Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Analysis and visualization of multi-modal protein distributions for structural proteome feature identification

Isabell Julia Orlishausen

Bioinformatics Program

Technical University of Munich

Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Analysis and visualization of multi-modal protein distributions for structural proteome feature identification

Analyse und Visualisierung multimodaler Proteinverteilungen zur Identifizierung struktureller Eigenschaften von Proteomen

Author: Isabell Julia Orlishausen
Supervisor: Prof. Dr. Burkhard Rost
Advisor: Tobias Olenyi, M.Sc.
Dr. Maria Littmann
Submitted: 15.05.2022

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 15.05.2022 Isabell Orlishausen

Acknowledgments

Thank you, Tobi, for using all the statistics lectures you attended to introduce me to these foreign concepts. Thank you for always being open for any questions, for the nice atmosphere in our meetings, for giving me the chance to further develop `ppprint`, and most importantly, thank you for repeatedly telling me to not stress.
Thank you, Maria, for lots of helpful advice in several situations and for inspiring me to write my thesis at ROSTLAB many months ago.
Thank you, Michael, for providing scripts for interfacing PredictProtein.
Thanks to my preferred "rubber duck" for being exactly that and the continuous support in many forms.
Thank you, Johannes, for advice and inspiration for my graphics and writing.
And finally, thank you, Maria, Tobi, and everyone involved in proof-reading parts of my thesis.

Abstract

Prediction of structure and other properties of proteins has reached accuracies enabling the omission of costly and time-expensive experiments. A multitude of tools specialized for the prediction of a single protein feature is combined in the meta-resource PredictProtein. We applied PredictProtein tools for the prediction of protein disorder, transmembrane helices, secondary structure, and binding sites for the complete Swiss-Prot proteomes of nine sample organisms. Comparing these organisms in the analysis of the whole-proteome distribution of each feature, we made several feature-specific observations and evaluated them against the backdrop of existing research. In most aspects, prokaryotic and eukaryotic organisms differed vastly from each other, while *Saccharomyces cerevisiae* showed varying similarity to both kingdoms. Supplementing our analysis, we provide forms of visualization for each investigated concept. PredictProtein so far does not offer the option of calculating and visualizing statistics when predicting for a collection of proteins, such as whole proteomes. Thus, we used our analysis scripts to develop `ppprint` as an extension for PredictProtein. Despite still being in development, `ppprint` can analyze PredictProtein prediction data for multiple proteomes and display the visualized comparison results in feature-specific dashboards.

Zusammenfassung

Die Vorhersage der Struktur und anderer Eigenschaften von Proteinen hat ein Level an Genauigkeit erreicht, dass das Aussparen von kostspieligen und zeitintensiven Experimenten ermöglicht. Eine Vielzahl an Vorhersage-Algorithmen, die sich auf die Prognose einer bestimmten Proteineigenschaft spezialisieren, wird in der Meta-Resource PredictProtein kombiniert. Wir haben PredictProtein Programme für die Vorhersage von Unordnung in Proteinen, Transmembranhelices, Proteinsekundärstruktur und Bindestellen von vollständigen Swiss-Prot Proteomen von neun Beispiel-Organismen angewandt. Beim Vergleich dieser Organismen innerhalb der Analyse von Verteilungen jeder Eigenschaft über ganze Proteome hinweg machten wir mehrere eigenschaftsspezifische Beobachtungen und evaluierten diese vor dem Hintergrund existierender Forschung. In den meisten Vergleichen unterschieden sich prokaryotische und eukaryotische Organismen stark voneinander, während *Saccharomyces cerevisiae* varierende Ähnlichkeit zu beiden Reichen zeigte. Ergänzend zu unserer Analyse stellen wir Visualiserungen for jedes untersuchte Konzept bereit. Bis zum jetzigen Zeitpunkt bietet PredictProtein nicht die Option, bei der Vorhersage einer Sammlung von Proteinen wie eines ganzen Proteoms Statistiken zu berechnen und zu visualisieren. Somit benutzten wir unsere Analyseskripte um `ppprint` als Erweiterung für PredictProtein zu entwickeln. Wenn auch noch sich in Entwicklung befindend kann `ppprint` PredictProtein Vorhersagedaten für mehrere Proteome analysieren und die visualisierten Ergebnisse des Vergleichs in eigenschaftsspezifischen Dashboards darstellen.

Contents

Part I.

Thesis

1. Introduction and Motivation

1.1. Background

One successful strategy to improve health in the history of humanity is designing drugs, which inhibit or repair malfunctioning biochemical pathways. Hence, past and current research aims to comprehend the underlying biophysical processes. Since proteins are essential to life, understanding their function is the key step towards this goal.

Protein function is widely determined by structure. Using X-Ray crystallography or nuclear magnetic resonance (NMR) spectroscopy to solve the structure of a protein requires years of work by highly-skilled specialists that operate expensive equipment. Thus, intensive research towards omitting the costly [1] experimental approach by predicting protein structure computationally has been done for decades. Recently, huge advances have been made in the field of protein structure prediction, with methods making successful predictions for more than 98 % of the human proteome [2]. However, in order to make statements regarding characteristics of proteins and even whole proteomes, exact three-dimensional structural predictions are not needed. Even before the achievement of high accuracy scores, powerful tools to predict various protein features were already available.

PredictProtein [3] is a collection of multiple specialized prediction tools available online. Most of these make use of evolutionary information and machine learning approaches. Recently, deep learning embeddings have been replacing the time-expensive multiple sequence alignment (MSA) step to predict structural and functional characteristics of proteins. Using only the protein sequence, the web-based resource is intuitive to use and offers a wealth of feature information corresponding to regions of the given sequence. Yet, any available form of analysis and visualization in PredictProtein is limited to the level of single protein predictions. When expanding the available data to whole proteomes, the distribution of protein features offers a tangible way of comparing organisms, which we explored in this thesis.

1.2. Related Work

Analysis and Comparison of Multiple Proteomes

Previously, work focused on a single protein feature [4]. Marot-Lassauzaie *et al.* restricted their in-depth analysis to the subcellular localization of proteins. They selected 10 model eukaryotes and used two prediction methods, LocTree3 [5] being part of PredictProtein, to compute their location spectrum. As a result of their analysis, they discovered that the proteome fractions in seven specified major cellular compartments do not show striking variation among the model organisms. However, the information containted in the fraction distribution sufficed to reflect established evolutionary relationships of the organisms. Two provided forms of visualization, one of them being a phylogenetic tree built from the location spectrum, demonstrate their findings on subcellular localization. We decided to broaden our approach and examine the proteome-wide distributions of multiple protein

features, while rather focusing on the aspect of visualization by providing several plots per feature.

Another method that is part of the PredictProtein collection is Meta-Disorder (MD) [6], a tool predicting regions of disorder in proteins. Since we also investigated aspects of disorder in our analysis, a detailed explanation of this protein feature will be given in the next chapter. Following MD, Schlessinger *et al.* used their method and others to further investigate protein disorder [7]. They focused on characterizing this novel feature in order to establish views supported by their analysis, which can be accepted for the time being. Thereby, they underlined the relevance of disorder as well as the proteome-wide study of a protein feature in general. In their research, they included a between-kingdoms comparison of aspects of disorder such as disorder composition. We reviewed the stated hypotheses and performed similar analyses for further properties regarding protein disorder. Yet, it was not our intention to define the protein feature, but rather to compare proteomes of different organisms while building on existing concepts. Here, we deemed it especially important to provide useful forms of visualization for each of the investigated properties of disorder and other protein features.

One way of functionally capturing the complexity of an organism is by defining it as the number of cell types it comprises. For a variety of organisms and tissues, Schad *et al.* extracted complexity values from a range of 1 for prokaryotes to 169 for *Homo sapiens* from literature [8]. A widely accepted view states that the number of genes generally does not correlate with the complexity of an organism [9]. In order to find other qualities of an organism that contribute to its complexity level, Schad *et al.* reviewed properties of proteins such as structural disorder by relating them to complexity. For their analysis, they used a mixture of direct predictions and data extracted from feature-specific databases. They found that several proteome traits, including the fraction of disorder or proteome size, somewhat correlate with complexity and differentiate the prokaryotic and eukaryotic kingdoms well. We likewise examined multiple protein features and their distribution on a per-proteome basis. While we also tried to find aspects of correlation within our data, we did not focus on investigating complexity. Thus, we developed a variety of forms of visualization instead of limiting the contained information to the display in scatter plots.

Mészáros *et al.* also explored protein disorder during the development of their prediction method [10]. ANCHOR makes use of properties of the general disorder prediction method IUPred [11] to predict disordered binding sites, short segments regulating the formation of a complex of a disordered protein and its interacting partner. These disordered region parts cannot form enough favorable intrachain interactions to maintain a well-structured form, but are able to gain energy through interacting with an external partner protein. Thereby, the work of Mészáros *et al.* combines two protein features: disorder and binding. Additional to the tool development, they also applied ANCHOR on a wide selection of Swiss-Prot proteomes across all kingdoms. They likewise provided multiple plots visualizing their findings, but, as they focused on developing ANCHOR, they restricted their work to the particular feature combination of disorder and binding. Instead of deepening our research on one single property of a protein, we performed our analysis for multiple protein features and used established PredictProtein tools to generate prediction data. However, we likewise deemed the combination of protein binding and disorder important to explore and compared our insights to their results using ANCHOR predictions.

Another approach for combining protein features is studying single-pass transmembrane proteins with respect to their subcellular localization [12]. Previously, it has been known that proteins targeted to the plasma membrane on average contain longer transmembrane

helices (TMHs) than proteins of cellular organelles. Singh & Mittal report that with increasing evolutionary complexity, the deviation in lengths of TMHs in different membrane types lessens. Instead of using predictions, they restricted their research to well-studied eukaryotic proteomes with known transmembrane domains and localizations. For our analysis, we made use of computationally generated protein feature data and could thus include more organisms in our comparison. Since we did not explore subcellular localizations of proteins, we could not review the claims made by Singh & Mittal. However, we likewise investigated the distribution of TMH lengths for an overlapping set of selected model organisms.

Tools for Visualization

A common shortcoming of the in-depth proteome analyses that were mentioned above is that despite great potential, the option for others to easily make use of the presented analysis methods for own research is not made available. Tools for proteomic prediction data are rare: Most available methods do not offer the comparison of multiple proteomes [13], only allow the display of feature information for individual proteins [3, 14] or are not available online [14]. ProteomeVis [15] is a web application for the proteome-wide study of protein structure and sequence evolution for multiple organisms. However, it does not use feature prediction but provides its own, manually curated data extracted from published papers and databases such as the Protein Data Bank (PDB) [16]. While focusing on graph-theoretical approaches, ProteomeVis enables the exploration of several proteome properties and metrics other than the distribution of protein features as mentioned above. For instance, it does not include protein disorder, since PDB does not feature proteins without a stable structure. Its dashboard-like design allows researchers lacking programmatic experience to perform proteome-wide analysis, yet without enabling a direct comparison between proteomes. Being restricted to two pre-loaded proteomes without the possibility for users to upload their own data, the method clearly has its drawbacks. To the best of our knowledge, a tool offering a dashboard-like comparison of multiple proteomes based on the distributions of several protein features, while enabling the incorporation of user-uploaded data and thus overcoming dependencies on experimental PDB data is yet to be published.

1.3. Overview

For a comparison on the level of proteomes, we first describe the generation of Predict-Protein prediction data and the selected combination of protein features. These include biophysical qualities such as disorder, transmembrane content and binding sites. Subsequent chapters analyze and extensively visualize extracted information in order to disclose properties hidden in the proteome feature distributions. The resulting web application `ppprint`, which will be introduced in the third chapter, enables researchers without data science expertise to likewise compare organisms by their proteomes. Derived insights into how single organisms or groups behave regarding crucial protein features can then enhance knowledge on their relationship and thus aid in research and drug development.

2. Material and Methods

A species can be characterized by its proteome. We have applied this concept in order to gain further insight into the relationships of different species. This chapter will describe the methodological principles of our comparison of several organisms on the basis of their proteomes. The first section will introduce how we used proteome sequencing data to predict several per-protein features. This was performed for multiple organisms, including *Homo sapiens*. We then analyzed and compared the distribution of these protein features on proteome level. This section will also specify feature-specific concepts that we will reference when presenting our results. General methods we made use of within our approach will be presented in the second section of this chapter. Analysis scripts can be found at `https://git.rostlab.org/orlishausen/proteome-analysis`.

2.1. Proteome Sequence Data

In this thesis, we compared multiple organisms according to the distribution of protein features among their proteomes. This section will provide information regarding the raw proteome sequencing data of the chosen sample organisms.

We selected the proteome of *Homo sapiens* as well as of six other eukaryotes for comparison. Since the extent of differentiation between the eukaryotic proteomes was initially uncertain, we added two prokaryotes to the list of organisms to be compared.

We extracted the raw sequencing files from Swiss-Prot [17] on October 1, 2021. Instead of including the commputationally analyzed part of UniProt [18] (TrEMBL [19]), only the manually annotated proteins from Swiss-Prot were used for the prediction and any following analysis. Thus, we increased the reliability of subsequent results, justifying potential apparent incompleteness of proteomes in comparison to TrEMBL numbers. Table 2.1 shows the selected organisms, their UniProt proteome identifiers (UPIDs) and the number of proteins referenced to the respective proteome.

Scientific Name	Trivial Name	UPID	Proteome Size
Homo sapiens	Human	UP000005640	20,371
Mus musculus	Mouse	UP000000589	17,077
Arabidopsis thaliana	Mouse-ear cress	UP000006548	16,036
Rattus norvegicus	Rat	UP000002494	8,131
Saccharomyces cerevisiae	Baker's yeast	UP000002311	6,049
Caenorhabditis elegans	Worm	UP000001940	4,264
Bacillus subtilis	Gram-positive bacterium	UP000001570	4,191
Drosophila melanogaster	Fruit fly	UP000000803	3,614
Escherichia coli	Gram-negative bacterium	UP000000558	2,033

Table 2.1.: Organisms and their Swiss-Prot proteomes selected for comparison. Names of prokaryotes are given in bold font.

We chose mouse and rat as sample organisms because of their close evolutionary relationship to human. To further ensure comparability, we selected the eukaryotes cress, yeast, worm, and fruit fly as well as the prokaryote *E. coli* since they are widely used as model organisms in research. *B. subtilis* is a gram-positive bacterium with a proteome of 98% reviewed sequences, completing the set of nine sample organisms as the second prokaryote.

2.2. Selected Proteome Features

This section will give a brief description of each selected protein feature, including our reasoning behind utilizing the respective property for our analysis and how the prediction results were generated.

PredictProtein Predictions

In order to compare the chosen organisms after gathering the raw proteome sequences, we sought to find their characteristics in form of the distribution of features attributed to the comprised proteins. Omitting the time-consuming and expensive experimental approach, we used computational protein feature predictions generated by PredictProtein. Through the meta-resource we could easily access a variety of tools, which predict specific structural and functional features from the amino acid sequence. Thousands of citations and millions of visits [20] justified the applicability of PredictProtein [20]. In addition to the direct link to its contributors at ROSTLAB this made it the resource of choice for protein feature predictions.

We accessed a local installation of PredictProtein via the internal ROSTLAB cluster. A more detailed overview of how the cluster was interfaced is shown in Figure II.1 in the appendix.

PredictProtein offers tools for the prediction of numerous structural and functional features. Due to the pre-defined time frame of this thesis, we could only include a fraction of them in our comparison. Nonetheless, future work on the excluded properties can build on our analysis. The following part provides an overview of the features selected for this thesis. Reasons for including protein disorder, transmembrane helices, secondary structure and binding sites will be given in the respective subsections.

2.2.1. Protein Disorder

Protein disorder comes in many shapes, including not only loops in secondary structure, but also molten globule domains and more. As a sup-ordinate definition applying to all types of disorder, disordered regions (DRs) in proteins can be defined as a those regions, which adopt two potentially widely different three-dimensional (3D) structures in isolation, when observed at two different points in time. Simply put, they lack a stable 3D structure.

Research of the past one and a half decade has discovered that protein disorder is conserved in evolution. Since the flexibility of DRs could be utilized to regulate complex protein interactions, the level of disorder prevalent in an organism's proteome could hint at its biological complexity [7]. We wanted to investigate this and other hypotheses relating disorder to binding sites, another protein feature. In PredictProtein, multiple original tools are combined in Meta-Disorder (MD) [6] in order to predict disordered regions in proteins. Thus, we decided to include protein disorder in our set of features to base our

comparison on.

The definition of protein disorder covers multiple phenomena regarding lack of structure, including local flexible loops. Those disorder loops behave differently than shorter loops in non-regular secondary structure, with length being a key factor for machine learning methods distinguishing the two. Although there is no scientific reality for a particular threshold level, it is biophysically safe to discern regular flexible loop regions in well-structured proteins from loopy disordered at a length of 30 residues. Thus, we decided to exclude any predicted disordered regions (DRs) with a length smaller than 30 from our data, unless stated otherwise.

Composition Schlessinger *et al.* defined the disorder composition per proteome as the fraction of proteins comprising at least one DR of length 30 or more [7].

