Functional Adaptations of the Bacterial Chaperone Trigger Factor to Extreme Environmental Temperatures.” *Environmental Microbiology* 17 (7): 2407–20.

Gohlke, Holger, Manfred Hendlich, and Gerhard Klebe. 2000. “Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions.” *Journal of Molecular Biology* 295 (2): 337–56.

Gomes, Cláudio M., and Patrícia F. N. Faísca. 2019. “Protein Folding: An Introduction.” In *Protein Folding: An Introduction*, edited by Cláudio M. Gomes and Patrícia F.N. Faísca, 1–63. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00882-0_1.

Gromiha, M. Michael, Manish C. Pathak, Kadhirvel Saraboji, Eric A. Ortlund, and Eric A. Gaucher. 2013. “Hydrophobic Environment Is a Key Factor for the Stability of Thermophilic Proteins.” *Proteins* 81 (4): 715–21. https://doi.org/10.1002/prot.24232.

Grossman, Murray. 2019. “Amyotrophic Lateral Sclerosis - a Multisystem Neurodegenerative Disorder.” *Nature Reviews. Neurology* 15 (1): 5–6. https://doi.org/10.1038/s41582-018-0103-y.

Gsponer, Jörg, Matthias E. Futschik, Sarah A. Teichmann, and M. Madan Babu. 2008. “Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation.” *Science* 322 (5906): 1365–68.

Gu, Jinge, Chen Wang, Rirong Hu, Yichen Li, Shengnan Zhang, Yunpeng Sun, Qiangqiang Wang, Dan Li, Yanshan Fang, and Cong Liu. 2021. “Hsp70 Chaperones TDP-43 in Dynamic, Liquid-like Phase and Prevents It from Amyloid Aggregation.” *Cell Research* 31 (9): 1024–27. https://doi.org/10.1038/s41422-021-00526-5.

Guan, Y., B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, et al. 2003. “Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China.” *Science (New York, N.Y.)* 302 (5643): 276–78. https://doi.org/10.1126/science.1087139.

Guerrero-Egido, Guillermo, Adrian Pintado, Kevin M. Bretscher, Luisa-Maria Arias-Giraldo, Joseph N. Paulson, Herman P. Spaink, Dennis Claessen, et al. 2024. “bacLIFE: A User-Friendly Computational Workflow for Genome Analysis and Prediction of Lifestyle-Associated Genes in Bacteria.” *Nature Communications* 15 (March):2072. https://doi.org/10.1038/s41467-024-46302-y.

Harrington, Lucas B., Ramesh K. Jha, Theresa L. Kern, Emily N. Schmidt, Gustavo M. Canales, Kellan B. Finney, Andrew T. Koppisch, Charlie E. M. Strauss, and David T. Fox. 2017. “Rapid Thermostabilization of Bacillus Thuringiensis Serovar Konkukian 97–27 Dehydroshikimate Dehydratase through a Structure-Based Enzyme Design and Whole Cell Activity Assay.” *ACS Synthetic Biology* 6 (1): 120–29. https://doi.org/10.1021/acssynbio.6b00159.

Hartl, F. Ulrich, Andreas Bracher, and Manajit Hayer-Hartl. 2011. “Molecular Chaperones in Protein Folding and Proteostasis.” *Nature* 475 (7356): 324–32. https://doi.org/10.1038/nature10317.

Heffernan, Rhys, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. 2017. “Capturing Non-Local Interactions by Long Short-Term Memory Bidirectional Recurrent Neural Networks for Improving Prediction of Protein Secondary Structure, Backbone Angles, Contact Numbers and Solvent Accessibility.” *Bioinformatics (Oxford, England)* 33 (18): 2842–49. https://doi.org/10.1093/bioinformatics/btx218.

Hidalgo, Aurelio, Lorena Betancor, Renata Moreno, Olga Zafra, Felipe Cava, Roberto Fernández-Lafuente, José M. Guisán, and José Berenguer. 2004. “Thermus Thermophilus as a Cell Factory for the Production of a Thermophilic Mn-Dependent Catalase Which Fails To Be Synthesized in an Active Form in Escherichia Coli.” *Applied and Environmental Microbiology* 70 (7): 3839–44. https://doi.org/10.1128/AEM.70.7.3839-3844.2004.

Hoffmann, Markus, Heike Hofmann-Winkler, and Stefan Pöhlmann. 2018. “Priming Time: How Cellular Proteases Arm Coronavirus Spike Proteins.” In *Activation of Viruses by Host Proteases*, edited by Eva Böttcher-Friebertshäuser, Wolfgang Garten, and Hans Dieter Klenk, 71–98. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-75474-1_4.

Holmes, Kathryn V. 2005. “Structural Biology. Adaptation of SARS Coronavirus to Humans.” *Science (New York, N.Y.)* 309 (5742): 1822–23. https://doi.org/10.1126/science.1118817.

Honig, Barry, and Lawrence Shapiro. 2020. “Adhesion Protein Structure, Molecular Affinities, and Principles of Cell-Cell Recognition.” *Cell* 181 (3): 520–35. https://doi.org/10.1016/j.cell.2020.04.010.

Hou, Yuxuan, Cheng Peng, Meng Yu, Yan Li, Zhenggang Han, Fang Li, Lin-Fa Wang, and Zhengli Shi. 2010. “Angiotensin-Converting Enzyme 2 (ACE2) Proteins of Different Bat Species Confer Variable Susceptibility to SARS-CoV Entry.” *Archives of Virology* 155 (10): 1563–69. https://doi.org/10.1007/s00705-010-0729-6.

Hu, Gang, Akila Katuwawala, Kui Wang, Zhonghua Wu, Sina Ghadermarzi, Jianzhao Gao, and Lukasz Kurgan. 2021. “flDPnn: Accurate Intrinsic Disorder Prediction with Putative Propensities of Disorder Functions.” *Nature Communications* 12 (1): 4438. https://doi.org/10.1038/s41467-021-24773-7.

Hu, Shenglan, Jieqiong Tan, Lixia Qin, Lingling Lv, Weiqian Yan, Hainan Zhang, BeiSha Tang, and Chunyu Wang. 2021. “Molecular Chaperones and Parkinson’s Disease.” *Neurobiology of Disease* 160 (December):105527. https://doi.org/10.1016/j.nbd.2021.105527.

Hua, S., and Z. Sun. 2001. “A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach.” *Journal of Molecular Biology* 308 (2): 397–407. https://doi.org/10.1006/jmbi.2001.4580.

Huang, Xiaoqiang, Robin Pearce, and Yang Zhang. 2020a. “De Novo Design of Protein Peptides to Block Association of the SARS-CoV-2 Spike Protein with Human ACE2.” *Aging (Albany NY)* 12 (12): 11263–76. https://doi.org/10.18632/aging.103416.