Proteome Disorder Content A similar measurement that, likewise to the composition, also assigns a single value to each proteome is the proteome content. In the context of disorder we define it as the sum of the number of residues in any DR of a proteome, divided by the total number of residues of all proteins in the proteome.

Disorder Diamonds and Spectra The concept of relative indices, which we will explain in detail in section 2.3, also allows to combine DR location and length information. The disorder diamond plot captures the mean relative start and end positions of all DRs centered at the respective y-value. The shape of the data naming the diamond plot is a trivial result from the definitions. As center positions move to the ends of the normalized protein ($y \in \{0.0, 1.0\}$), the absolute maximum of start and end positions, being relative to the center, decreases. For instance, a DR centered at the relative index of 0.9 can have a maximum total width of $0.1 + 0.1 = 0.2$, being restricted by the end of the protein. Disordered regions centered around 0.5 however can span the whole protein and thus reach a width of $0.5 + 0.5 = 1.0$, from protein start to end.

In order to be able to extract more findings from the diamond plot, we introduced the required comprehensibility by reducing redundancy. The naturally symmetric nature of the diamond plot can be viewed as distracting, which is why we decided to drop the start values. After rotating the remaining half of the diamond, the former y-axis became the new x-axis. We found that the bin means suffice for the visual comparison, since the background binning bars do not extensively add to the plot and can thus be regarded as redundant. This made it possible to plot the error bars as shaded intervals behind each curve, which connects the means and gives the newly formed spectrum plot its name.

2.2.2. Transmembrane Helices

Proteins located in the cellular membrane constitute a subgroup of proteins only making up less than a third of an organism's proteome [21]. Still, these proteins fulfill numerous tasks that are essential for the living cell. Their transmembrane nature facilitates specific binding by helping with ligand recognition, regulates transport of substances across the membrane and enables communication between cells. The majority of transmembrane proteins (TMPs) contain helices (TMHs) crossing the lipid bi-layer. Since 57% of drug targets are transmembrane proteins [22], research regarding the distribution of TMPs and their helices among different organisms is highly relevant. PredictProtein uses TMSEG [23] to detect TMPs, predict their topologies and locate the present TMHs. We decided to

further investigate transmembrane helices by including them in our set of features for the comparison of the selected sample organisms.

The potentially multiple parts of the protein that pass through the lipid bi-layer predominantly adopt helical structures. Naturally, these helices require a stretch of amino acids of a certain composition as well as minimum length in order to be formed. Studies have tried to determine a biologically safe length threshold [24]. To capture biological realities, we used 12 amino acids as a minimum length for these transmembrane helices (TMHs) in accordance with the predictions produced by TMSEG [23].

Signal peptides (SPs) are short stretches at the N-terminus of proteins that are used for protein targeting. Since these peptides usually get clipped prior to the insertion of the TMP into the membrane, we excluded these parts, predicted by TMSEG, from our data. After potential SP clipping and insertion of the protein, the TMP has a certain orientation depending on the location of the protein N-terminus. We also made use of these topological predictions by TMSEG.

Protein Classes Besides single-pass transmembrane proteins, TMPs can also be classified as "multi-pass" when comprising multiple membrane-spanning TMHs. All non-TMPs are classified as "globular" because of their folded conformation.

Transmebrane Content Per Protein For the analysis of protein classes, we partitioned the proteome of an organism into three classes depending on the number of transmembrane helices (TMHs) of the contained proteins. By incorporating this number when investigating TMPs, we defined a straightforward measurement on the basis of segments, in this case, TMHs. This bears the strong dependency on how the predictor defined present TMHs. For instance, a stretch of amino acids can be viewed as one or two TMHs, depending on whether the predictions by TMSEG did not or did classify a sub-sequence as an in-between loop. By that, a few amino acid residues decide whether to raise the count of the single-pass or multi-pass fraction. Here, we define the transmembrane content per TMP. Through measuring the content using prediction data on the level of residues instead of segments, we trade simplicity of whole TMHs for a decrease in dependency on exact predictions. Because of the low total share of TMPs per proteome, we decided to exclude globular proteins from this part of our analysis.

Orientation of Transmembrane Proteins After being inserted into the cellular membrane, the orientation of a TMP can be determined based on the location of its N-terminal end. Intuitively, we labeled TMPs with the N-terminus on the inner or outer side of the cell as "cytoplasmic" or "extracellular", respectively. TMPs that have a signal peptide or an amino acid sequence that begins with a transmembrane helix were classified as "Membrane".

Topological Fractions of Transmembrane Proteins Earlier in this section, we explained the reason for our analysis progressing from straightforward protein classes, relying on classifications based on segments, to TM content that we calculated using residue-level data. We likewise made use of per-residue topology data by analyzing the mean component fractions across all proteins for each sample proteome. While losing simplicity by advancing from segment to residue data, we simultaneously reduce the complexity of separate, explicit information regarding the number and lengths of TMHs by combining the two.

2.2.3. Secondary Structure

A protein can fold into its native shape even when it is isolated from the cell [25], implying that the amino acid sequence alone determines the structure of a protein. On a lower complexity level than estimating the 3D shape of a protein, this discovery also allows the prediction of secondary structure elements, using only the sequence. The structural elements as specified in the 8-state DSSP assignment [26] can be simplified into three groups: helix (H), strand (E) and other states (O).

By nature, protein disorder is a complex protein feature. It comprises several forms of disorder and, in comparison to other protein properties predicted by PredictProtein, can be viewed as a rather novel protein feature. Since transmembrane proteins only make up about a quarter of an organism's proteome, transmembrane helices constitute a feature that merely applies to parts of the sample data. In order to incorporate the whole proteome while exploring an established protein feature, we expanded our comparison to secondary structure since it is a relevant principal property of a protein. As we focused on investigating aspects of disorder, transmembrane helices and ligand binding, the respective sections regarding the investigation of secondary structure will serve as an overview of possible starting points for a more detailed analysis.

Being able to accurately predict secondary structure has been a key step on humanity's way to understanding how proteins fold and thus, decoding the function of a protein. Its relevance is supported by the recent inclusion of ProtT5-sec [27] into the PredictProtein service. ProtT5-sec belongs to a novel kind of secondary structure prediction methods that achieves high accuracies without using evolutionary information in the form of multiple sequence alignments. Since ProtT5-sec is not included in the local cluster installation of PredictProtein, we used RePROF [28] to predict whether an amino acid residue belongs to a H, E or O part of the protein. As a locally forming structural element, helices require a minimum of four consecutive compatible residues to be formed. Thus, we chose four amino acid residues as a minimum length for helical regions predicted by RePROF.

Structural Fractions As for topological regions of transmembrane proteins, we investigated the distribution of helical and strand residues detached from segment-wise definitions. Here, we define the helix, strand or "other" content of a protein as the number of helical, strand or "other" residues, respectively, divided by the protein length. Applying this definition, we calculated the mean levels of helix, strand and "other" content for each proteome.

2.2.4. Protein Binding

The vast majority of proteins are involved in processes of binding to other molecules during their lifespan [29]. Determining presence and position of potential protein, DNA, or RNA binding sites therefore is an integral part of understanding the function of proteins. The predominant purpose of drugs is influencing the way a protein is interacting with its ligand, which is strongly dictated by the binding partner preference as well as the involved binding residues of the drug complex. Knowledge of the distribution of binding behaviors of eukaryotic and prokaryotic proteomes might further reveal underlying variety in interaction processes within the different kingdoms.

Binding sites are three-dimensional complexes shaped by few amino acid residues that can be far apart in sequence, yet close in 3D structure. Enforced by the lack of experimentally verified data [30], this makes it challenging to predict residues involved in binding. Despite

the difficulty of getting accuracy scores equally high as those for predicting structural properties, we decided on analyzing binding in order to expand our comparison to the level of functional features. The analysis of protein disorder, transmembrane helices and aspects of secondary structure covers central parts of the structural protein feature space. PredictProtein also includes methods for the prediction of properties directly describing aspects of protein function. Completing the set of features covered by our analysis, we explored protein, DNA and RNA binding sites predicted by ProNA2020 [31]. Since the cluster version of PredictProtein runs an older version, we used ProNA2019 for predicting binding sites. Recent advances in predicting binding at ROSTLAB [32] underpin the significance of this protein feature.

For the analysis of combined features (section 3.5 in chapter 3), we used our binding data in a comparison to predictions by Mészáros *et al.* [10], who chose a minimum length of six amino acid residues for each binding region. In order to exclude biophysically questionable predictions and provide comparable results, we adapted the same minimum length threshold. Since ProNA2019 classifies the reliability of its predictions into three classes, we additionally restricted our analysis to regions of the highest reliability index class.

Binding Category Fractions In the context of binding we define the region content per protein as the sum of the number of residues in any binding region in a protein, divided by the protein length. For the fractions, we calculated the means per binding category.

Using protein disorder: Fraction of disordered residues used in binding The fraction of disordered residues used in binding consists of the number of the sum of the lengths of all protein-binding regions (PBRs) overlapping disordered regions, divided by all disordered residues. Here, the definition of an overlap includes all situations where a residue of a DR concurrently is part of a PBR.

2.3. Analysis and Visualization Methods

As described in the previous section, we used PredictProtein to generate prediction data of four selected protein properties for nine whole Swiss-Prot proteomes (see Table 2.1). In the following, we will give an overview of methods that were part of our analysis.

2.3.1. General Methods

In order to make use of raw PredictProtein output data, we adapted existing ROSTLAB parsing scripts for each individual feature output file. PredictProtein specifies feature information in the form of regions relating to the amino acid sequence of a single protein. Thus, it is possible to assign most of our provided visualization to either of two categories. One respects protein grouping per proteome, for example the distribution of the number of feature regions per protein. The other contains plots showing information that is detached from protein grouping, such as the distribution of the feature region lengths. For plotting, we used the `Matplotlib` [33] and `seaborn` [34] libraries.

2.3.2. Methods Applied During Plotting

In the following, we will define concepts used during our analysis.

Relative Indices Indices of feature regions can take values in $[1, n]$ with n being the length of the respective protein. Comparing absolute lengths or indices of regions across proteins suffers from the missing information of the protein lengths. The underlying total number of amino acids can be incorporated using the concept of relative indices, which is illustrated in Figure 2.1. Here, we normalized absolute indices ranging from 1 to n into an interval of $[0, 1]$ relative to protein length. Thus, we could transform feature region lengths as well as start and end indices of feature regions from proteins with an arbitrary amount of amino acids into decimal values from the same interval.

Figure 2.1.: Concept of indices relative to protein length, ranging from 0 (beginning of the protein) to 1 (protein end). For a protein of length n as indicated on the left, a feature region ranging from $n/2$ to $3n/4$ can be transformed into a region ranging from 0.5 to 0.75 using relative indices.

Statistical Significance For several forms of visualization we provide information about the statistical significance of our results, given that the plot style allows an non-obstructive display. For the standard error (SE) calculation, we estimated the underlying distribution using bootstrapping. We created 1000 artificial versions of each proteome that were sampled with replacement from the real values. For each artificial proteome, we partitioned the data into the exact same bins as for the real proteomes and calculated the bin-wise measurement, resulting in 1000 artificial values per bin. The standard deviation (SD) of each bin serves as an estimate for SE. For each bin value $\overline{x}$, we constructed the 95% CIs for a significance level of $\alpha = 0.025$ as follows:

$$[\overline{x} - t_\alpha(DOF) \cdot SE, \overline{x} + t_\alpha(DOF) \cdot SE] \tag{2.1}$$

Assuming that the artificial values are normally distributed for each bin as suggested by the Central Limit Theorem, we used the 95 quantile of the Student's t-distribution ($t_\alpha(DOF)$), leading to a symmetric two-sided confidence interval (CI). For proteome sizes n as given in Table 2.1, $n - 1$ degrees of freedom (DOF) resulted in a value of 1.96 for $t_\alpha(DOF)$.

For plots other than histograms, which display a binned distribution, we calculated CIs for single bars per proteome while skipping the binning procedure. Otherwise, we followed the above-mentioned bootstrapping method with the same values for the parameters.

Kullback-Leibler Divergence Oftentimes, the proteome-wide distribution of a protein feature can be displayed as a histogram, as performed for protein and region lengths, number of regions in a protein and more. We calculated the Kullback-Leibler (KL) divergence for each ordered pair of proteomes (a, b) to provide numeric information in addition to the visual comparison of the histogram bars and potential computed Kernel Density Estimates (KDEs). The KL divergence is an asymmetric statistical measurement of the distance of two probability distributions such as the binned protein feature distributions. It can take values of $[0, \infty)$ with 0 representing equality. We computed the KL-divergence as formally

defined in Equation 2.2 using SciPy [35]. Here, p and q are the distributions of n bins for proteomes a and b, respectively.

$$D_{KL}(p||q) = \sum_{i=1}^{n} (p(x_i) \cdot log(\frac{p(x_i)}{q(x_i)}) \tag{2.2}$$

Since the denominator in Equation 2.2 will be equal to 0 if at least one bin value in q is 0, we added pseudo counts of $exp(-12)$ to each bin value. Thereby, we prevented ∞-values as a result for the KL divergence.

Cross-Correlation We found that the distribution of certain qualities shows resemblance to audio signals when displaying it in form of a line function. In signal processing, cross-correlation (CC) is used as a measure of similarity between two signals, capturing their lag or displacement relative to each other. Since the definition of CC is ambiguous, Equation 2.3 specifies how we defined the CC of two discrete distributions r, v corresponding to proteomes a, b. We calculated CC using `NumPy` [36] as in the definition in Equation 2.3. Here, l is the lag and $conj$ is the complex conjugate function, following signal processing convention. The used function `numpy.correlate` outputs a vector of the same length as the distributions r, v, containing values of $(-\infty, +\infty)$. These correspond to the correlation values for each possible lag. The index of the maximum CC value indicates the lag of the two distributions. A maximum at 0 corresponds to an offset of 0 and thus, irrespective of their immediate similarity as estimated by KL, means that the two signals or distributions are not shifted but already as alike as possible.

$$CC_{rv}(l) = \sum_{m=-\infty}^{\infty} (r[m+l] \cdot conj(v[m])) \tag{2.3}$$

The resemblance of the data to signal spectra is why in certain cases, we used CC to measure how well two distributions match up with each other as an alternative to KL. Thus, CC captures different aspects of similarity while still supplementing the visual part of our comparison.

3. Results of Analysis and Visualization of Sample Proteomes

Using protein feature prediction data, we compared nine organisms on the basis of their proteomes. Differences in data set size, equal to the number of proteins in a proteome, or in the average length of a protein captured in the distribution of protein lengths, can influence any absolute statistics produced in the analysis. Therefore, we performed a preceding analysis in order to evaluate potential biases arising from unequal underlying distributions of principal properties such as proteome size and protein length. We then applied our analysis methods in the comparison of the selected sample organisms and developed various forms of visualization. This chapter will provide our feature-specific findings and finish by discussing the individual results in a unifying context.

Preceding Data Analysis for Bias Identification

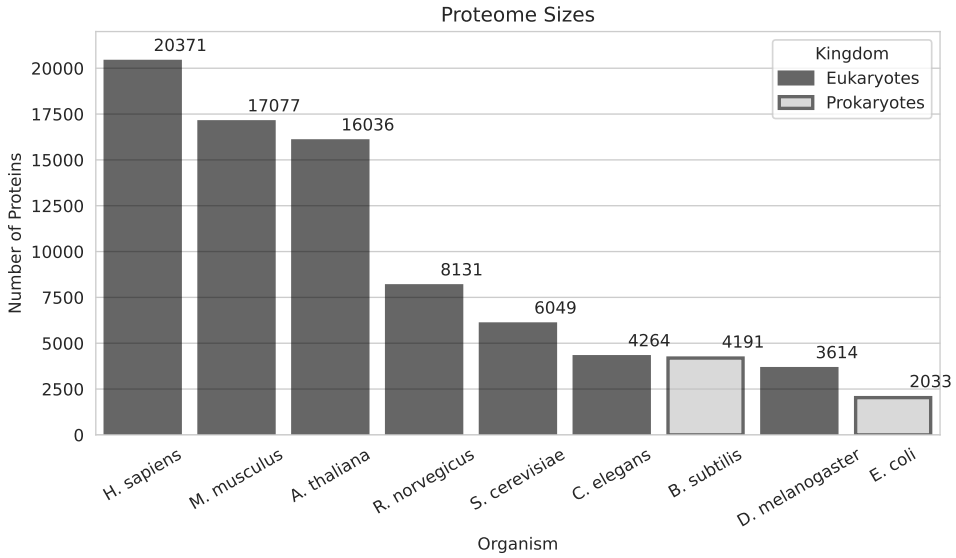

Figure 3.1.: Number of proteins contained in the Swiss-Prot proteome of the respective organism. Exact numbers as given in Table 2.1 are shown above the top-right corner of each bar.

The data set size for the Swiss-Prot proteomes (see Table 2.1) varies almost by an order of magnitude in the most extreme case, *Homo sapiens* and *E. coli*. To correct for this variation, most visualizations that will be presented in following chapters present numbers relative to the proteome size. Still, sparse data may hold more variability and lead to less stable distributions. Thus, any required smoothing has more potential to alter the distributions of the respective proteomes. To give an overview of the existing differences in data set dimension, Figure 3.1 displays the size of each proteome as given in Table 2.1.

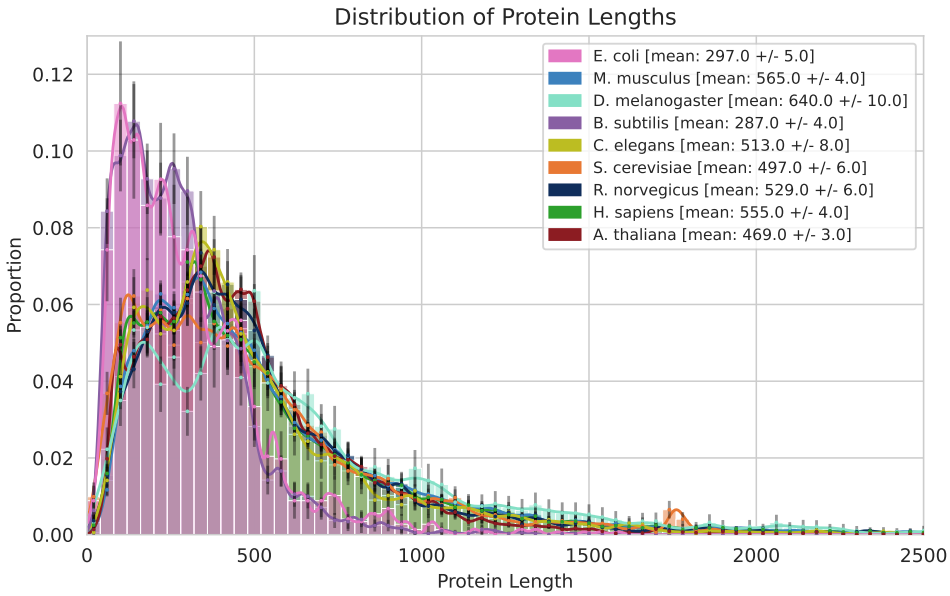

Figure 3.2.: Distribution of protein lengths with computed Kernel Density Estimate (KDE) and error bars corresponding to the 95% confidence intervals. Bin values are relative to the respective proteome size. We found that for every proteome but one, less than 1% of all its proteins are longer than 2500 amino acids. For *Drosophila melanogaster*, this holds true for less than 1.8%. To allow enough details to be visible in the present data, we chose a cutoff of 2500 residues.

The protein length distribution (Figure 3.2) already separates prokaryotes and eukaryotes. While among the eukaryotes, some variations are present up to a length of 500, they follow an increasingly similar distribution for longer proteins. An exception occurs at a length of around 1700, where *Saccharomyces cerevisiae* proteins are overrepresented. This might be due to a comparatively large number of Ty1 and Ty2 retrotransposon polyproteins registered in Swiss-Prot, potentially arising from the extensive study of Ty transposons in yeast [37]. Overall, the eukaryotic distributions are clearly differing from the prokaryotes, which show significantly less proteins with more than 500 amino acids. This is also captured in the difference in median protein length (see legend of Figure 3.2). Brocchieri *et al.* previously analyzed proteomes of the two kingdoms and noted similar median lengths [38]. We could confirm that on average, a prokaryotic protein is shorter than one of eukaryotic origin. In many cases, this variation between the two kingdoms could be found in other concepts of our analysis that utilize protein lengths.

As Figure 3.2 demonstrates, protein length can vary heavily within the same proteome. Thus, absolute indices and resulting absolute feature region length can have vastly different interpretations depending on the length of the protein that the region is part of. For example, a disordered region (DR) of length 40 in an average *E. coli* protein (mean length 297.5) hardly equals a DR of the same length from a 40-residue protein in terms of meaning. While the DR only comprises roughly 13% of the average protein, it makes the other one a fully disordered protein. The caveat of this approach of defining indices relative to protein length is the loss of the smallest biologically supported unit, the amino acid. Without it, it may be harder to interpret potential results. However, incorporating protein length information provides additional insights as demonstrated. Because of the underlying distribution of Figure 3.2, this also favors a separation of prokaryotic and eukaryotic distributions. The concept of relative indices is explained in Figure 2.1.

3.1. Protein Disorder

Composition In Figure 3.3, we analyzed the disorder composition using the definition made by Schlessinger *et al.* [7]. We could substantiate and visualize that the disorder

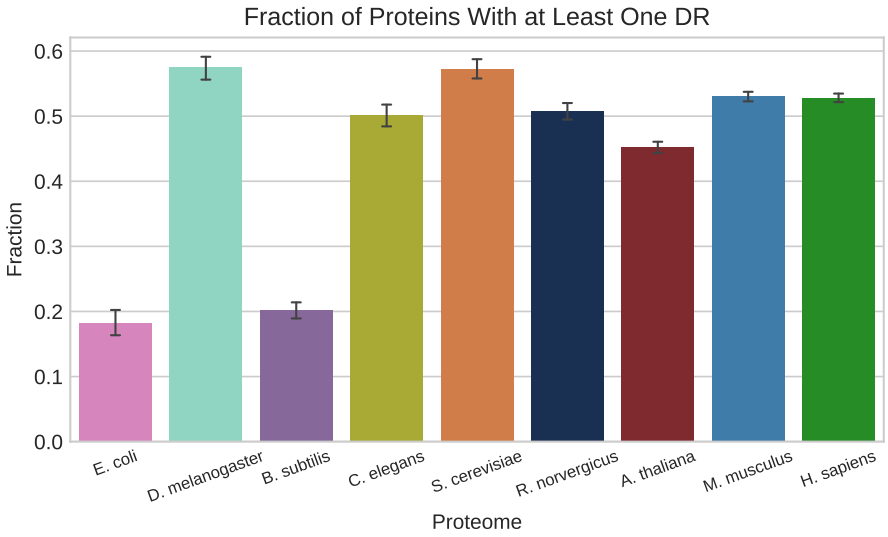

Figure 3.3.: **Disorder composition** among sample proteomes. The relative composition of a proteome is defined as the number of proteins containing at least one long DR normalized by proteome size. Error bars indicate 95% CIs.

composition of a proteome is related to the complexity of its kingdom, as both prokaryotes show significantly lower values than any eukaryote. This observation hints at the suggestion by Schlessinger *et al.* of protein disorder as a biological tool of an organism to expand in complexity. While our data can explain the advance from prokaryotes to eukaryotes utilizing disorder, the composition fails to accurately display the difference in complexity between the eukaryotes. Here, the single-cell organism yeast (*S. cerevisiae*) ranks higher than rat (*R. norvegicus*) for instance. This clarifies that the simplicity of measurement such as the composition limits its potential to fully describe a biological phenomenon as complex as protein disorder.

Proteome Disorder Content The proteome disorder content preserves the ranking order according to the composition (see Figure II.4 in the appendix). Thus, it contributes to the view of an unequal distribution of disorder among organisms and its involvement in the difference in complexity between kingdoms. However, we noticed that using only proteins that contain DRs for normalization reverses the ranking of eukaryotes compared to prokaryotes (see Figure II.5 in the appendix). This observation raises awareness of the fact that the exact definition of a measurement, including aspects of normalization, has a notable effect on biological conclusions to be drawn from the data. We deemed our original way of defining the proteome content as more reasonable as the normalization captures more information. The result of applying each definition can be referred to in the appendix (Figure II.4, Figure II.5).

Number of Disordered Regions The separation of kingdoms can also be observed when plotting the distribution of the number of DRs per protein. In addition to the

visual comparison, we produced a heatmap depicting the KL-divergence of the two resulting distributions (Figure 3.4). Prokaryotes had significantly more proteins without any disordered region, almost doubling the proportion in eukaryotes. This is reversed for proteins with at least one disordered region, which are invariably overrepresented among all eukaryotic model organisms. The KL heatmap supports this distribution-wide separation of kingdoms numerically and demonstrates that the distribution of the number of DRs per protein varies significantly more between organisms of different kingdoms than among kingdoms. However, the observation of prokaryotes having less DRs per protein on

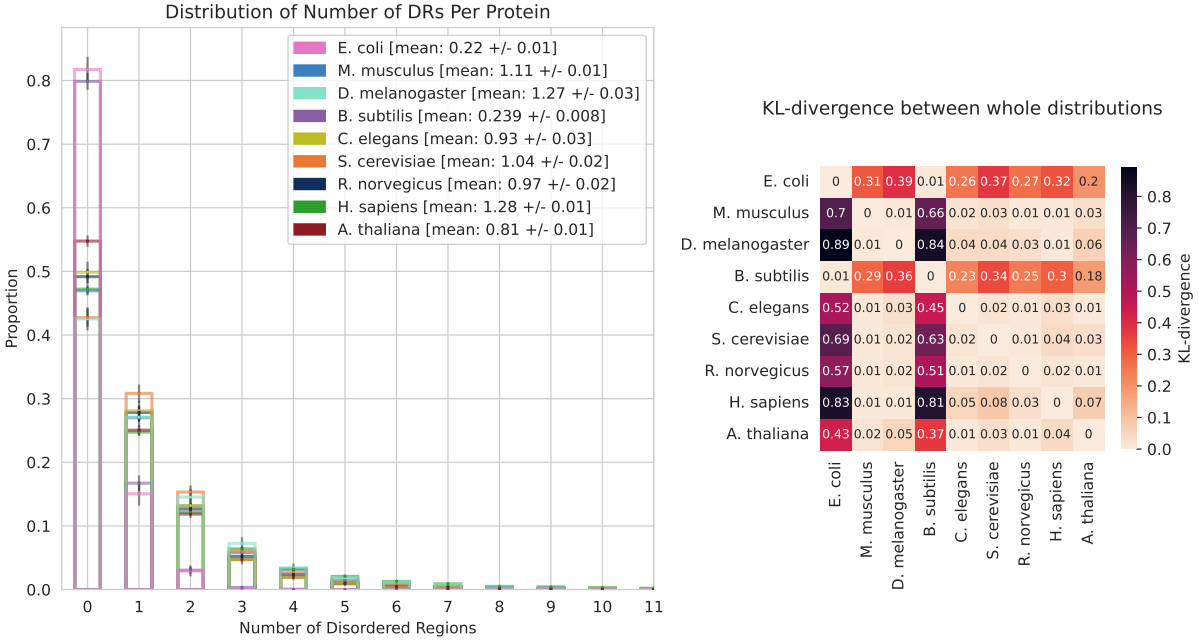

Figure 3.4.: **Number of DRs per protein** (a) and **KL-divergence** between whole binned distributions heatmap (b). **(a)** Numbers are relative regarding proteome size. Error bars represent 95% CIs. We chose eleven as a cutoff, since more than 99% of proteins do not contain more than eleven DRs. **(b)** KL-divergence calculated as described in chapter 2. Dark coloring indicates a high divergence.

average might also be influenced by the underlying protein length distribution (Figure 3.2). Assuming similar DR lengths in the kingdoms, shorter prokaryotic proteins could imply less "space" for DRs. Another possible reason for this observation is the fact that prokaryotes, especially *E. coli*, are very well researched. Potentially, extensive research has explored prokaryotic proteins in detail and thus excluded false-positive DRs in the data that was used for Meta-Disorder [6] training, decreasing the number of predicted DRs.

Disordered Region Lengths In order to investigate this, we analyzed the data on the level of regions, after proteome and protein level. Independent of the protein a region belongs to, we visualized the distribution of the lengths of DRs in a proteome in Figure 3.5. Bins and a region minimum length requirement of six instead of 30 amino acids were chosen in accordance with the parameter selection of Mészáros *et al.* [10]. Although we used Meta-Disorder [6] instead of IUPred [11] for disorder predictions, we could produce similar distributions of DR lengths for both kingdoms.
Prokaryotic DRs tend to be shorter than DRs in eukaryotic proteins, even when sticking to our initial minimum length threshold of 30 amino acids (Figure II.3). Thus, the observation

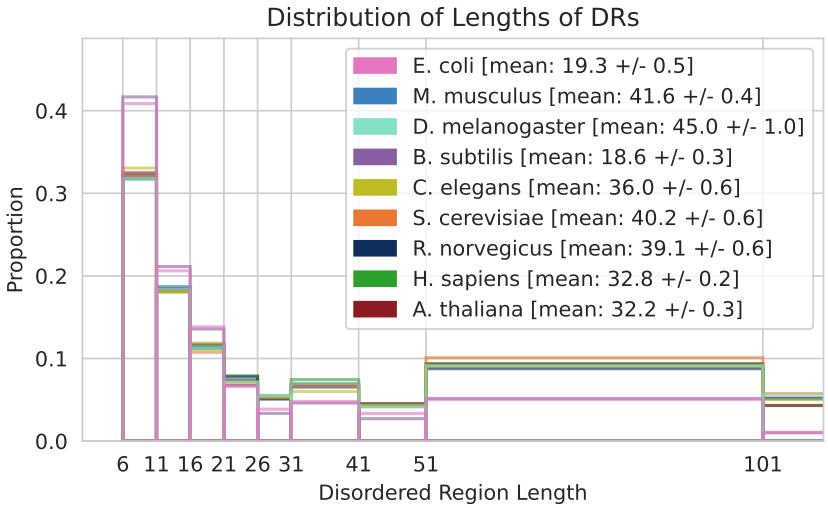

Figure 3.5.: **Absolute disordered region lengths.** Binned values are relative regarding proteome size. Bins were chosen to equal the selection made by Mészáros *et al.*

of prokaryotic proteins comprising less DRs on average cannot be exclusively linked to the underlying protein length distribution restricting the number of DRs a protein can contain. Yet, when applying the concept of relative indices (see chapter 2), eukaryotic disordered regions are smaller when being viewed relative to protein length as depicted in Figure 3.6. This implies that potentially, protein length still heavily influences the presence of DRs. Irrespectively, absolute disordered region length increases with kingdom complexity.

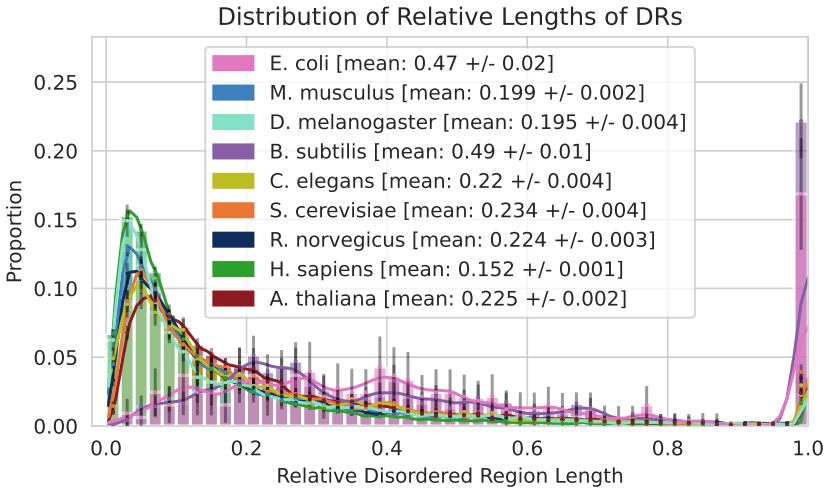

Figure 3.6.: **Relative disordered region lengths** with KDE and error bars indicating 95% CIs. Binned values are relative regarding proteome size.

In addition, prokaryotes have more spanning or nearly spanning disordered regions in comparison to eukaryotes and thus, more fully or nearly fully disordered proteins relative to proteome size Figure 3.6. At the same time, proteins of prokaryotic organisms are more often without any DR than eukaryotic proteins Figure 3.4.

Spread of Disordered Regions Relative region lengths build on the concept of indices relative to protein lengths. Since this enables a comparison across proteins of different lengths, we analyzed the within-protein location of DRs in Figure 3.7. Here, visually closely clustered lines corresponding to the different proteomes can be grouped together. Again, the two prokaryotes separate themselves from the eukaryotic proteomes. Interestingly, regions in worm and yeast as well as rat, mouse, human, and fruit fly respectively are spread highly similarly. Cress shows approximately equal similarity to both eukaryotic groups. All follow a development that is contrary to the distribution of prokaryotic DRs, which are underrepresented in the first quarter of the normalized protein ([0.0, 1.0]). Yet, prokaryotes contain slightly more DRs located in the two inner quarters.

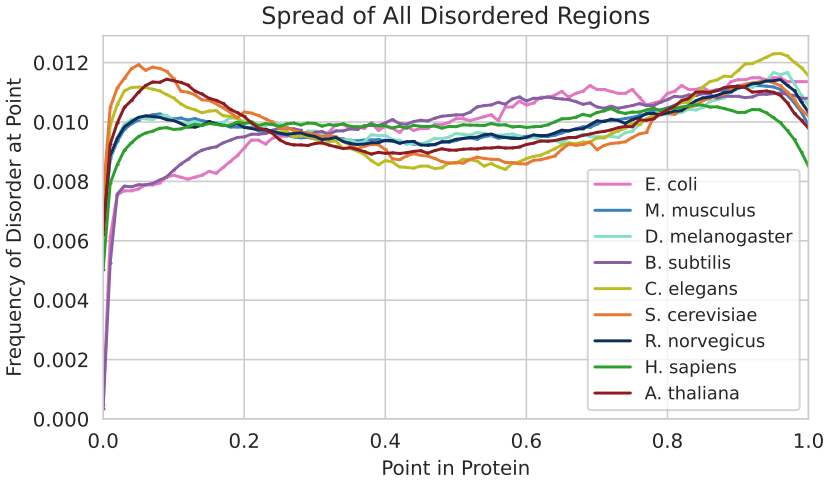

Figure 3.7.: Abundance of **disordered residues at relative protein indices** normalized by the total number of disordered residues in a proteome.