———. 2020b. “EvoEF2: Accurate and Fast Energy Function for Computational Protein Design.” *Bioinformatics (Oxford, England)* 36 (4): 1135–42. https://doi.org/10.1093/bioinformatics/btz740.

Hulswit, R. J. G., C. a. M. de Haan, and B.-J. Bosch. 2016. “Coronavirus Spike Protein and Tropism Changes.” *Advances in Virus Research* 96:29–57. https://doi.org/10.1016/bs.aivir.2016.08.004.

Jaenicke, Rainer. n.d. “Protein Stability and Protein Folding.” In *Ciba Foundation Symposium 161 - Protein Conformation*, 206–21. John Wiley & Sons, Ltd. Accessed March 16, 2024. https://doi.org/10.1002/9780470514146.ch13.

Jiang, Xinglin, Haibo Zhang, Jianming Yang, Min Liu, Hongru Feng, Xiaobin Liu, Yujin Cao, Dexin Feng, and Mo Xian. 2013. “Induction of Gene Expression in Bacteria at Optimal Growth Temperatures.” *Applied Microbiology and Biotechnology* 97 (12): 5423–31. https://doi.org/10.1007/s00253-012-4633-8.

Jiang, Xuejun, Hyun-Eui Kim, Hongjun Shu, Yingming Zhao, Haichao Zhang, James Kofron, Jennifer Donnelly, et al. 2003. “Distinctive Roles of PHAP Proteins and Prothymosin-α in a Death Regulatory Pathway.” *Science* 299 (5604): 223–26. https://doi.org/10.1126/science.1076807.

Johnson, Mark, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L. Madden. 2008. “NCBI BLAST: A Better Web Interface.” *Nucleic Acids Research* 36 (Web Server issue): W5-9. https://doi.org/10.1093/nar/gkn201.

Jumper, J., Tunyasuvunakool, K., Kohli, P., and Hassabis, D. n.d. “Computational Predictions of Protein Structures Associated with COVID-19.” Accessed December 1, 2022. https://www.deepmind.com/open-source/computational-predictions-of-protein-structures-associated-with-covid-19.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kaffe-Abramovich, Tamar, and Ron Unger. 1998. “A Simple Model for Evolution of Proteins towards the Global Minimum of Free Energy.” *Folding and Design* 3 (5): 389–99. https://doi.org/10.1016/S1359-0278(98)00052-2.

Kai, Mihoko. 2016. “Roles of RNA-Binding Proteins in DNA Damage Response.” *International Journal of Molecular Sciences* 17 (3): 310. https://doi.org/10.3390/ijms17030310.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kaur, Jashandeep, Arbind Kumar, and Jagdeep Kaur. 2018. “Strategies for Optimization of Heterologous Protein Expression in E. Coli: Roadblocks and Reinforcements.” *International Journal of Biological Macromolecules* 106 (January):803–22. https://doi.org/10.1016/j.ijbiomac.2017.08.080.

Kelaini, Sophia, Celine Chan, Victoria A Cornelius, and Andriana Margariti. 2021. “RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease.” *Biology* 10 (5): 366. https://doi.org/10.3390/biology10050366.

Khan, Salman Ali, Komal Zia, Sajda Ashraf, Reaz Uddin, and Zaheer Ul-Haq. 2021. “Identification of Chymotrypsin-like Protease Inhibitors of SARS-CoV-2 via Integrated Computational Approach.” *Journal of Biomolecular Structure and Dynamics* 39 (7): 2607–16. https://doi.org/10.1080/07391102.2020.1751298.

Kim, Jun Soo, and Arun Yethiraj. 2011. “Crowding Effects on Protein Association: Effect of Interactions between Crowding Agents.” *The Journal of Physical Chemistry B* 115 (2): 347–53. https://doi.org/10.1021/jp107123y.

Kim, Youngchang, Lance Bigelow, Maria Borovilos, Irina Dementieva, Erika Duggan, William eschenfeldt, Catherine Hatzos, et al. 2008. “High-Throughput Protein Purification for X-Ray Crystallography and NMR.” In *Advances in Protein Chemistry and Structural Biology*, edited by Andrzej Joachimiak, 75:85–105. Structural Genomics, Part A. Academic Press. https://doi.org/10.1016/S0065-3233(07)75003-9.

Kinjo, Akira R., and Shoji Takada. 2002. “Effects of Macromolecular Crowding on Protein Folding and Aggregation Studied by Density Functional Theory: Dynamics.” *Physical Review E* 66 (5): 051902. https://doi.org/10.1103/PhysRevE.66.051902.

Kiss, Gert, Nihan Çelebi-Ölçüm, Rocco Moretti, David Baker, and KN Houk. 2013. “Computational Enzyme Design.” *Angewandte Chemie International Edition* 52 (22): 5700–5725.

Kloczkowski, A., K. -L. Ting, R. L. Jernigan, and J. Garnier. 2002. “Protein Secondary Structure Prediction Based on the GOR Algorithm Incorporating Multiple Sequence Alignment Information.” *Polymer* 43 (2): 441–49. https://doi.org/10.1016/S0032-3861(01)00425-6.

Koga, Nobuyasu, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B Acton, Gaetano T Montelione, and David Baker. 2012. “Principles for Designing Ideal Protein Structures.” *Nature* 491 (7423): 222.

Koh, Hye Yeon, Hyun Park, Jun Hyuck Lee, Se Jong Han, Young Chang Sohn, and Sung Gu Lee. 2017. “Proteomic and Transcriptomic Investigations on Cold-Responsive Properties of the Psychrophilic Antarctic Bacterium Psychrobacter Sp. PAMC 21119 at Subzero Temperatures.” *Environmental Microbiology* 19 (2): 628–44. https://doi.org/10.1111/1462-2920.13578.

Kolonko, Marta, Dominika Bystranowska, Michał Taube, Maciej Kozak, Mark Bostock, Grzegorz Popowicz, Andrzej Ożyhar, and Beata Greb-Markiewicz. 2020. “The

Intrinsically Disordered Region of GCEprotein Adopts a More Fixed Structure by Interacting with the LBD of the Nuclearreceptor FTZ-F1." *Cell Communication and Signaling* 18 (1): 180. https://doi.org/10.1186/s12964-020-00662-2.

Konaté, Mariam M, Germán Plata, Jimin Park, Dinara R Usmanova, Harris Wang, and Dennis Vitkup. 2019. "Molecular Function Limits Divergent Protein Evolution on Planetary Timescales." Edited by Nir Ben-Tal, Diethard Tautz, and Nir Ben-Tal. *eLife* 8 (September):e39705. https://doi.org/10.7554/eLife.39705.

Konermann, Lars. 2012. "Protein Unfolding and Denaturants." In *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0003004.pub2.