Liu *et al.* [39] noted that certain regions connecting protein domains are more abundant in eukaryotes, most probably because of a higher number of multi-domain proteins. They describe these linker regions as significantly more disordered in the eukaryotic kingdom, and suggest this phenomenon as a reason for the systematic increase in disorder from prokaryotes to eukaryotes, which we have noted as well. Assuming this argument, our results propose that these linkers could be located predominantly in the first quarter of eukaryotic proteins.

Disorder Diamonds and Spectra Using relative indices, we combined DR location and length information. Figure 3.8 displays the mean relative start and end positions of all DRs centered at a specific position in the protein. A more detailed explanation of how to read the diamond plot is given in the caption.

The mean start and end value distributions of eukaryotic organisms tended to be closer to the y-axis in Figure 3.8, containing smaller values as opposed to prokaryotic distributions. The latter showed stronger variability in mean values for different center positions, presumably because of a smaller sample size. Visually, prokaryotes seem to have wider (*i. e.* relatively longer) DRs for several y (center) values, building onto observations made from Figure 3.6.
Following our reasoning when explaining our process from diamond to spectrum plots in chapter 2, we extended our analysis of DR length and location in Figure 3.9.

The widest regions appeared at center positions around 0.5 in the spectrum plot, most

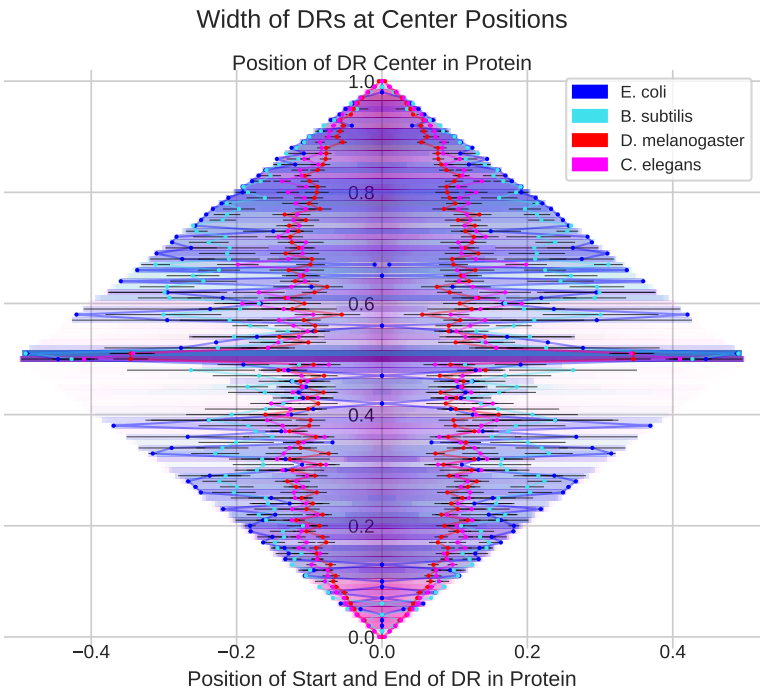

Figure 3.8.: **Disorder diamond plot** for four sample proteomes. DRs were grouped into 0.2-wide bins according to the position of their relative center index. Points represent mean start (negative x-axis values) and end (positive x-axis values) for all DRs centered at y. For instance, a protein-spanning DR has a relative center index of y=0.5. Start and end positions relative to the center position result in the minimum and maximum x values, -0.5 and 0.5 respectively. In order to ensure comparability irrespective of proteome size, we chose colors of equal saturation and brightness and specified a level of bin transparency that corrects for different numbers of contained regions. Error bars indicate 95% CIs. Proteome colors were selected based on kingdom.

probably because of protein-spanning DRs that, trivially, have 0.5 as their center position. This finding is supported by previous observations such as for Figure 3.6, demonstrating that prokaryotes appear to have more fully disordered proteins than eukaryotes. In addition, peaks and valleys tend to correlate among and further, across kingdoms. Most notably, gaps that occurred because no DRs were centered at these positions correlate across all proteomes. Because of the lack of an obvious explanation for this phenomenon, it remains an open question and requires further analysis. One way of measuring the visual correlation is calculating the cross-correlation (CC) for each pair of proteomes (Figure 3.9). CC distinctly separates the pairings based on the combination of their kingdoms. The curve for the prokaryotic pairing, *B. subtilis* and *E. coli*, is set apart from the curves for the inter-kingdom pairings, while the eukaryotic curve for fruit fly and worm falls below all others. This division of curves in three groups could be found for every combination of 4 out of the selected sample proteomes. Thus, we could split the two-dimensional space as depicted in Figure 3.9 in three zones, corresponding to prokaryotic, mixed and eukaryotic pairings. Interestingly, an exception from this rule merely occured for mixed pairings involving the human proteome, which appeared in the eukaryotic zone.

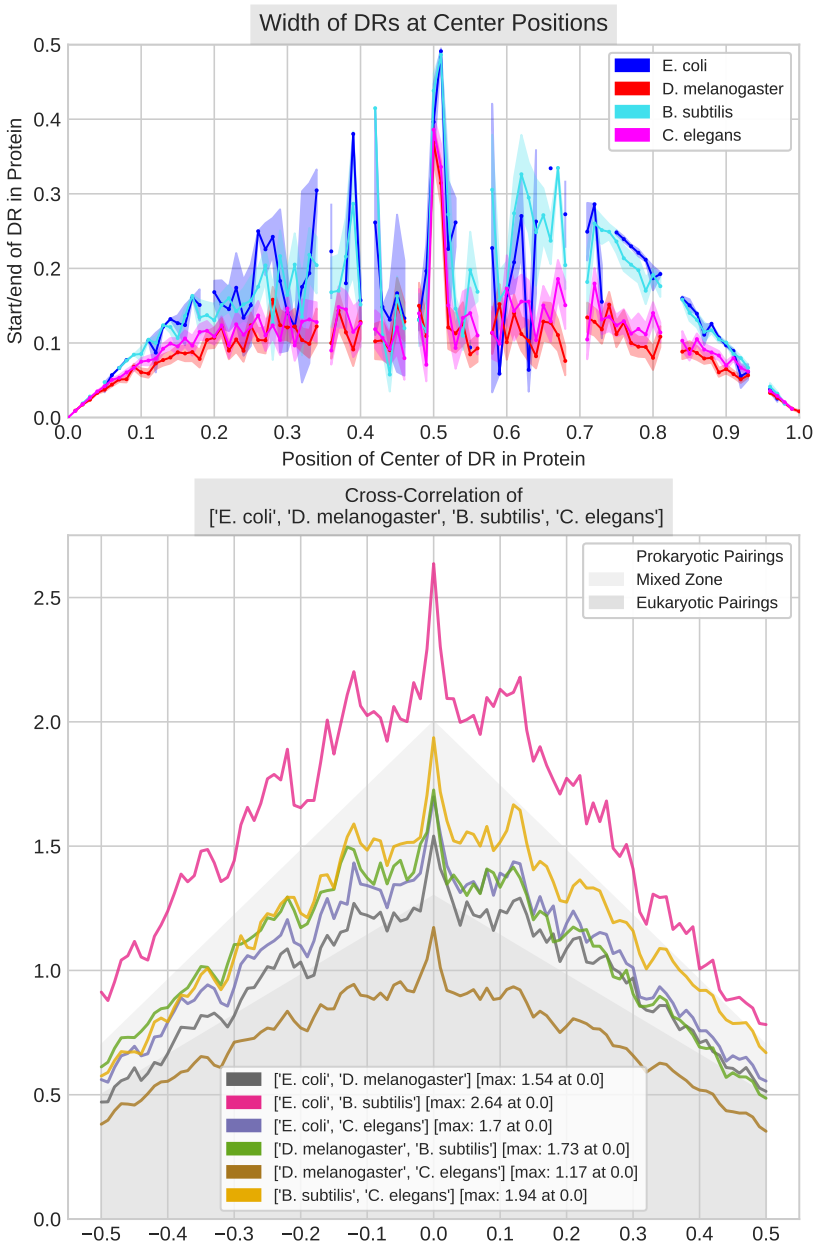

Figure 3.9.: **Disorder spectrum plot** (a) and pairwise **cross-correlation** (b) for four sample proteomes. **(a)** As in Figure 3.8, means for the distance of relative start/end positions to the DR center were computed for each center position on the x-axis. The shaded area behind each curve indicates the 95% CI. To prevent visual overload because of small proteome size, we excluded positions without any present DR centers from the plot. Fail-safe CI and cross-correlation (CC) calculations were executed using pseudo-counts instead. **(b)** CC curves can be grouped according to the kingdom combination of the respective proteome pair. Corresponding zones are explained in the legend. Spectrum plots such as Figure II.2 for other combinations of four out of the nine sample proteomes can be found in the appendix (Part II).

3.2. Transmembrane Helices

Protein Classes We confirmed established views [21] by observing that among most organisms, a fraction not exceeding 25% of the proteins of a proteome are transmembrane proteins (TMPs). Additionally, we note that the two sample prokaryotes show remarkably similar proportions of all three classes as depicted in Figure 3.10, differing from every eukaryotic distribution. For a visualization for all organisms, refer to the appendix. Here, we could confirm observations made by Liu & Rost [21] regarding the absence of a correlation between organism complexity and TMP proportion. Notably, eukaryotes seem to contain less TMPs relative to proteome size than prokaryotes, with rat being the only exception. Likewise to all other proteomes of its kingdom, noticeably more single-pass TMPs are present in rat, equalling the respective fraction in cress. Yet on the contrary to the other eukaryotes, the fraction of multi-pass TMPs in rat almost compares to prokaryotic levels. By including rat in our analysis, we affirm existing findings: The fraction of TMPs varies more among kingdoms than between them and thus invalidates itself for explaining differences in complexity. Still, in Figure 3.10, the two kingdoms can be consistently discerned based on the fraction of single-pass TMPs.

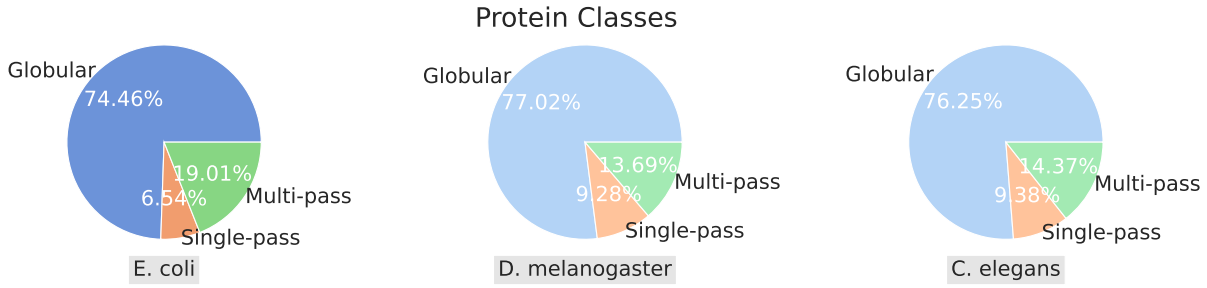

Figure 3.10.: **Proportions of protein classes** among three sample proteomes. We used darker colors for prokaryotic data. Accordingly, pie-charts for all sample organisms can be found in the appendix (Figure II.6).

Transmembrane Content Per Protein The highest proportion of eukaryotic TMPs appeared at a transmembrane content level of less than 0.1 (Figure 3.11). For prokaryotes, TMPs with a content of around 0.6 comprised the biggest proportion, with CIs not overlapping the eukaryotic CIs. In addition, the average eukaryotic TMP has a transmembrane content of around 0.2, half the value for the average prokaryotic TMP with a content of around 0.4. These significant differences clearly discern the two kingdoms. Since the definition of transmembrane content makes use of the protein length, this separation could be linked to the underlying distribution (Figure 3.2) displaying the shorter total lengths of prokaryotic proteins. Potentially, the combination of the number of TMHs and their lengths could result in a similar absolute number of TMH residues per protein for both kingdoms. For eukaryotes, these findings imply either long loop regions in-between TMHs, or extensive tails on the cytoplasmic or extracellular side of the membrane.

Orientation of Transmembrane Proteins We note that likewise to the analysis of protein classes earlier in this section, general remarks about differences between eukaryotes and prokaryotes can be made despite the occurrence of exceptions. Apart from yeast, the distribution of TMP orientations differs vastly between the two kingdoms. While

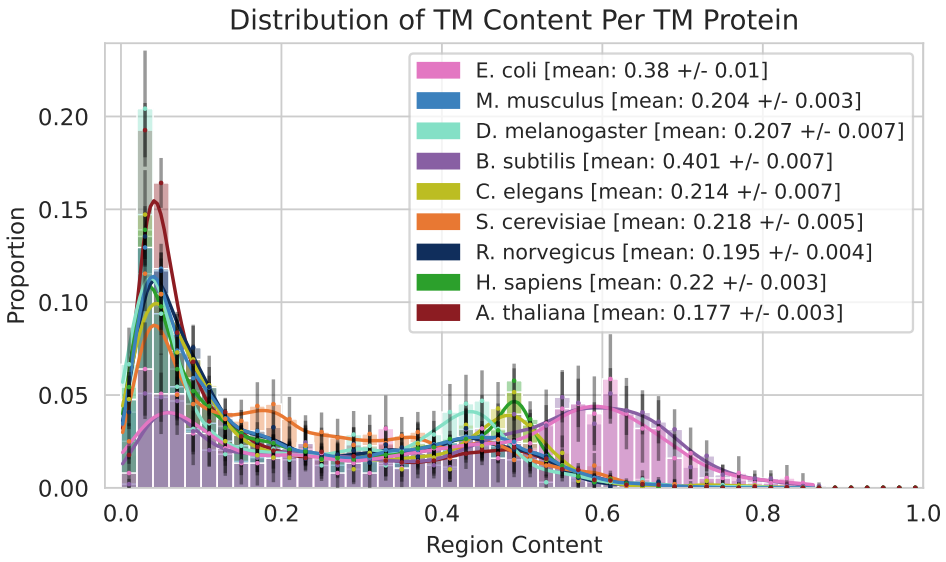

Figure 3.11.: **Transmembrane content per transmembrane protein**, defined as the number of TMH residues divided by the protein length, with KDE. Numbers are given relative to proteome size. Error bars indicate 95% CIs.

cytoplasmic TMPs are more abundant in prokaryotes, all eukaryotes reveal a greater extracellular fraction. The latter observation also holds true for yeast. Yet, a much smaller fraction of the "Membrane" class together with more cytoplasmic TMPs are the reason why the distribution for *S. cerevisiae* looks similar to the prokaryotic samples. In addition, we found that eukaryotes have less "Membrane"-oriented TMPs. Here, fruit fly deviates from the norm due to showing a class fraction even higher than the prokaryotic counterparts. Overall, we observe that the kingdoms can be distinguished based on their level of extracellular TMPs and, excluding yeast, most reliably based on their cytoplasmic TMP fraction.

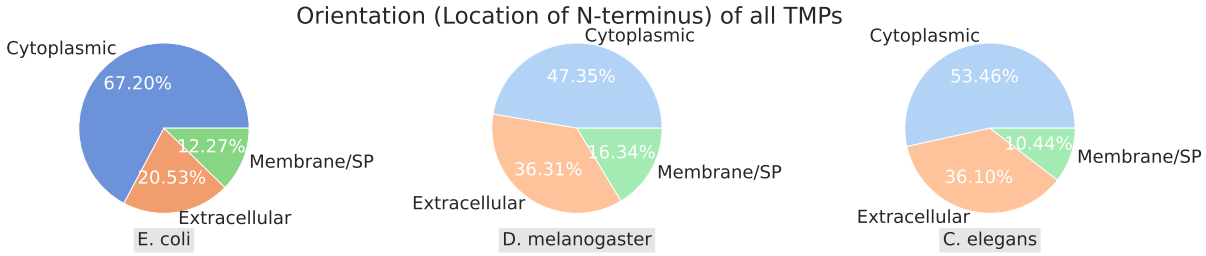

Figure 3.12.: **Proportions of TMP orientations** among three sample proteomes. We used darker colors for prokaryotic data. Accordingly, pie-charts for all sample organisms can be found in the appendix (Figure II.7).

Topological Fractions of Transmembrane Proteins The two kingdoms demonstrate substantial differences in the mean topological fractions of their transmembrane proteins (Figure 3.13). Prokaryotes appear to comprise a significantly bigger fraction of membrane residues. At the same time, the fraction of outside residues is underrepresented in comparison to the eukaryotic counterparts, which also have bigger "inside" fractions. These observations are consistent and significant across all sample organisms included in

our comparison. Since prokaryotic proteins are shorter on average, the results can have implications comparable to our analysis of the transmembrane content per protein. It is possible that for both kingdoms, TMH length and frequency per protein together produce similar absolute numbers of membrane residues per protein. Assuming this reality, the mean fraction of membrane residues in TMPs would yield higher numbers for prokaryotic proteomes because of the underlying protein length distribution. Irrespectively, these "relative" definitions constitute a solid way of discerning eukaryotes and prokaryotes, supported by consistent statistical significance.

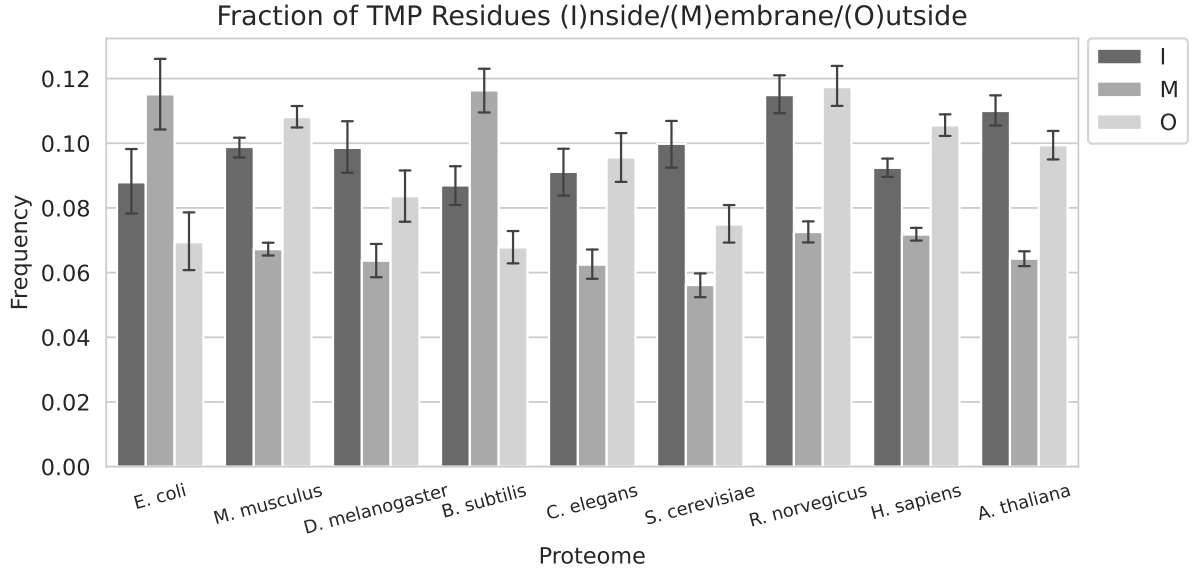

Figure 3.13.: **Fractions of residues** corresponding to TMP topology. These are defined as the number of residues per protein predicted to be part of the cytoplasmic (inside), membrane, or extracellular (outside) component, divided by the protein length. We calculated the means per component for each proteome. Error bars indicate 95% CIs. A more detailed view of the underlying distributions for each component class is given in Figure II.8.

Number of Transmembrane Helices The transmembrane helix as a unit is still a biologically valid segment and intuitive to grasp. Thus, as for disordered regions, we analyzed the distribution of the number of TMHs per transmembrane protein and visualized it in form of a histogram (Figure 3.14).

Here, we could reproduce findings of Liu & Rost who observed imbalances in the number of TMHs for the two kingdoms [21]. We also note that 7-TM proteins, TMPs with 7 TMHs, are significantly more abundant in certain eukaryotes such as *C. elegans*. As suggested by Liu & Rost, this observation can be linked to an over representation of smell receptors in worm. In addition, we confirm that numerically, the organisms comprising the greatest proportions for 6-TM and 12-TM proteins are prokaryotes. Potentially, this can be explained by transporter proteins predominantly comprising twelve TMHs or two domains of six TMHs [21, 40]. However, the differences for these numbers of TMHs were not significant. The appendix includes a more detailed depiction of the differences per number of TMHs for each pair of proteomes. For instance, Figure 3.15 demonstrates that *C. elegans* has significantly higher proportions of single-pass and 7-TM proteins as well

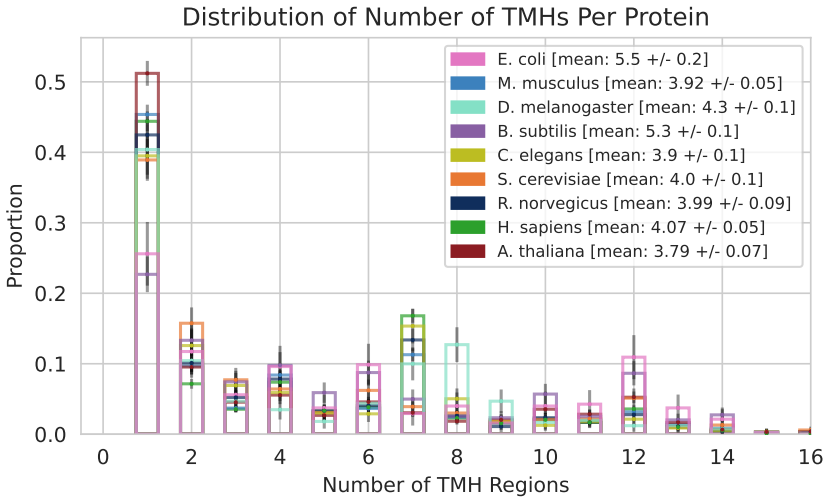

Figure 3.14.: **Number of TMHs per protein.** For improved visibility, we excluded globular proteins. Numbers are given relative to the total number of TMPs contained in the respective proteome. Error bars indicate 95% CIs. We chose 16 as a cutoff, since more than 99% of TMPs do not contain more than 16 TMHs.

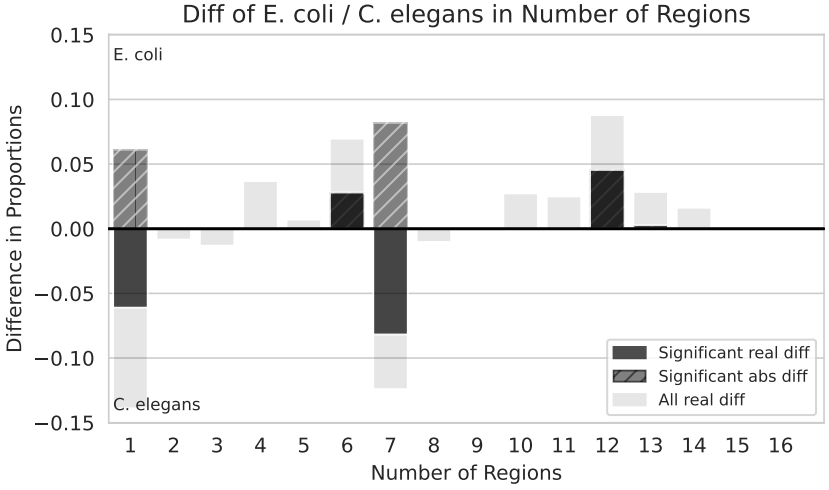

Figure 3.15.: Difference in number of TMHs per protein for two sample proteomes. Dark bars indicate significant differences, striped shading represents the absolute value of significant differences. All other differences in proportions within the 95% CIs of both proteomes for the respective bin are given in light grey. Plots visualizing the pairwise differences for other pairings of sample proteomes can be seen in Figure II.9.

as less 6-TM and 12-TM proteins in comparison to *E. coli*. Earlier in this section, we discovered that in most cases, prokaryotes have a higher fraction of multi-pass TMPs. In accordance to this, we demonstrate that overall, the average prokaryotic TMP in our analysis had one TMH helix more than its eukaryotic counterpart, as indicated in the legend of Figure 3.14.

Our analysis yielded that between proteomes, not only differences of significance between proportions of TMPs with a certain number of TMHs exist. Different numbers of helices

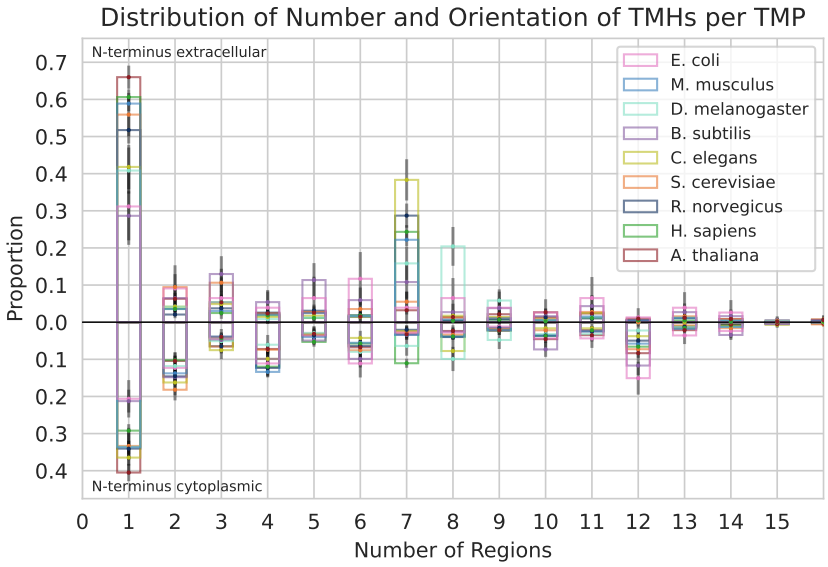

Figure 3.16.: **Number of TMHs per protein with topologies.** For improved visibility, we excluded globular proteins, TMPs with an amino terminus located in the membrane and TMPs containing signal peptides. Numbers are given relative to the total number of TMPs of the respective topological orientation contained in the respective proteome. Error bars indicate 95% CIs. We chose 16 as a cutoff, since more than 99% of TMPs do not contain more than 16 TMHs.

per protein also favored varying TMP orientations, as defined in Figure 3.12, for different sample proteomes. In Figure 3.16, we visualized the proportions of two topological orientations, extracellular and cytoplasmic, for varying numbers of TMHs contained in a transmembrane protein. Many transporters belong to the class of 12-TM proteins [40]. A principal 12-TM transporter family typically favors a cytoplasmic C-terminus, and consequently a cytoplasmic N-terminus [41]. This phenomenon could explain the over representation of cytoplasmic TMPs with twelve TMHs in our data. Again, we thereby could substantiate findings made by Liu & Rost who recorded the majority of prokaryotic 12-TM proteins to have a N-terminus located on the cytoplasmic side of the membrane [21]. Likewise, we found that eukaryotic 7-TM proteins predominantly adopt an extracellular orientation, aligning with established knowledge of G protein-coupled receptors (GPCRs) [42]. We thus affirm the potential of topological and TMH abundance information of transmembrane proteins regarding the mapping of biophysical realities and differences between organisms and kingdoms. In addition, we could widely validate the observation made by Liu & Rost, demonstrating that the majority of TMPs have less than four transmembrane helices. In our analysis, this statement holds true for all eukaryotic proteomes. Figure 3.17 visualizes our results in the form of the cumulative distribution of the number of TMHs per protein. For prokaryotes, we found that most TMPs contain less than five instead of four transmembrane helices. This agrees with our finding of a higher average number of TMHs in prokaryotic proteomes.

Transmembrane Helix Length Distribution Complementing our extensive analysis regarding the number of transmembrane helices per protein, we explored the distribution of TMH lengths. Figure 3.18 shows proportion of absolute TMH length values.

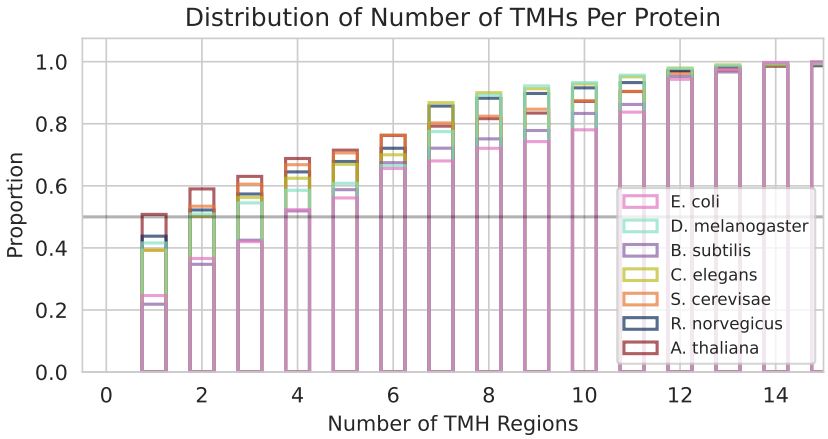

Figure 3.17.: **Cumulative distribution of the number of TMHs per protein.** For improved visibility, we excluded globular proteins. Cumulative numbers are given relative to the total number of TMPs contained in the respective proteome. A horizontal line at y=0.5 indicates the threshold for a majority in the cumulative display. We chose 16 as a cutoff, since more than 99% of TMPs do not contain more than 16 TMHs.

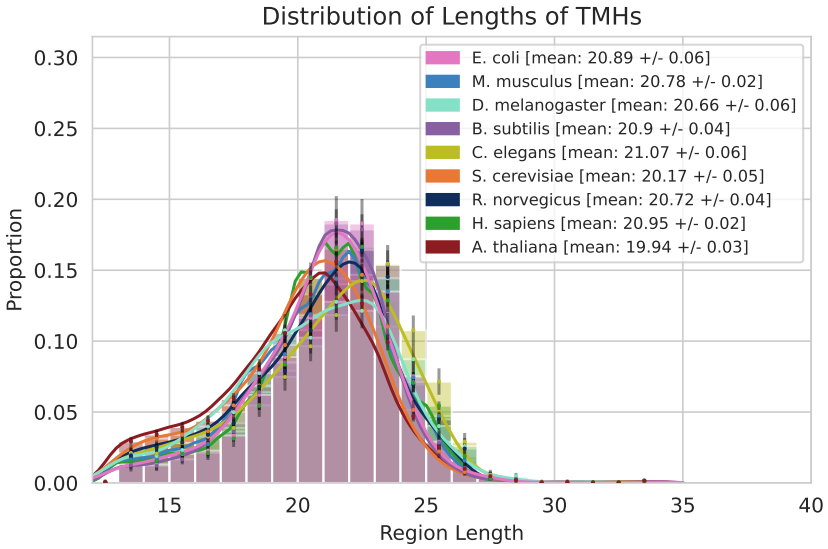

Figure 3.18.: **Absolute TMH lengths** with KDE. Binned values are relative to proteome size. Error bars indicate 95% CIs.

Unlike the results from our disordered region length analysis, we cannot find substantial TMH length differences between the two kingdoms. Visually, our data agrees with previous findings of Liu & Rost, as they observed most TMHs to comprise 17 to 33 residues. Based on our analysis, we could narrow this interval. For all proteomes, the proportions of TMHs of 19 to 24 or 20 to 25 amino acids comprise the majority of all transmembrane helices. In addition, the average length of a transmembrane helix appears to be uniformly distributed across proteomes, as indicated in the legend of Figure 3.18. We therefore report TMH lengths to be comparable across a variety of organisms of different kingdoms.

3.3. Secondary Structure

Secondary Structure Elements in Proteins By predicting secondary structure, RePROF partitions each protein into segments of helices, strands or other structural elements. For each proteome, we collected the number of proteins containing at least one β-strand or helix region, respectively. We visualized the resulting confusion matrix in form of a heat map relative to proteome size in Figure 3.19.

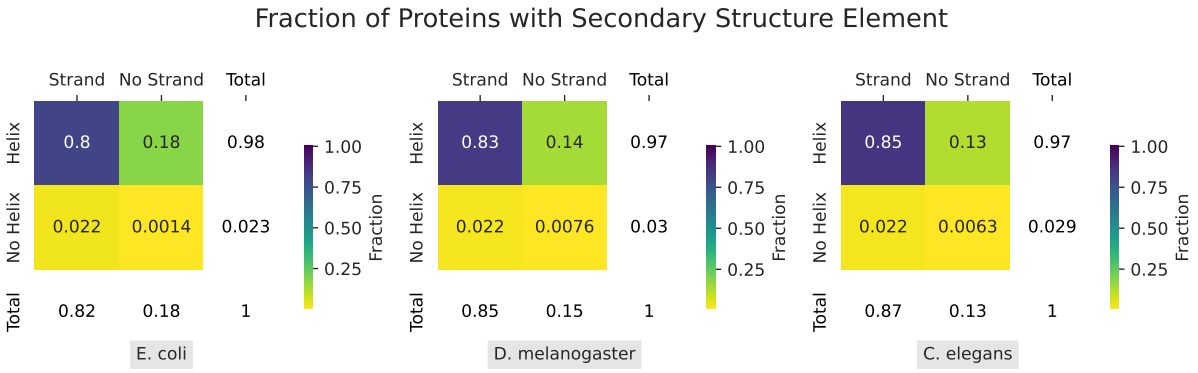

Figure 3.19.: **Presence of secondary structure elements in proteins** among three sample proteomes. The decimal numbers displayed in the three by three square indicate the number of proteins, which consist of at least one sheet or helix region, respectively. Numbers are given relative to proteome size. Accordingly, heat maps for all sample organisms can be found in Figure II.10.