Krefft, Daria, Aliaksei Papkov, Agnieszka Zylicz-Stachula, and Piotr M. Skowron. 2017. "Thermostable Proteins Bioprocesses: The Activity of Restriction Endonuclease-Methyltransferase from Thermus Thermophilus (RM.TthHB27I) Cloned in Escherichia Coli Is Critically Affected by the Codon Composition of the Synthetic Gene." *PLoS ONE* 12 (10). https://doi.org/10.1371/journal.pone.0186633.

Krivov, Georgii G., Maxim V. Shapovalov, and Roland L. Dunbrack. 2009. "Improved Prediction of Protein Side-Chain Conformations with SCWRL4." *Proteins: Structure, Function, and Bioinformatics* 77 (4): 778–95. https://doi.org/10.1002/prot.22488.

Kruglikov, Alibek, Mohan Rakesh, Yulong Wei, and Xuhua Xia. 2021. "Applications of Protein Secondary Structure Algorithms in SARS-CoV-2 Research." *Journal of Proteome Research* 20 (3): 1457–63. https://doi.org/10.1021/acs.jproteome.0c00734.

Kruglikov, Alibek, Yulong Wei, and Xuhua Xia. 2022. "Proteins from Thermophilic Thermus Thermophilus Often Do Not Fold Correctly in a Mesophilic Expression System Such as Escherichia Coli." *ACS Omega* 7 (42): 37797–806. https://doi.org/10.1021/acsomega.2c04786.

Kruglikov, Alibek, and Xuhua Xia. 2024. "Mesophiles vs. Thermophiles: Untangling the Hot Mess of Intrinsically Disordered Proteins and Growth Temperature of Bacteria." *International Journal of Molecular Sciences* 25 (4): 2000. https://doi.org/10.3390/ijms25042000.

Kurgan, Lukasz, Gang Hu, Kui Wang, Sina Ghadermarzi, Bi Zhao, Nawar Malhis, Gábor Erdős, Jörg Gsponer, Vladimir N. Uversky, and Zsuzsanna Dosztányi. 2023. "Tutorial: A Guide for the Selection of Fast and Accurate Computational Tools for the Prediction of Intrinsic Disorder in Proteins." *Nature Protocols*, September. https://doi.org/10.1038/s41596-023-00876-x.

Kurzbach, Dennis, Thomas C. Schwarz, Gerald Platzer, Simone Höfler, Dariush Hinderberger, and Robert Konrat. 2014. "Compensatory Adaptations of Structural Dynamics in an Intrinsically Disordered Protein Complex." *Angewandte Chemie (International Ed. in English)* 53 (15): 3840–43. https://doi.org/10.1002/anie.201308389.

Kuznetsova, Irina M., Konstantin K. Turoverov, and Vladimir N. Uversky. 2014. "What Macromolecular Crowding Can Do to a Protein." *International Journal of Molecular Sciences* 15 (12): 23090–140. https://doi.org/10.3390/ijms151223090.

Kwon, S. Chul, Tuan Anh Nguyen, Yeon-Gil Choi, Myung Hyun Jo, Sungchul Hohng, V. Narry Kim, and Jae-Sung Woo. 2016. "Structure of Human DROSHA." *Cell* 164 (1–2): 81–90. https://doi.org/10.1016/j.cell.2015.12.019.

Lake, J. A. 1994. "Reconstructing Evolutionary Trees from DNA and Protein Sequences: Paralinear Distances." *Proceedings of the National Academy of Sciences of the United States of America* 91 (4): 1455–59. https://doi.org/10.1073/pnas.91.4.1455.

Lan, Jun, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, et al. 2020. "Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor." *Nature* 581 (7807): 215–20. https://doi.org/10.1038/s41586-020-2180-5.

Lee, Michael S., Freddie R. Salsbury Jr., and Charles L. Brooks III. 2004. "Constant-pH Molecular Dynamics Using Continuous Titration Coordinates." *Proteins: Structure, Function, and Bioinformatics* 56 (4): 738–52. https://doi.org/10.1002/prot.20128.

Lee, Robin van der, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, et al. 2014. "Classification of Intrinsically Disordered Regions and Proteins." *Chemical Reviews* 114 (13): 6589–6631. https://doi.org/10.1021/cr400525m.

Leis, Benedikt, Angel Angelov, Markus Mientus, Haijuan Li, Vu T. T. Pham, Benjamin Lauinger, Patrick Bongen, et al. 2015. "Identification of Novel Esterase-Active Enzymes from Hot Environments by Use of the Host Bacterium Thermus Thermophilus." *Frontiers in Microbiology* 6. https://www.frontiersin.org/articles/10.3389/fmicb.2015.00275.

Lennicke, Claudia, Jette Rahn, Nadine Heimer, Rudolf Lichtenfels, Ludger A. Wessjohann, and Barbara Seliger. 2016. "Redox Proteomics: Methods for the Identification and Enrichment of Redox-Modified Proteins and Their Applications." *PROTEOMICS* 16 (2): 197–213. https://doi.org/10.1002/pmic.201500268.

Lewis, Hal A., Xun Zhao, Chi Wang, J. Michael Sauder, Isabelle Rooney, Brian W. Noland, Don Lorimer, et al. 2005. "Impact of the deltaF508 Mutation in First Nucleotide-Binding Domain of Human Cystic Fibrosis Transmembrane Conductance Regulator on Domain Folding and Structure." *The Journal of Biological Chemistry* 280 (2): 1346–53. https://doi.org/10.1074/jbc.M410968200.

Li, Lingyan, Mifang Ren, Yueqiang Xu, Cheng Jin, Wenhao Zhang, and Xiuzhu Dong. 2020. "Enhanced Glycosylation of an S-Layer Protein Enables a Psychrophilic Methanogenic Archaeon to Adapt to Elevated Temperatures in Abundant Substrates." *FEBS Letters* 594 (4): 665–77. https://doi.org/10.1002/1873-3468.13650.

Li, Wenhui, Thomas C. Greenough, Michael J. Moore, Natalya Vasilieva, Mohan Somasundaran, John L. Sullivan, Michael Farzan, and Hyeryun Choe. 2004. "Efficient Replication of Severe Acute Respiratory Syndrome Coronavirus in Mouse Cells Is Limited by Murine Angiotensin-Converting Enzyme 2." *Journal of Virology* 78 (20): 11429–33. https://doi.org/10.1128/JVI.78.20.11429-11433.2004.

Liang, Fu-Cheng, Rita P. -Y. Chen, Chun-Cheng Lin, Kuo-Ting Huang, and Sunney I. Chan. 2006. "Tuning the Conformation Properties of a Peptide by Glycosylation and Phosphorylation." *Biochemical and Biophysical Research Communications* 342 (2): 482–88. https://doi.org/10.1016/j.bbrc.2006.01.168.

Liu, Jinfeng, and Burkhard Rost. 2004. "CHOP Proteins into Structural Domain-like Fragments." *Proteins* 55 (3): 678–88. https://doi.org/10.1002/prot.20095.

López-López, Olalla, María-Esperanza Cerdán, and María-Isabel González-Siso. 2015. "Thermus Thermophilus as a Source of Thermostable Lipolytic Enzymes." *Microorganisms* 3 (4): 792–808. https://doi.org/10.3390/microorganisms3040792.

Lu, Guangwen, Qihui Wang, and George F. Gao. 2015. "Bat-to-Human: Spike Features Determining 'host Jump' of Coronaviruses SARS-CoV, MERS-CoV, and Beyond." *Trends in Microbiology* 23 (8): 468–78. https://doi.org/10.1016/j.tim.2015.06.003.

Lu, Jia-Hai, Ding-Mei Zhang, Guo-Ling Wang, Zhong-Min Guo, Juan Li, Bing-Yan Tan, Li-Ping Ou-Yang, Wen-Hua Ling, Xin-Bing Yu, and Nan-Shan Zhong. 2005. "Sequence Analysis and Structural Prediction of the Severe Acute Respiratory Syndrome Coronavirus Nsp5." *Acta Biochimica Et Biophysica Sinica* 37 (7): 473–79. https://doi.org/10.1111/j.1745-7270.2005.00066.x.

Lund, O., Nielsen, M., Lundegaard, C., and Worning, P. 2002. "CPH Models 2.0: X3M a Computer Program to Extract 3D Models." In .

Luo, Shitong, Jiaqi Guan, Jianzhu Ma, and Jian Peng. 2021. "A 3D Generative Model for Structure-Based Drug Design." In *Advances in Neural Information Processing Systems*, 34:6229–39. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/hash/314450613369e0ee72d0da7f6fee773c-Abstract.html.

Luo, Yin, Buyong Ma, Ruth Nussinov, and Guanghong Wei. 2014. "Structural Insight into Tau Protein's Paradox of Intrinsically Disordered Behavior, Self-Acetylation Activity, and Aggregation." *The Journal of Physical Chemistry Letters* 5 (17): 3026–31. https://doi.org/10.1021/jz501457f.

Macdonald, Bryanne, Shannon McCarley, Sundus Noeen, and Alan E. van Giessen. 2016. "β-Hairpin Crowding Agents Affect α-Helix Stability in Crowded Environments." *The Journal of Physical Chemistry. B* 120 (4): 650–59. https://doi.org/10.1021/acs.jpcb.5b10575.

Mao, Albert H., Scott L. Crick, Andreas Vitalis, Caitlin L. Chicoine, and Rohit V. Pappu. 2010. "Net Charge per Residue Modulates Conformational Ensembles of Intrinsically Disordered Proteins." *Proceedings of the National Academy of Sciences* 107 (18): 8183–88. https://doi.org/10.1073/pnas.0911107107.

Mardis, Elaine R. 2006. "Anticipating the 1,000 Dollar Genome." *Genome Biology* 7 (7): 112. https://doi.org/10.1186/gb-2006-7-7-112.

Mate, Diana M., Noé R. Rivera, Esther Sanchez-Freire, Juan A. Ayala, José Berenguer, and Aurelio Hidalgo. 2020. "Thermostability Enhancement of the Pseudomonas Fluorescens Esterase I by in Vivo Folding Selection in Thermus Thermophilus." *Biotechnology and Bioengineering* 117 (1): 30–38. https://doi.org/10.1002/bit.27170.

McGuffin, L. J., K. Bryson, and D. T. Jones. 2000. "The PSIPRED Protein Structure Prediction Server." *Bioinformatics (Oxford, England)* 16 (4): 404–5. https://doi.org/10.1093/bioinformatics/16.4.404.

McPherson, Alexander. 1985. "[7] Crystallization of Proteins by Variation of pH or Temperature." In *Methods in Enzymology*, 114:125–27. Elsevier.

Migliori, Valentina, Sameer Phalke, Marco Bezzi, and Ernesto Guccione. 2010. "Arginine/Lysine-Methyl/Methyl Switches: Biochemical Role of Histone Arginine Methylation in Transcriptional Regulation." *Epigenomics* 2 (1): 119–37. https://doi.org/10.2217/epi.09.39.

Miklos, Andrew C., Mohona Sarkar, Yaqiang Wang, and Gary J. Pielak. 2011. "Protein Crowding Tunes Protein Stability." *Journal of the American Chemical Society* 133 (18): 7116–20. https://doi.org/10.1021/ja200067p.

Mikol, Vincent, and Richard Giegé. 1989. "Phase Diagram of a Crystalline Protein: Determination of the Solubility of Concanavalin A by a Microquantitation Assay." *Journal of Crystal Growth* 97 (2): 324–32.

Miotto, Mattia, Pier Paolo Olimpieri, Lorenzo Di Rienzo, Francesco Ambrosetti, Pietro Corsi, Rosalba Lepore, Gian Gaetano Tartaglia, and Edoardo Milanetti. 2019. “Insights on Protein Thermal Stability: A Graph Representation of Molecular Interactions.” *Bioinformatics* 35 (15): 2569–77. https://doi.org/10.1093/bioinformatics/bty1011.

Necci, Marco, Damiano Piovesan, and Silvio C. E. Tosatto. 2021. “Critical Assessment of Protein Intrinsic Disorder Prediction.” *Nature Methods* 18 (5): 472–81. https://doi.org/10.1038/s41592-021-01117-3.

Nedialkova, Danny D., and Sebastian A. Leidel. 2015. “Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity.” *Cell* 161 (7): 1606–18. https://doi.org/10.1016/j.cell.2015.05.022.

Niehaus, F., C. Bertoldo, M. Kähler, and G. Antranikian. 1999. “Extremophiles as a Source of Novel Enzymes for Industrial Application.” *Applied Microbiology and Biotechnology* 51 (6): 711–29. https://doi.org/10.1007/s002530051456.

Ninh, Pham Huynh, Kohsuke Honda, Takaaki Sakai, Kenji Okano, and Hisao Ohtake. 2015. “Assembly and Multiple Gene Expression of Thermophilic Enzymes in Escherichia Coli for in Vitro Metabolic Engineering.” *Biotechnology and Bioengineering* 112 (1): 189–96. https://doi.org/10.1002/bit.25338.

Pancsa, Rita, Denes Kovacs, and Peter Tompa. 2019. “Misprediction of Structural Disorder in Halophiles.” *Molecules* 24 (3): 479. https://doi.org/10.3390/molecules24030479.

Pandey, Anil K., Himal K. Ganguly, Sudipta Kumar Sinha, Kelly E. Daniels, Glenn P. A. Yap, Sandeep Patel, and Neal J. Zondlo. 2023. “An Inherent Difference between Serine and Threonine Phosphorylation: Phosphothreonine Strongly Prefers a Highly Ordered, Compact, Cyclic Conformation.” *ACS Chemical Biology* 18 (9): 1938–58. https://doi.org/10.1021/acschembio.3c00068.