We noted multiple findings that applied to all compared organisms. First, while almost all proteins consisted of at least one helix, around 80% of each proteome simultaneously also contained at least one strand element. Second, our data showed noticeably more proteins that had helical but no strand components than vice versa. In addition, eukaryotes had less proteins without strands than prokaryotes, particularly resulting in a smaller subsection of these proteins that also contained helices (top center heat map value in Figure 3.19). Interestingly, we could reproduce this qualitative trend with older RePROF cache prediction data. Using cache data of a subset of six of our sample proteomes, we noted greater differences in these numbers separating prokaryotes and eukaryotes. We likewise visualized these results in the appendix (Figure II.11). Each combination of a difference in alignment methods, RePROF prediction algorithms or proteome sequencing data could account for the varying distributions. When using new prediction data as displayed in Figure 3.19, prokaryotes and eukaryotes did not separate as well as before. Still, while the two kingdoms differentiated less, the qualitative trends we found remained.

Structural Fractions Figure 3.20 displays that we refer to as structural fractions, as explained in chapter 2. Among kingdoms, all sample proteomes showed highly similar mean content values. However, on the contrary to our previous analysis of the presence of whole structural regions in proteins, the structural fractions allow a clear separation of the two kingdoms. Prokaryotes showed a slightly higher strand content, a clear over representation of helical content and noticeably less relative residues that were classified as "other".

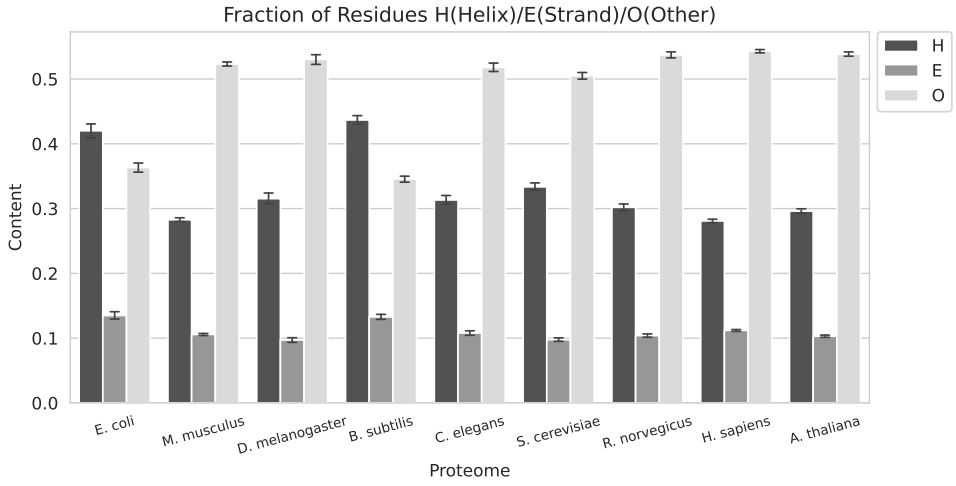

Figure 3.20.: **Fractions of residues** corresponding to secondary structure. These are defined as the number of residues per protein predicted to be located in a helical (H), strand (E) or other (O) segment, divided by the protein length. We calculated the means per component for each proteome. Error bars indicate 95% CIs. A more detailed view of the underlying distributions for each component class can be referred to in Figure II.14.

Helical and Strand Content Relation Instead of viewing helix and strand predictions as independent entities, we collected per-protein helix and strand content values, ranging from 0.0 to 1.0. In order to visualize this relationship for each proteome, we plotted both values in a two-dimensional space.

As previously, the data showed certain similarities between the two kingdoms. While a maximum of 50% of a protein's residues were sheet residues, helical content reached levels of up to 1.0. Visually, the eukaryotic proteomes filled a greater area in Figure 3.21. In particular, we observed that the prokaryotic distributions were more focused around greater values for the protein-wise helix content. For the same density estimate settings, eukaryotic estimates occupied values of smaller sheet- and especially helix content. This implies a higher number of "other" residues per protein, explaining the overall higher fraction of "other" residues on a proteome level, as depicted in Figure 3.20. Yet in contrast to the proteome-level fractions, analyzing helix and sheet content results in not one, but two eukaryotic clusters. Fruit fly, worm and yeast covered the whole spectrum of possible helix content values, but came short of greater sheet content for proteins with very few helical residues. The latter was not the case for the second eukaryotic cluster consisting of cress, rat, mouse and human, which comprise the proteomes with the highest number of proteins. This potentially limits the plot by linking it to proteome size. However, this did not prove to be the sole characteristic determining the spread and area of the kernel density estimates, since proteomes of comparable size such as *B. subtilis* and *D. melanogaster* or *R. norvergicus* and *S. cerevisiae* clearly differ in distributions. Thus, the relationship of helix and sheet content appears to cluster the selected sample organisms into two eukaryotic and one separate, prokaryotic group.

Spread of Secondary Structure Regions Apart from per-protein content, there are more ways of enabling the combination of protein-specific information. Making use of the relative definition of start and end indices of secondary structure regions, we explored the

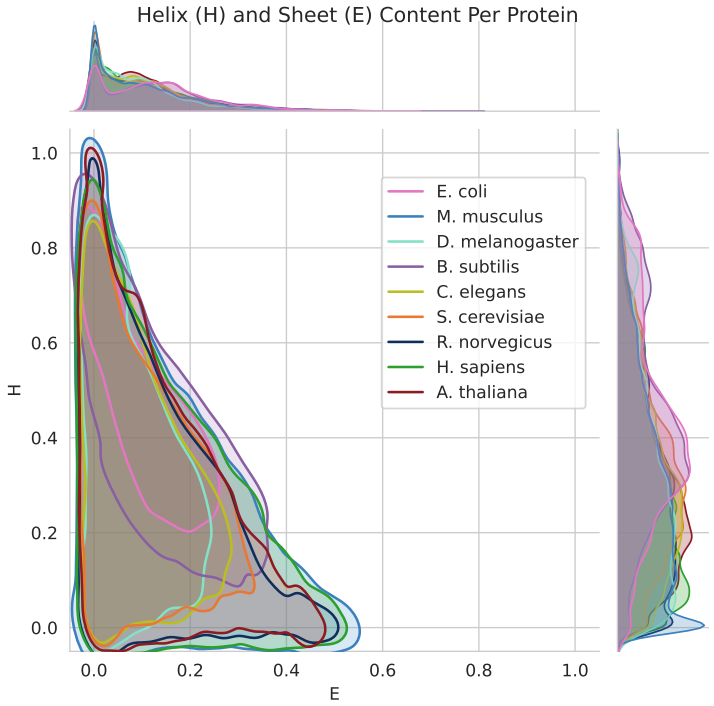

Figure 3.21.: **Helix and sheet content per protein**. Here, each data point is used to calculate the KDE, replacing the traditional scatter plot. On the sides, the underlying distribution of the content per protein is shown for the two secondary structure elements.

location of helical and strand elements relative to protein length. In Figure 3.22, each line represents the spread of regions of the two kinds of secondary structure for each proteome.

In terms of the spread of regions, eukaryotic helices followed distributions that appeared almost uniform for most relative points in the protein. Here, we noted yeast as an exception, behaving neither particularly similar towards organisms of its kingdom nor towards others. The two prokaryotes produced highly similar distributions. They differed from the eukaryotes most noticeably in the first 0.1-wide part of the normalized protein, which showed a stronger presence of prokaryotic helices. Since this first stretch of the amino acid sequence constitutes the characteristic location of signal peptides, we cautiously suggest that secretory protein carrying signal sequences could be more abundant in prokaryotes. A solid explanation for our observation however requires further research. Further, the spread of strand regions were not as easy to cluster into groups. The two prokaryotes show distinct small frequencies of strand residues at most relative points. However, since the concerned values fall within a range of about 0.002 to 0.008, all noted differences can be viewed as minuscule.

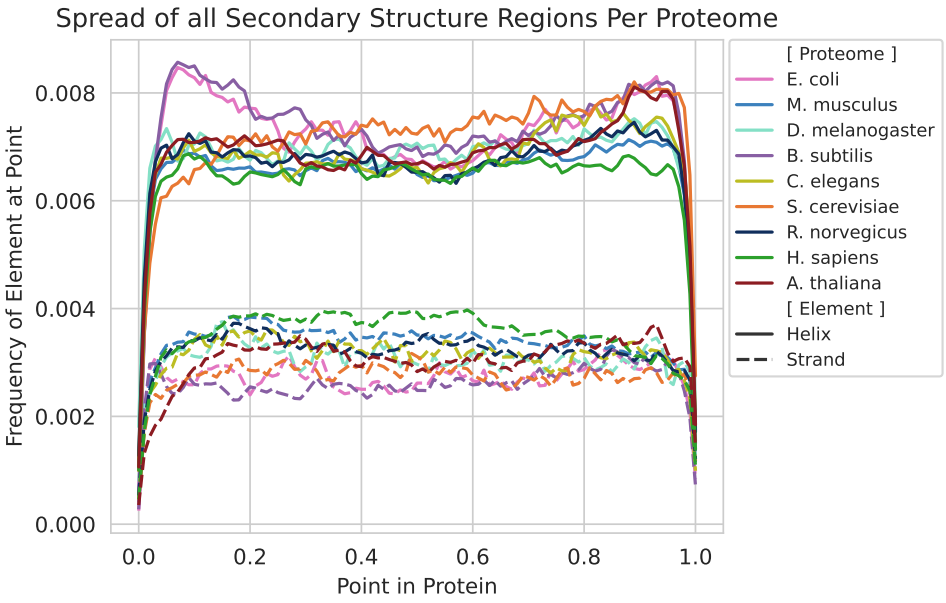

Figure 3.22.: Abundance of **helix and strand residues at relative protein indices** normalized by the total number of helix or strand residues, respectively, in a proteome.

3.4. Binding Sites

Binding Elements in Proteins As a starting point of capturing differences in binding behavior, we reduced the complexity of the feature to the principal affinity of interaction with varying partners. Figure 3.23 visualizes the per-proteome distribution of the presence of at least one DNA, RNA or protein binding region in a protein.

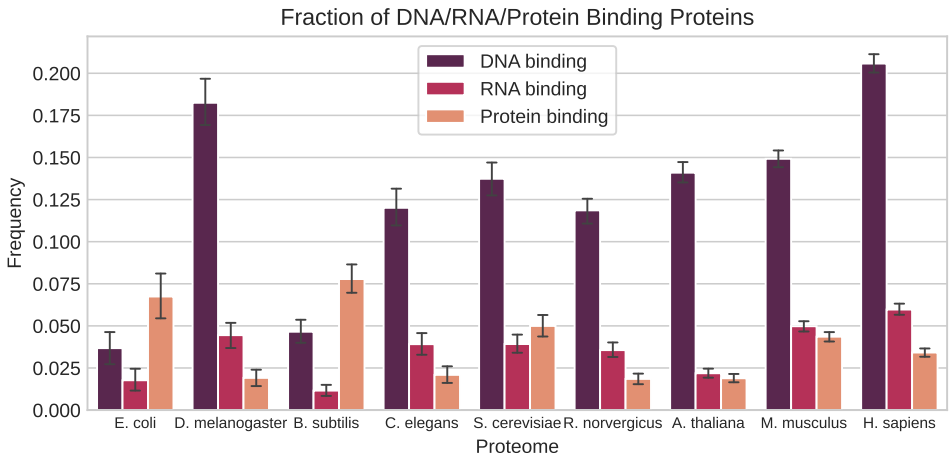

Figure 3.23.: **Fractions of proteins** binding DNA, RNA or other protein molecules for each sample proteome. Error bars indicate 95% CIs.

We observed that on average, eukaryotic proteins were more often involved in binding interactions than their prokaryotic counterparts. Differences in DNA and RNA binding contributed most to this imbalance. In particular, human and fly showed the greatest fraction of DNA-binding proteins. On the contrary to the unbound state of prokaryotic DNA, eukaryotic DNA is organized into compact structures, which include histone proteins. Adding to the actual histones bound to the DNA, this wrapping makes regulation more difficult. The underlying complex mechanisms of interacting with histones and making the DNA accessible require the use of a multitude of DNA-binding proteins. Potentially, eukaryotic chromatin handling explains parts gap between the kingdoms.

In addition, mostly non-overlapping confidence intervals suggest an over representation of protein-binding proteins in prokaryotes in comparison to eukaryotes. Instead of arranging as multi-domain complexes, prokaryotic proteins usually consist of a single domain. Thus, in order to fulfill equally complex tasks within the cell, some interactions require prokaryotic proteins to first make up for their lack of "own" additional functional domains by forming a greater complex. This assembly process requires the involved prokaryotic proteins to be able to bind to each other. We speculate that this variation in basic protein structure within the two kingdoms could explain aspects of the differences in fractions of protein-binding proteins.

While the above findings do apply to all proteomes and thus separate the two kingdoms, we noticed that, likewise to previous parts of our analysis, yeast differentiates from the other eukaryotic sample organisms. In terms of numbers of proteins relative to proteome size, the latter show a profile of DNA-binding $\geq$ RNA-binding $\geq$ protein-binding. For prokaryotes, the order is protein-binding $\geq$ DNA-binding $\geq$ RNA-binding. Yeast deviates from both patterns. More proteins bind other proteins than they bind RNA, yet, DNA-binding proteins remain the greatest fraction. This adds to our observation of the switching behavior of *S. cerevisiae*. With varying analyses and features, the single-cell eukaryote

appears more similar to prokaryotes, its own kingdom or behaves differently all together.

Since the relative number of DNA and protein-binding proteins shows the most significant variation between the two kingdoms, we temporarily excluded binding to RNA molecules from our analysis. We extended our previous statistic by one dimension, investigating the fractions of proteins that bind to either RNA or proteins, neither or both. Figure 3.24 visualizes respective numbers for three sample proteomes.

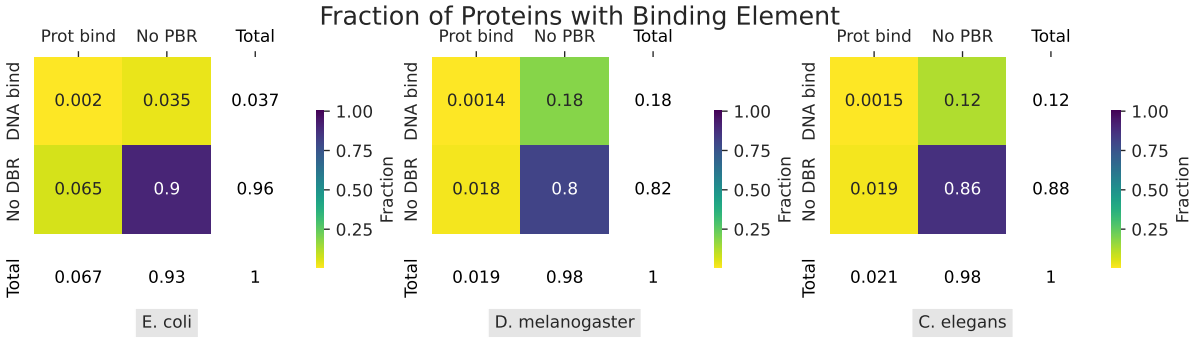

Figure 3.24.: **Presence of DNA and protein binding elements in proteins** among three sample proteomes. The decimal numbers displayed in the three by three matrix indicate the number of proteins, which consist of at least one region of the respective binding category. Numbers are given relative to proteome size. Accordingly, heat maps for all sample organisms can be found in Figure II.12.

As noted before, eukaryotes comprised more binding proteins than prokaryotes. While the latter had greater fractions of protein-binding, DNA-binding was even more distinctly over represented among eukaryotes. In addition, we found that the relative number of proteins exclusively binding DNA molecules differs the most between the two kingdoms, showing noticeably greater values for eukaryotic organisms. Here, the importance of a minimum length cutoff gets demonstrated. For six of the sample proteomes, we investigated the same fractions for the data without previously filtering to binding regions with at least six amino acid residues (see appendix Figure II.13). Without the minimum length restriction, ProNA2019 prediction analysis resulted in greater fractions of DNA binding proteins for all organisms. The increase in the relative number of protein-binding proteins for prokaryotes only was even more noticeable. This shows that introducing a threshold for region lengths affects downstream analyses. Although filtering potential biophysically unsafe predictions of very short regions decreases the chance of drawing unfitting conclusions from the data, it is crucial to perform minimum length threshold determination with caution.

Binding Category Fractions Above, we explored the affinity of proteins regarding different kinds of binding partner molecules. However, ProNA2019 also predicts the location of actual sites involved in the binding process. The concept of feature content per protein, e. g. the number of region residues divided by protein length, allows to soften the definition of distinct binding sites. Adding to our protein-level analysis in the above, the mean content per proteome, as depicted in Figure 3.25, can give further insight into an organism's binding behavior.