Pauling, L., R. B. Corey, and H. R. Branson. 1951. “The Structure of Proteins; Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain.” *Proceedings of the National Academy of Sciences of the United States of America* 37 (4): 205–11. https://doi.org/10.1073/pnas.37.4.205.

Pavlović-Lažetić, Gordana M., Nenad S. Mitić, Jovana J. Kovačević, Zoran Obradović, Saša N. Malkov, and Miloš V. Beljanski. 2011. “Bioinformatics Analysis of Disordered Proteins in Prokaryotes.” *BMC Bioinformatics* 12 (1): 66. https://doi.org/10.1186/1471-2105-12-66.

Pearce, Robin, Xiaoqiang Huang, Dani Setiawan, and Yang Zhang. 2019. “EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function.” *Journal of Molecular Biology* 431 (13): 2467–76. https://doi.org/10.1016/j.jmb.2019.02.028.

Peng, Zhenling, Jing Yan, Xiao Fan, Marcin J. Mizianty, Bin Xue, Kui Wang, Gang Hu, Vladimir N. Uversky, and Lukasz Kurgan. 2015. “Exceptionally Abundant Exceptions: Comprehensive Characterization of Intrinsic Disorder in All Domains of Life.” *Cellular and Molecular Life Sciences* 72 (1): 137–51. https://doi.org/10.1007/s00018-014-1661-9.

Platzer, Gerald, Mark Okon, and Lawrence P. McIntosh. 2014. “pH-Dependent Random Coil 1H, 13C, and 15N Chemical Shifts of the Ionizable Amino Acids: A Guide for Protein pKa Measurements.” *Journal of Biomolecular NMR* 60 (2): 109–29. https://doi.org/10.1007/s10858-014-9862-y.

Plotkin, Joshua B., and Grzegorz Kudla. 2011. “Synonymous but Not the Same: The Causes and Consequences of Codon Bias.” *Nature Reviews Genetics* 12 (1): 32–42. https://doi.org/10.1038/nrg2899.

Privalov, Peter L. 1990. “Cold Denaturation of Protein.” *Critical Reviews in Biochemistry and Molecular Biology* 25 (4): 281–306. https://doi.org/10.3109/10409239009090612.

Rahman, M. Mahafuzur, and Christofer Lendel. 2021. “Extracellular Protein Components of Amyloid Plaques and Their Roles in Alzheimer’s Disease Pathology.” *Molecular Neurodegeneration* 16 (1): 59. https://doi.org/10.1186/s13024-021-00465-0.

Rajapaksha, Prasangi, Ankit Pandeya, and Yinan Wei. 2020. “Probing the Dynamics of AcrB Through Disulfide Bond Formation.” *ACS Omega* 5 (34): 21844–52. https://doi.org/10.1021/acsomega.0c02921.

Ramírez-Palma, Lillian G., Adrián Espinoza-Guillén, Fabiola Nieto-Camacho, Alexis E. López-Guerra, Virginia Gómez-Vidales, Fernando Cortés-Guzmán, and Lena Ruiz-Azuara. 2021. “Intermediate Detection in the Casiopeina–Cysteine Interaction Ending in the Disulfide Bond Formation and Copper Reduction.” *Molecules* 26 (19): 5729. https://doi.org/10.3390/molecules26195729.

Randolph, Lois, George Parra, and David Libich. 2021. “Understanding the Effects of Phosphorylation on the Structural Dynamics of EWS-FLI1 and Its Self-Associative Behavior.” *The FASEB Journal* 35 (S1). https://doi.org/10.1096/fasebj.2021.35.S1.04950.

Rangarajan, Nivedita, Prakash Kulkarni, and Sridhar Hannenhalli. 2015. “Evolutionarily Conserved Network Properties of Intrinsically Disordered Proteins.” *PLoS ONE* 10 (5): e0126729. https://doi.org/10.1371/journal.pone.0126729.

Rathi, PrakashChandra, Hans Wolfgang Höffken, and Holger Gohlke. 2014. “Quality Matters: Extension of Clusters of Residues with Good Hydrophobic Contacts Stabilize (Hyper)Thermophilic Proteins.” *Journal of Chemical Information and Modeling* 54 (2): 355–61. https://doi.org/10.1021/ci400568c.

Rehman, Saima, Lubov S. Grigoryeva, Katherine H. Richardson, Paula Corsini, Richard C. White, Rosie Shaw, Theo J. Portlock, et al. 2020. “Structure and Functional Analysis of the Legionella Pneumophila Chitinase ChiA Reveals a Novel Mechanism of Metal-Dependent Mucin Degradation.” *PLoS Pathogens* 16 (5): e1008342. https://doi.org/10.1371/journal.ppat.1008342.

Rohl, Carol A., and Andrew J. Doig. 1996. “Models for the 310-helix/Coil, Π-helix/Coil, and A-helix/310-helix/Coil Transitions in Isolated Peptides.” *Protein Science* 5 (8): 1687–96.

Rosano, Germán L., and Eduardo A. Ceccarelli. 2014. “Recombinant Protein Expression in Escherichia Coli: Advances and Challenges.” *Frontiers in Microbiology* 5 (April):172. https://doi.org/10.3389/fmicb.2014.00172.

Ross, P.D., and M.V. Rekharsky. 1996. “Thermodynamics of Hydrogen Bond and Hydrophobic Interactions in Cyclodextrin Complexes.” *Biophysical Journal* 71 (4): 2144–54. https://doi.org/10.1016/S0006-3495(96)79415-8.

Rost, B., C. Sander, and R. Schneider. 1994. “PHD--an Automatic Mail Server for Protein Secondary Structure Prediction.” *Computer Applications in the Biosciences: CABIOS* 10 (1): 53–60. https://doi.org/10.1093/bioinformatics/10.1.53.

Russo, Anna, Sara La Manna, Ettore Novellino, Anna Maria Malfitano, and Daniela Marasco. 2016. “Molecular Signaling Involving Intrinsically Disordered Proteins in Prostate

Cancer." *Asian Journal of Andrology* 18 (5): 673–81. https://doi.org/10.4103/1008-682X.181817.
Sahdev, Sudhir, Sunil K. Khattar, and Kulvinder Singh Saini. 2008. "Production of Active Eukaryotic Proteins through Bacterial Expression Systems: A Review of the Existing Biotechnology Strategies." *Molecular and Cellular Biochemistry* 307 (1): 249–64. https://doi.org/10.1007/s11010-007-9603-6.
Sakahira, Hideki, Peter Breuer, Manajit K. Hayer-Hartl, and F. Ulrich Hartl. 2002. "Molecular Chaperones as Modulators of Polyglutamine Protein Aggregation and Toxicity." *Proceedings of the National Academy of Sciences* 99 (suppl_4): 16412–18. https://doi.org/10.1073/pnas.182426899.
Sali, A., and T. L. Blundell. 1993. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* 234 (3): 779–815. https://doi.org/10.1006/jmbi.1993.1626.
Samiotakis, Antonios, and Margaret S. Cheung. 2011. "Folding Dynamics of Trp-Cage in the Presence of Chemical Interference and Macromolecular Crowding. I." *The Journal of Chemical Physics* 135 (17): 175101. https://doi.org/10.1063/1.3656691.
Sanger, F. 1952. "The Arrangement of Amino Acids in Proteins." In *Advances in Protein Chemistry*, edited by M. L. Anson, Kenneth Bailey, and John T. Edsall, 7:1–67. Academic Press. https://doi.org/10.1016/S0065-3233(08)60017-0.
Santofimia-Castaño, Patricia, Bruno Rizzuti, Ángel L. Pey, Philippe Soubeyran, Miguel Vidal, Raúl Urrutia, Juan L. Iovanna, and José L. Neira. 2017. "Intrinsically Disordered Chromatin Protein NUPR1 Binds to the C-Terminal Region of Polycomb RING1B." *Proceedings of the National Academy of Sciences of the United States of America* 114 (31): E6332–41. https://doi.org/10.1073/pnas.1619932114.
Santofimia-Castaño, Patricia, Bruno Rizzuti, Yi Xia, Olga Abian, Ling Peng, Adrián Velázquez-Campoy, José L. Neira, and Juan Iovanna. 2020. "Targeting Intrinsically Disordered Proteins Involved in Cancer." *Cellular and Molecular Life Sciences* 77 (9): 1695–1707. https://doi.org/10.1007/s00018-019-03347-3.
Sarkar, Tapati, Sukhen Das, Antara De, Papiya Nandy, Shiladitya Chattopadhyay, Mamta Chawla-Sarkar, and Ashesh Nandy. 2015. "H7N9 Influenza Outbreak in China 2013: In Silico Analyses of Conserved Segments of the Hemagglutinin as a Basis for the Selection of Peptide Vaccine Targets." *Computational Biology and Chemistry* 59 Pt A (December):8–15. https://doi.org/10.1016/j.compbiolchem.2015.08.003.
Sato, Yu, Kenji Okano, Hiroyuki Kimura, and Kohsuke Honda. 2020. "TEMPURA: Database of Growth TEMPeratures of Usual and RAre Prokaryotes." *Microbes and Environments* 35 (3). https://doi.org/10.1264/jsme2.ME20074.
Schlee, Sandra, and Joachim Reinstein. 2002. "The DnaK/ClpB Chaperone System from Thermus Thermophilus." *Cellular and Molecular Life Sciences CMLS* 59 (10): 1598–1606.
Schüler, Andreas, and Erich Bornberg-Bauer. 2016. "Evolution of Protein Domain Repeats in Metazoa." *Molecular Biology and Evolution* 33 (12): 3170–82. https://doi.org/10.1093/molbev/msw194.
Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, et al. 2020. "Improved Protein Structure Prediction Using Potentials from Deep Learning." *Nature* 577 (7792): 706–10. https://doi.org/10.1038/s41586-019-1923-7.

Seniya, Chandrabhan, Ghulam Jilani Khan, Richa Misra, Vaibhav Vyas, and Shruti Kaushik. 2014. "In-Silico Modelling and Identification of a Possible Inhibitor of H1N1 Virus." *Asian Pacific Journal of Tropical Disease* 4 (January):S467–76. https://doi.org/10.1016/S2222-1808(14)60492-8.

Shang, Jian, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. 2020. "Structural Basis of Receptor Recognition by SARS-CoV-2." *Nature* 581 (7807): 221–24. https://doi.org/10.1038/s41586-020-2179-y.

Shang, Yun, Dami Yang, Yunmi Ha, Ju Yeon Lee, Jin Young Kim, Man-Ho Oh, and Kyoung Hee Nam. 2021. "Open Stomata 1 Exhibits Dual Serine/Threonine and Tyrosine Kinase Activity in Regulating Abscisic Acid Signaling." *Journal of Experimental Botany* 72 (15): 5494–5507. https://doi.org/10.1093/jxb/erab225.

Shao, Hui, Wenmin Huang, Luisana Avilan, Véronique Receveur-Bréchot, Carine Puppo, Rémy Puppo, Régine Lebrun, Brigitte Gontero, and Hélène Launay. 2021. "A New Type of Flexible CP12 Protein in the Marine Diatom Thalassiosira Pseudonana." *Cell Communication and Signaling* 19 (1): 38. https://doi.org/10.1186/s12964-021-00718-x.

Sharir-Ivry, Avital, and Yu Xia. 2017. "The Impact of Native State Switching on Protein Sequence Evolution." *Molecular Biology and Evolution* 34 (6): 1378–90. https://doi.org/10.1093/molbev/msx071.

Smolarczyk, Tomasz, Irena Roterman-Konieczna, and Katarzyna Stapor. n.d. "Protein Secondary Structure Prediction: A Review of Progress and Directions." *Current Bioinformatics* 15 (2): 90–107.

Socci, Nicholas D., José Nelson Onuchic, and Peter G. Wolynes. 1998. "Protein Folding Mechanisms and the Multidimensional Folding Funnel." *Proteins: Structure, Function, and Bioinformatics* 32 (2): 136–58. https://doi.org/10.1002/(SICI)1097-0134(19980801)32:2<136::AID-PROT2>3.0.CO;2-J.

Söding, Johannes, Andreas Biegert, and Andrei N. Lupas. 2005. "The HHpred Interactive Server for Protein Homology Detection and Structure Prediction." *Nucleic Acids Research* 33 (Web Server issue): W244-248. https://doi.org/10.1093/nar/gki408.

Sørensen, Hans Peter, and Kim Kusk Mortensen. 2005. "Advanced Genetic Strategies for Recombinant Protein Expression in Escherichia Coli." *Journal of Biotechnology* 115 (2): 113–28. https://doi.org/10.1016/j.jbiotec.2004.08.004.

Soulages, Jose L., Kangmin Kim, Christina Walters, and John C. Cushman. 2002. "Temperature-Induced Extended Helix/Random Coil Transitions in a Group 1 Late Embryogenesis-Abundant Protein from Soybean." *Plant Physiology* 128 (3): 822–32.

Stadler, Andreas M., Franz Demmel, Jacques Ollivier, and Tilo Seydel. 2016. "Picosecond to Nanosecond Dynamics Provide a Source of Conformational Entropy for Protein Folding." *Physical Chemistry Chemical Physics* 18 (31): 21527–38. https://doi.org/10.1039/C6CP04146A.