Adding to similar observations on the protein level, the value for mean PBR (defined in the legend of Figure 3.25) content among prokaryotes was significantly higher than for most eukaryotes. Simultaneously, eukaryotic proteomes showed significantly greater levels

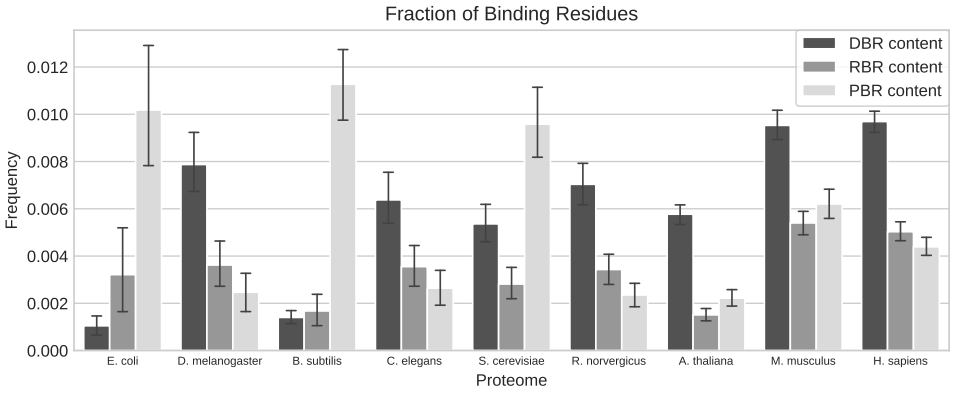

Figure 3.25.: **Fractions of residues** corresponding to three kinds of binding components (DBR = DNA-binding region, RBR = RNA-binding region, PBR = protein-binding region). These are defined as the mean content per protein. We calculated the means per binding category. Error bars indicate 95% CIs.

of mean DBR content. The confidence intervals of RBR and PBR content overlapped for all eukaryotes but one. Again, the profile of mean binding region content for yeast differed vastly from all other organisms, as its mean PBR content was higher than the DNA-binding equivalent. Here, *S. cerevisiae* showed PBR content of prokaryotic levels. This was not true for the fraction of protein-binding proteins (see Figure 3.23), which was rather similar to eukaryotic dimensions. We suggest that PBRs in yeast are either longer, or more often located within the same protein.

In addition, the mean binding content sum for all categories was the most high in mouse and human. The two eukaryotes also had the highest values for mean DBR content, supported by statistical significance. In terms of binding partner preference, mouse did not show the same over representation of DNA-binding as human. Assuming the same logic as for yeast, we conclude that DBRs in *M. musculus* are either longer or more frequent per protein than DBRs in *H. sapiens*.

3.5. Combining Features: Relationship Plots

The distribution of protein disorder, transmembrane helices, secondary structure elements and binding sites all characterize a proteome from different perspectives. Yet, it is clear that these features are far from independent of each other. Certain secondary structure elements might enable the three-dimensional folding of specific binding pockets. A specific protein might lack the energy to maintain a stable structure, but might give up their disordered state upon binding to another molecule. In order to be inserted into the cellular membrane, most proteins require well-structured hydrophobic helices that can cross the lipid bi-layer. Can they still harbor structurally disordered elements? We investigated this and other questions touching on the subject of combining protein properties. The following section presents starting points for a cross-feature analysis in order to further characterize the selected sample proteomes.

3.5.1. Protein Disorder and Transmembrane Helices

As established in previous chapters, transmembrane proteins (TMPs) consist of one or multiple helices domains crossing the cellular membrane. Intuitively, this structural requirement leaves little room for regions that do not adopt a stable three-dimensional form. In Figure 3.26, we visualized simultaneous presence of disordered regions and transmembrane helices for three sample organisms.

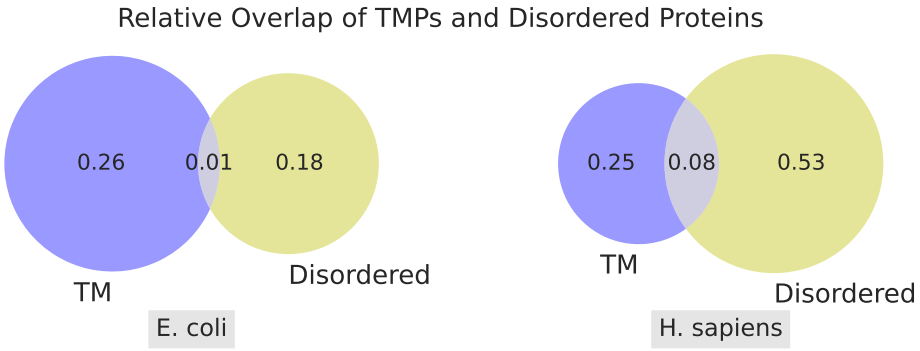

Figure 3.26.: **Presence of transmembrane helices and protein disorder within proteins** for three sample proteomes. The decimal numbers displayed in each segment indicate the number of proteins containing at least one region of disorder (yellow), transmembrane helix (blue) or both (intersection). Numbers are given relative to proteome size. Accordingly, venn diagrams for all sample organisms can be found in the appendix (Figure II.15).

As expected, the intersection of proteins comprising disordered regions (DRs) and TMPs is small, yet, it does exist. The fraction of TMPs was rather similar for proteomes of both kingdoms, which highly differed in their proportion of disordered proteins. As presented earlier (see section 3.1), noticeably less proteins from prokaryotes showed elements of disorder. Thus, less than 50% of each of the prokaryotic proteomes could be displayed in Figure 3.26. We found that the intersection of TMPs and disordered proteins was even smaller for prokaryotes. For instance, only about 1% of *E. coli* proteins contained DRs and TMHs. Since these types of regions cannot overlap by definition, their concurrent presence is influenced by the length of the respective protein. Since the average prokaryotic protein is shorter than its eukaryotic counterpart, less space for the presence of these two

structural opposites is available. This could potentially explain the smaller fraction of intersecting proteins in the prokaryotic proteomes.

3.5.2. Protein Disorder and Binding Sites

One of the key parts of our comparison is the analysis of the distribution of protein disorder (see section 3.1) among the selected sample proteomes. Since structure determines function and disorder describes the absence of a well-ordered structure, the function of DRs constitutes a research area with the potential to disrupt many adopted biological views. In order to investigate the functional activity of disordered proteins, Mészáros *et al.* developed a prediction algorithm for the classification of disordered protein binding regions [10]. Since PredictProtein provides predictions for protein disorder as well as binding sites, we combined these two features. The following subsection will present analysis results and provide comparative information regarding findings of Mészáros *et al.*.

Residue Fractions The combination of a structural feature with a functional feature allowed the definition of another concept. Figure 3.27 provides a visualization of the relationship between the fraction of disordered residues used in binding (DUBs) and the relative number of amino acids within DRs for each proteome.

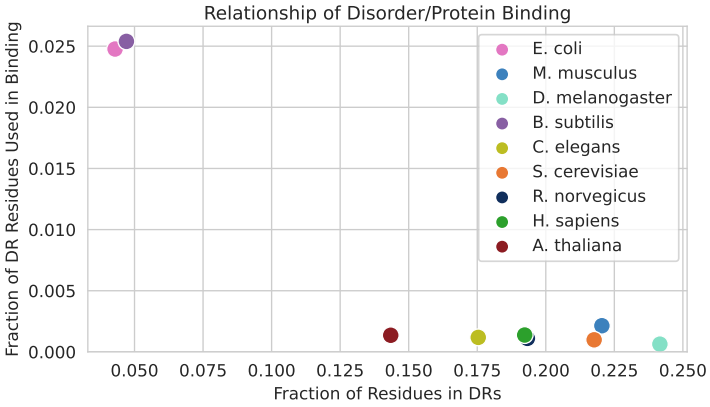

Figure 3.27.: Relationship of **the fraction of disordered residues used in binding** on the y-axis and **the disorder content per proteome** on the x-axis. Concepts are defined in chapter 2.

We found that combining both features in the fraction of DUBs separates the two kingdoms in a particularly solid way. The two sample prokaryotes showed highly similar values that were noticeably greater than for any eukaryote. Adding the disorder content per proteome as a second dimension further distances the two kingdoms. Mészáros *et al.* report comparable values for the analysis of their disorder predictions by IUPred [11] for a variety of Swiss-Prot proteomes. However, applying their developed prediction tool ANCHOR for the location of disordered binding sites, results for the fraction of DUBs vastly differ. While ProNA2019 binding region predictions result in a maximum of about 2.5% for the fraction of DUBs, Mészáros *et al.* present values ranging from about 30% up to over 90%. In addition, the fraction of DUBs is higher for prokaryotes as for eukaryotes instead of vice versa, as reported by Mészáros *et al.*. Since we used the same minimum length threshold for PBRs, the reason for this cannot be a large number of smaller binding regions predicted by ANCHOR. We already include the whole PBR length if it overlaps

with a disordered region. Although their definition of DUBs may differ from ours, it is unlikely to be even wider in terms of residue inclusion. Thus, we further included the analysis of the number of PBRs per DR as well as the distribution of the disordered protein binding region (DPBR) lengths.

Number of Protein Binding Regions Per Disordered Region Figure 3.28 displays the distribution of the number of PBRs per DR. As an overlap criterion we used the definition from Figure 3.27. Notably, whole distributions were highly similar to each other among the

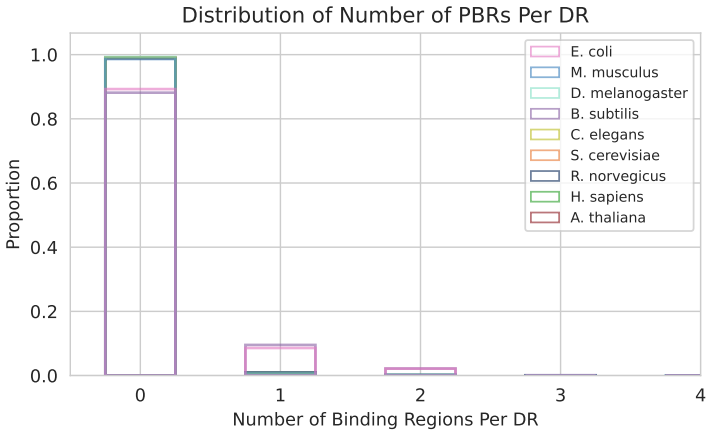

Figure 3.28.: **Distribution of the number of protein binding regions per disordered region.** Bin values are given relative to proteome size. A PBR is part of a DR if they overlap, as defined in the legend of Figure 3.27. The cutoff of four was selected since it was the maximum number of PBRs per DR.

same kingdom. Yet, our data differed from the predictions made by ANCHOR. Mészáros *et al.* report two as the most frequent number of predicted binding sites within long (equal to our definition, at least 30 residues) disordered regions. We found that only about 10% of the prokaryotic DRs contained one or more protein binding sites. In eukaryotes, the respective fraction slimmed down to about 1% of DRs. Thus, our results cannot confirm ANCHOR prediction analysis for either prokaryotic nor eukaryotic organisms. For both eukaryotes and prokaryotes, we found distinctly less PBRs overlapping DRs in our data. Since we only included binding predictions of the highest of three reliability classes, these observations of variation could be explained by a more rigid result filtering as part of our analysis.

Disordered Protein Binding Region Lengths To conclude our analysis of the combination of disorder and protein binding, we compared the distribution of DPBR lengths to predictions made by ANCHOR. Again, we defined an overlap of a DR and a PBR as previously and used the length of the respective binding site prediction. Figure 3.29 shows the binned distribution as provided by Mészáros *et al.*.

Here, our data is distributed approximately similarly to ANCHOR predictions. For smaller region lengths, prokaryotic bin proportions were greater than eukaryotic numbers. This observation was reversed for longer DPBRs. However, we noted smaller differences in proportions between the two kingdoms than ANCHOR. For instance, above 50% of both eukaryotic and prokaryotic DPBRs had between six and eleven amino acid residues, whereas Mészáros *et al.* report less than 45% for eukaryotic DPBRs of that size. In general,

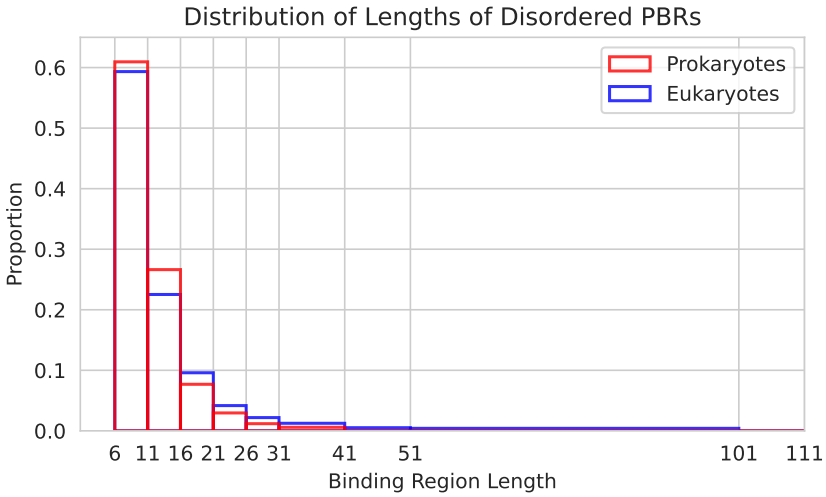

Figure 3.29.: **Distribution of disordered protein binding region (DPBR) lengths** grouped by kingdoms. Binning was performed in accordance to ANCHOR prediction analysis [10]. Bin values are given relative to proteome size.

ANCHOR appears to predict longer disordered binding sites, which partially explained the higher relative fraction as visualized in Figure 3.27. As for previous findings, we suspect that filtering our predictions by reliability class might have affected number and length of analyzed binding sites. Since we only included the most reliably predicted regions, we hope that our analysis had more success at capturing biological realities.

3.6. Discussion of General Findings

Because of the large number of visualizations for multiple protein features, we structured this thesis to include applicable comparisons to previous work directly ensuing the presentation of the respective result. We hoped to ensure a clear assignment of evaluation and results to the respective concepts. Thus, instead of discussing individual results, this section will pick up feature-specific findings and relate them in a unifying context.

Yeast Behaves Differently Than Other Eukaryotes

At multiple points during our analysis, we noted that *S. cerevisiae* showed a distribution that noticeably and oftentimes significantly differed from other eukaryotes. This behavior was noticed across all selected proteome features.

In spectrum plots regarding protein disorder, cross-correlation of pairings of yeast and another organism showed larger values. For instance, eukaryotic pairings involving yeast covered the top part of the eukaryotic zone (Figure 3.30). When yeast was paired with a prokaryotic, the CC line was close to the border to the prokaryotic zone. Here, the single-cell eukaryote showed strong similarities to the behavior of prokaryotic organisms.

This theme appeared for transmembrane helices as well. Yeast had an extraordinarily high fraction of transmembrane proteins (TMPs) with the N-terminus on the cytoplasmic side of the membrane, paired with a lower fraction of TMPs with an extracellular orientation. Compared to the other sample organisms, yeast showed a distribution more similar to the two prokaryotes than the other eukaryotes (Figure II.7). The latter however do contain more TMPs with a signal peptide or where the N-terminus is located within the membrane. For these proteins, yeast showed values close to several other eukaryotes.

We also found that for some concepts, yeast deviated from all other organisms. In Figure 3.22, we investigated the spread of secondary structure elements across the normalized protein. We found that for helices, eukaryotes displayed a rather similar profile, while prokaryotes followed a differentiating distribution that was consistent across the kingdom. The frequency of helices in yeast was as low as for the other eukaryotes for the first quarter of the normalized protein. For the rest of the protein however, yeast separated itself from all other proteomes, before showing values comparable to the prokaryotic organisms for the final quarter of the normalized protein.

The notion of yeast representing eukaryotic as well as prokaryotic qualities was also present within the fraction of binding residues per proteome (Figure 3.25). Here, yeast showed an average DNA-binding region (DBR) content well within the confidence intervals of several other eukaryotes. Simultaneously, the fraction of protein-binding region residues was significantly greater than for all other organisms of the eukaryotic kingdom and instead overlapping with prokaryotic confidence intervals. Thus, yeast resembled both eukaryotes and prokaryotes by displaying a unique profile.