Sun, Peter D., Christine E. Foster, and Jeffrey C. Boyington. 2004. "Overview of Protein Structural and Functional Folds." *Current Protocols in Protein Science* 35 (1). https://doi.org/10.1002/0471140864.ps1701s35.

Teoh, C. P., P. Lavin, N. A. Yusof, M. González-Aravena, N. Najimudin, Y. K. Cheah, and C. M. V. L. Wong. 2023. "Transcriptomics Analysis Provides Insights into the Heat Adaptation Strategies of an Antarctic Bacterium, Cryobacterium Sp. SO1." *Polar Biology* 46 (3): 185–97. https://doi.org/10.1007/s00300-023-03115-x.

Terwilliger, Thomas C., David Stuart, and Shigeyuki Yokoyama. 2009. “Lessons from Structural Genomics.” *Annual Review of Biophysics* 38:371–83. https://doi.org/10.1146/annurev.biophys.050708.133740.

Teufl, Magdalena, Charlotte U. Zajc, and Michael W. Traxlmayr. 2022. “Engineering Strategies to Overcome the Stability–Function Trade-Off in Proteins.” *ACS Synthetic Biology* 11 (3): 1030–39. https://doi.org/10.1021/acssynbio.1c00512.

The UniProt Consortium. 2021. “UniProt: The Universal Protein Knowledgebase in 2021.” *Nucleic Acids Research* 49 (D1): D480–89. https://doi.org/10.1093/nar/gkaa1100.

Thole, Joseph F., Christopher A. Waudby, and Gary J. Pielak. 2023. “Disordered Proteins Mitigate the Temperature Dependence of Site-Specific Binding Free Energies.” *The Journal of Biological Chemistry* 299 (3): 102984. https://doi.org/10.1016/j.jbc.2023.102984.

Thompson, J D, D G Higgins, and T J Gibson. 1994. “CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.” *Nucleic Acids Research* 22 (22): 4673–80.

Torrisi, Mirko, Manaz Kaleel, and Gianluca Pollastri. 2018. “Porter 5: Fast, State-of-the-Art Ab Initio Prediction of Protein Secondary Structure in 3 and 8 Classes.” *bioRxiv*, October, 289033. https://doi.org/10.1101/289033.

Urhan, Aysun, Bianca-Maria Cosma, Ashlee M. Earl, Abigail L. Manson, and Thomas Abeel. 2023. “SAP: Synteny-Aware Gene Function Prediction for Bacteria Using Protein Embeddings.” *bioRxiv*, November, 2023.05.02.539034. https://doi.org/10.1101/2023.05.02.539034.

Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. “Why Are ‘Natively Unfolded’ Proteins Unstructured under Physiologic Conditions?” *Proteins* 41 (3): 415–27. https://doi.org/10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7.

Uversky, Vladimir N. 2011. “Intrinsically Disordered Proteins from A to Z.” *The International Journal of Biochemistry & Cell Biology* 43 (8): 1090–1103. https://doi.org/10.1016/j.biocel.2011.04.001.

Vella, F. 1992. “Introduction to Protein Structure: By C Branden and J Tooze. Pp 302. Garland Publishing, New York. 1991.” *Biochemical Education* 20 (2): 122. https://doi.org/10.1016/0307-4412(92)90132-6.

Vogt, Gerhard, and Patrick Argos. 1997. “Protein Thermal Stability: Hydrogen Bonds or Internal Packing?” *Folding and Design* 2 (June):S40–46. https://doi.org/10.1016/S1359-0278(97)00062-X.

Wang, Juan, Kai Liu, Ruirui Xing, and Xuehai Yan. 2016. “Peptide Self-Assembly: Thermodynamics and Kinetics.” *Chemical Society Reviews* 45 (20): 5589–5604. https://doi.org/10.1039/C6CS00176A.

Wang, Sheng, Jian Peng, Jianzhu Ma, and Jinbo Xu. 2016. “Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields.” *Scientific Reports* 6 (January):18962. https://doi.org/10.1038/srep18962.

Wang, Zhiyong, Feng Zhao, Jian Peng, and Jinbo Xu. 2010. “Protein 8-Class Secondary Structure Prediction Using Conditional Neural Fields.” In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 109–14. https://doi.org/10.1109/BIBM.2010.5706547.

Waterhouse, Andrew, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, et al. 2018. “SWISS-MODEL: Homology Modelling of Protein Structures and Complexes.” *Nucleic Acids Research* 46 (Web Server issue): W296–303. https://doi.org/10.1093/nar/gky427.

Wennerström, Håkan, and Mikael Oliveberg. 2022. “On the Osmotic Pressure of Cells.” *QRB Discovery* 3 (July):e12. https://doi.org/10.1017/qrd.2022.3.

Westphal, Adrie H., Astrid A. Geerke-Volmer, Carlo P. M. van Mierlo, and Willem J. H. van Berkel. 2017. “Chaotropic Heat Treatment Resolves Native-like Aggregation of a Heterologously Produced Hyperthermostable Laminarinase.” *Biotechnology Journal* 12 (6): 1700007. https://doi.org/10.1002/biot.201700007.

Wetzler, Diana E., Federico Fuchs Wightman, Hernan A. Bucci, Jimena Rinaldi, Julio J. Caramelo, Norberto D. Iusem, and Martiniano M. Ricardi. 2018. “Conformational Plasticity of the Intrinsically Disordered Protein ASR1 Modulates Its Function as a Drought Stress-Responsive Gene.” *PLoS ONE* 13 (8): e0202808. https://doi.org/10.1371/journal.pone.0202808.

Whisstock, James C, and Arthur M Lesk. 2003. “Prediction of Protein Function from Protein Sequence and Structure.” *Quarterly Reviews of Biophysics* 36 (3): 307–40.

Wibmer, Constantinos Kurt, Frances Ayres, Tandile Hermanus, Mashudu Madzivhandila, Prudence Kgagudi, Brent Oosthuysen, Bronwen E. Lambson, et al. 2021. “SARS-CoV-2 501Y.V2 Escapes Neutralization by South African COVID-19 Donor Plasma.” *Nature Medicine* 27 (4): 622–25. https://doi.org/10.1038/s41591-021-01285-x.

Winklhofer, Konstanze F, Jörg Tatzelt, and Christian Haass. 2008. “The Two Faces of Protein Misfolding: Gain- and Loss-of-Function in Neurodegenerative Diseases.” *The EMBO Journal* 27 (2): 336–49. https://doi.org/10.1038/sj.emboj.7601930.

Wood, EJ. 1996. “Structure in Protein Chemistry: By J Kyte. Pp 606. Garland Publishing, New York & London. 1995. $62 ISBN 0-8153-1701-8.” *Biochemical Education* 24 (1): 68–69. https://doi.org/10.1016/S0307-4412(96)80028-8.

Wood, Jeanette M, Jürgen Maibaum, Joseph Rahuel, Markus G Grütter, Nissim-Claude Cohen, Vittorio Rasetti, Heinrich Rüger, Richard Göschke, Stefan Stutz, and Walter Fuhrer. 2003. “Structure-Based Design of Aliskiren, a Novel Orally Effective Renin Inhibitor.” *Biochemical and Biophysical Research Communications* 308 (4): 698–705.

Wright, Peter E., and H. Jane Dyson. 2015. “Intrinsically Disordered Proteins in Cellular Signalling and Regulation.” *Nature Reviews Molecular Cell Biology* 16 (1): 18–29. https://doi.org/10.1038/nrm3920.

Wu, Sitao, and Yang Zhang. 2007. “LOMETS: A Local Meta-Threading-Server for Protein Structure Prediction.” *Nucleic Acids Research* 35 (10): 3375–82. https://doi.org/10.1093/nar/gkm251.

Wu, Yi I., Daniel Frey, Oana I. Lungu, Angelika Jaehrig, Ilme Schlichting, Brian Kuhlman, and Klaus M. Hahn. 2009. “A Genetically Encoded Photoactivatable Rac Controls the Motility of Living Cells.” *Nature* 461 (7260): 104–8. https://doi.org/10.1038/nature08241.

Xia, Xuhua. 2018. “DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution.” *Molecular Biology and Evolution* 35 (6): 1550–52. https://doi.org/10.1093/molbev/msy073.

———. 2020. “Phylogeny-Based Comparative Methods.” In *A Mathematical Primer of Molecular Phylogenetics*. Apple Academic Press.

Xie, Hongbo, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Vladimir N. Uversky, and Zoran Obradovic. 2007. "Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions." *Journal of Proteome Research* 6 (5): 1882–98. https://doi.org/10.1021/pr060392u.

Xiong, Youling L. 1997. "Protein Denaturation and Functionality Losses." In *Quality in Frozen Food*, edited by Marilyn C. Erickson and Yen-Con Hung, 111–40. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-5975-7_8.

Xue, Bin, Roland L. Dunbrack, Robert W. Williams, A. Keith Dunker, and Vladimir N. Uversky. 2010. "PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids." *Biochimica et Biophysica Acta* 1804 (4): 996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011.

Yabukarski, Filip, Justin T. Biel, Margaux M. Pinney, Tzanko Doukov, Alexander S. Powers, James S. Fraser, and Daniel Herschlag. 2020. "Assessment of Enzyme Active Site Positioning and Tests of Catalytic Mechanisms through X-Ray–Derived Conformational Ensembles." *Proceedings of the National Academy of Sciences* 117 (52): 33204–15. https://doi.org/10.1073/pnas.2011350117.

Yampolsky, Lev Y., and Arlin Stoltzfus. 2005. "The Exchangeability of Amino Acids in Proteins." *Genetics* 170 (4): 1459–72. https://doi.org/10.1534/genetics.104.039107.

Yan, Jing, Marcin J. Mizianty, Paul L. Filipow, Vladimir N. Uversky, and Lukasz Kurgan. 2013. "RAPID: Fast and Accurate Sequence-Based Prediction of Intrinsic Disorder Content on Proteomic Scale." *Biochimica Et Biophysica Acta* 1834 (8): 1671–80. https://doi.org/10.1016/j.bbapap.2013.05.022.

Yang, Jianyi, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. 2020. "Improved Protein Structure Prediction Using Predicted Interresidue Orientations." *Proceedings of the National Academy of Sciences of the United States of America* 117 (3): 1496–1503. https://doi.org/10.1073/pnas.1914677117.

Yang, Jianyi, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. 2015. "The I-TASSER Suite: Protein Structure and Function Prediction." *Nature Methods* 12 (1): 7.

Yang, Yuedong, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. 2018. "Sixty-Five Years of the Long March in Protein Secondary Structure Prediction: The Final Stretch?" *Briefings in Bioinformatics* 19 (3): 482–94. https://doi.org/10.1093/bib/bbw129.

Yi, T. M., and E. S. Lander. 1993. "Protein Secondary Structure Prediction Using Nearest-Neighbor Methods." *Journal of Molecular Biology* 232 (4): 1117–29. https://doi.org/10.1006/jmbi.1993.1464.

Yoon, Mi-Kyung, Diana M. Mitrea, Li Ou, and Richard W. Kriwacki. 2012. "Cell Cycle Regulation by the Intrinsically Disordered Proteins P21 and P27." *Biochemical Society Transactions* 40 (5): 981–88. https://doi.org/10.1042/BST20120092.

Yuan, Xin, and Christopher Bystroff. 2007. "Protein Contact Map Prediction." In *Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization*, edited by Ying Xu, Dong Xu, and Jie Liang, 255–77. BIOLOGICAL AND MEDICAL PHYSICS BIOMEDICAL ENGINEERING. New York, NY: Springer. https://doi.org/10.1007/978-0-387-68372-0_8.

Zacchi, Lucía F., Hui-Chuan Wu, Samantha L. Bell, Linda Millen, Adrienne W. Paton, James C. Paton, Philip J. Thomas, Michal Zolkiewski, and Jeffrey L. Brodsky. 2014. "The BiP Molecular Chaperone Plays Multiple Roles during the Biogenesis of TorsinA, an AAA+ ATPase Associated with the Neurological Disease Early-Onset Torsion Dystonia*." *Journal of Biological Chemistry* 289 (18): 12727–47. https://doi.org/10.1074/jbc.M113.529123.

Zhang, Buzhong, Jinyan Li, and Qiang Lü. 2018. "Prediction of 8-State Protein Secondary Structures by a Novel Deep Learning Architecture." *BMC Bioinformatics* 19 (1): 293. https://doi.org/10.1186/s12859-018-2280-5.

Zhang, Qiangfeng Cliff, Donald Petrey, Lei Deng, Li Qiang, Yu Shi, Chan Aye Thu, Brygida Bisikirska, Celine Lefebvre, Domenico Accili, and Tony Hunter. 2012. "Structure-Based Prediction of Protein–Protein Interactions on a Genome-Wide Scale." *Nature* 490 (7421): 556.

Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. "A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin." *Nature* 579 (7798): 270–73. https://doi.org/10.1038/s41586-020-2012-7.

Zylicz-Stachula, Agnieszka, Olga Zolnierkiewicz, Katarzyna Sliwinska, Joanna Jezewska-Frackowiak, and Piotr M. Skowron. 2014. "Modified 'One Amino Acid-One Codon' Engineering of High GC Content TaqII-Coding Gene from Thermophilic Thermus Aquaticus Results in Radical Expression Increase." *Microbial Cell Factories* 13 (1): 7. https://doi.org/10.1186/1475-2859-13-7.