The compatibility of *S. cerevisiae* and prokaryotes as well as its divergence from other eukaryotes has been noted in previous research. During their genomic analysis, Karlin *et al.* found that for predicted highly expressed genes, the genome of yeast paralleled typical bacterial genomes, while still featuring eukaryotic cytoskeletal genes [43]. In the field of recombinant production, Preisler *et al.* recently argued that using *S. cerevisiae* as host for expression results in higher quantities of eukaryotic and prokaryotic transmembrane proteins [44]. Kachroo *et al.* discovered that all enzymes of the heme biosynthesis pathway in yeast could be replaced by prokaryotic orthologs with a merely minimal decrease in growth rates. On the level of proteomes, Ambri *et al.* report the successful application of

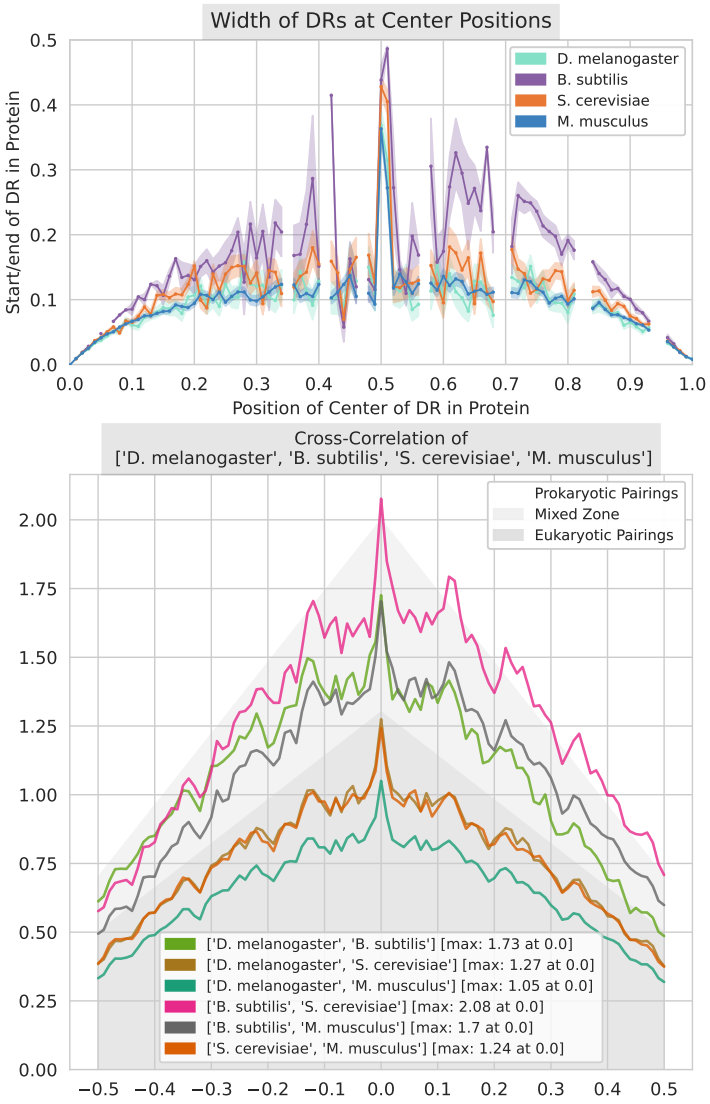

Figure 3.30.: **Disorder spectrum plot** (a) and pairwise **cross-correlation** (b) for four sample proteomes including yeast. **(a)** Means for the distance of relative start/end positions to the DR center were computed for each center position on the x-axis. The shaded area behind each curve indicates the 95% CI. We excluded positions without any present DR centers from the plot. CI and cross-correlation (CC) calculations were executed using pseudo-counts. **(b)** CC curves can be grouped according to the kingdom combination of the respective proteome pair. Corresponding zones are explained in the legend. Spectrum plots for other combinations of four out of the nine sample proteomes can be found in the appendix (Part II).

prokaryotic transcriptional activators as biosensors in yeast [45]. Since we found that the mean protein-binding content of yeast resembles prokaryotic levels, it appears possible that members of the "enormous reservoir of ligand-binding transcriptional regulators" [45] found in prokaryotes can be utilized as biosensors in an organism that shows high similarity in protein binding. In sum, evidence of *S. cerevisiae* resembling prokaryotic characteristics or deviating from the eukaryotic consensus has been generated for an abundance of specific biological properties. Our analysis adds to this by demonstrating this notion for aspects

of several protein features on the level of whole proteome distributions, as congregated in the above section.

Several Concepts Allow the Separation of Kingdoms

Throughout our analysis, we defined and visualized a multitude of concepts that separated eukaryotes from prokaryotes. Certain features and respective forms of visualization made this particularly clear. The underlying protein length distribution (Figure 3.2) massively contributes to the separation according to concepts that make use of information relative to protein length.

For protein disorder, we investigated the distribution of the number of disordered regions (DR) per protein. While eukaryotic proteins contained more DRs on average (Figure 3.4), prokaryotes had a higher proportion of proteins without any disorder. This divergent behavior of the two kingdoms could be captured in the pairwise KL-divergence. While the divergence of the two prokaryotes as well as among the eukaryotes was minimal, distances across kingdoms took noticeably larger values.

In addition to the analysis of the number of DRs per protein, the disorder spectrum concept as displayed in Figure 3.9 further enabled the characterization of each kingdom. Since we found that prokaryotes had relatively wider DRs than eukaryotes, we could capture the pairwise distance of the disorder spectra of two proteomes using cross-correlation. Prokaryotic, eukaryotic and mixed-kingdom pairings produced three different approximate value ranges. Thus, the two-dimensional space created when visualizing CC values could be partitioned into zones according to the kingdoms of the proteomes comprising the respective pair.

When analyzing transmembrane helices (TMHs), further kingdom-separating concepts could be defined. In particular, we could find consistent differences in the proportions for certain numbers of TMHs contained in transmembrane proteins (TMPs) (Figure 3.28, Figure 3.17). Prokaryotes appeared to have a smaller fraction of single-pass TMPs. Yet, they showed greater distributions for 6-TM and 12-TM proteins than eukaryotes. Most of the latter on the other hand contained more single-pass TMPs as well as 7-TM proteins, which we linked to the more numerous existence of G protein-coupled receptors (GPCRs). These separating differences were also visible when performing a pairwise comparison as visualized in Figure II.9.

The fractions of DNA, RNA, and protein-binding proteins per proteome resulted in vastly different profiles for the two kingdoms (Figure 3.23). Overall, the relative number of proteins interacting in binding was lower for *E. coli* and *B. subtilis* than for eukaryotes. The difference was most clear when comparing the much smaller proportion of DNA-binding proteins. In addition, significantly more protein-binding proteins appeared to exist in the two prokaryotes. While some variations within data of the eukaryotic kingdom did occur, these two findings safely separated prokaryotes from eukaryotes.

Notably, this separation of kingdoms in our data and the respective visualizations could get reproduced or even reinforced when combining two features in a concept. Comparing the overlap of the presence of disorder simultaneously to transmembrane helices (Figure 3.26) or relating the fraction of disordered residues used in binding to the disorder content of a proteome (Figure 3.27) resulted in further forms of visualization indicating a distinct characterization of each kingdom.
As shown in the above section, we explored a variety of concepts that allow the separation of eukaryotes and prokaryotes. One of the most fundamental biological principles, being the existence of kingdoms, is preserved in PredictProtein [3] predictions for the selected features.

Thus, the investigated concepts have the potential to be used for the kingdom classification for proteome prediction data for which the true kingdom is unknown. When applied to PredictProtein predictions, the developed forms of visualization and measurements provide a tool for researchers lacking the required data analysis skills for comparing their proteome against other known ones and drawing conclusions regarding the organism of their samples. In chapter 1, we established that a common shortcoming of comparable analyses of multiple proteomes is that the analysis techniques do not get "recycled" in form of a usable tool enabling others to perform similar comparisons. Concurrently, other available tools either do not allow the comparison of multiple proteomes [13], are not available as an easy-to-use website [14], or are restricted to pre-loaded data [15]. In order to provide the opportunity of using our methods for further research, we made use of our analysis scripts for the development of `ppprint`, an extension service for PredictProtein predictions.

4. Dashboards for Proteome Feature Predictions: `ppprint`

PredictProtein [3] is a collection of a multitude of protein feature prediction tools. While for collections of proteins such as whole proteomes, generating predictions requires a local installation of PredictProtein, the service can be accessed online for single sequences. Here, interactive dashboards present the results for several protein features including those selected in our comparison in chapter 3. Few forms of visualizations display basic information such as the location of the feature regions, for instance transmembrane helices, within the concerned protein. Since predictions were produced only for the single sequence, no distributions as for collections of proteins can be calculated. However, for multiple sequences up to whole proteomes and thus using PredictProtein locally instead of the web service, no additional statistics or feature distributions are provided. In this chapter, we introduce `ppprint`[1] as a potential extension for PredictProtein. Although `ppprint` is still being developed, we will present its current state and existing functionalities. First, we will address its purpose, followed by a description of the internal data handling. This chapter will finish with a explanation of how the user can finally interact with `ppprint`.

4.1. Purpose and Features

The aim of `ppprint` is to calculate trends and distributions among whole-proteome data as exemplary performed and presented in chapter 3. Analysis results as well as comparisons of up to four proteomes are then accessible via per-feature dashboards provided by the web application. Users can upload their own data or use available sample proteomes to examine underlying patterns by exploring the provided forms of visualization for protein disorder, transmembrane helices and protein binding sites. Derived insights into how single organisms or groups behave regarding crucial protein features can then enhance knowledge on their relationship and thus aid in research and drug development. Because of its non-final state, the `ppprint` website can currently only be loaded locally following the instructions as given in the git repository. In later stages of development, the web application will be accessible as an online service.

4.2. Implementation

The current implementation of `ppprint` can handle three different scenarios: data upload, creation of a new comparison and the display of the feature dashboards. Inside `ppprint`, the three requests that can occur are processed separately. Figure 4.5 provides an overview of the sequential handling of each type of request made by the user. Incorporating exclusively scripts written in Python, we built `ppprint` using the Django web framework

[1] `https://git.rostlab.org/orlishausen/ppprint`

[46]. In the front-end, we used Bootstrap [47] for polishing of our HTML templates. It is important to note that here, we describe the internal processes, while possibilities available to the user will be explained in the subsequent section.

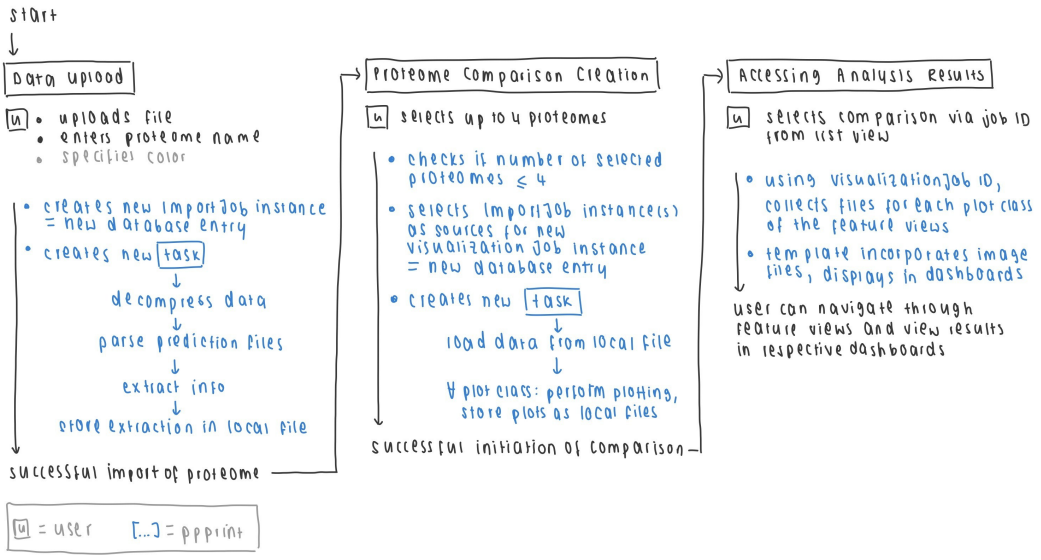

Figure 4.1.: Overview of how `ppprint` handles requests the user can make. Arrows indicate a typical run-through when using the application. Blue writing represents internal `ppprint` handling.

Data Upload Upon upload of PredictProtein prediction data, a new row in the database will be created, which is represented by an instance of the Django model `ImportJob`. Besides creation date and job status, the model references the proteome name as well as an associated color for the display. In addition, the upload of data causes the creation of a new task, which will be handled by the worker. Here, we use Celery as a task queue. Within the new `run_import_job()` task, `ppprint` parses the provided feature prediction files and extracts all data required for downstream analysis. Finally, the resulting data, grouped by feature and base (per-protein or per-region information), is stored locally in per-proteome files.

Proteome Comparison Creation When the user requests a new comparison of proteomes, the request is handled by a `SelectionForm`. The latter incorporates a validator setting the maximum number of proteomes for selection to four in order to prevent the unintended visual overloading of provided figures. The user selected one or multiple `ImportJob` model instances as sources for the newly created `VisualizationJob`. Likewise to the data upload process, a new task gets created and handled by the task queue. The `run_visualization_job()` task loads the required data from the previously locally stored files. After data collection, `ppprint` performs plotting for every plot class that was specified internally, each mapping to exactly one image to be displayed. The resulting `.png` plot files are stored locally.

Display of Analysis Results in Dashboards Following the selection of a specific `VisualizationJob` via its identifier representing the internal primary key, the image file for each plot class gets collected. The files get handed to the HTML template corresponding to one of the possible result views: a feature-independent overview, as well as the detail

pages for protein disorder, transmembrane helices, and protein binding sites. The template incorporates all image files that belong to the respective view type and displays them all in form of simple dashboards.

4.3. Usage

Here, we describe how the user can interact with `ppprint`. Since currently, the service is not yet available as a externally hosted website, instructions on how to set up `ppprint` locally can be found in the `README.md` file located in the git repository.

Uploading Data

In order to make PredictProtein predictions available for comparison, the user has to upload them via the upload page (Figure 4.2), which can be reached using the navigation bar. Since for most proteomes of model organisms, Gigabytes of prediction files may have been generated, the user can only upload a compressed `.tar` file with an extension such as `.xz`, `.gz`. For a successful data import, the user also has to enter a name and may pick a color for the display of analysis results. When the user has selected a file and filled out the remaining two fields, the data can be submitted for internal parsing and extraction.

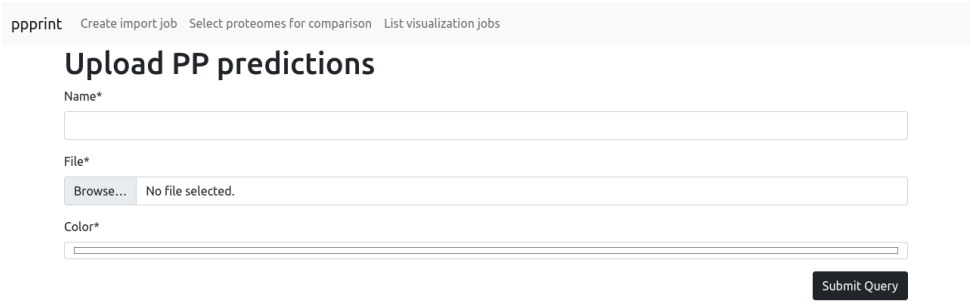

Figure 4.2.: Data upload view as currently implemented in `ppprint`. The user can enter a name for the proteome to be uploaded, select the compressed file and specify a color for plotting.

Comparing Proteomes and Viewing Visualization of Analysis Results

The success of an upload can be verified via the proteome selection page (Figure 4.3). When the specified name of the uploaded proteome appears in the list of proteomes selectable for comparison, data import and extraction has finished successfully. The user can then choose one or multiple proteomes and submit them for analysis. Currently, the maximum number of proteomes that can be simultaneously selected is four, ensuring that there is no visual overload in the final plot figures.

Once the user has submitted a set of proteomes for analysis, an overview of the created comparisons can be accessed by navigating to the listing page (Figure 4.4). Here, the creation date of the job and its status, which can be "created", "running", "success", and "failure" are given for each created job. The user can then request the final visualization of analysis results by clicking on the respective comparison identifier, internally corresponding to the primary key of the `VisualizationJob`. If the status of the job is not "success", the list of plots ready to be displayed may be incomplete and either requires a page refresh

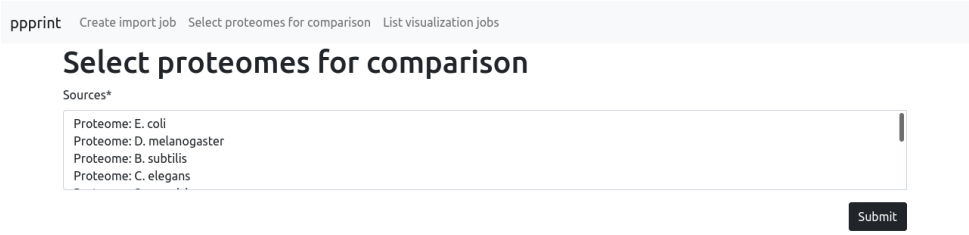

Figure 4.3.: Selection of proteomes for the creation of a new comparison as currently implemented in ppprint. The user can select up to four proteomes to be included in the analysis.

("created"/"running") or starting a new comparison ("failure"), for analysis results to be shown completely.

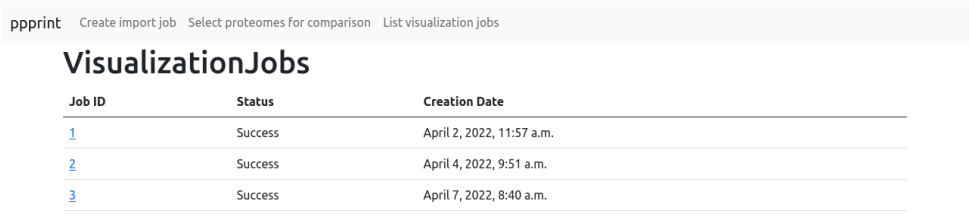

Figure 4.4.: Listing of created comparisons as currently implemented in ppprint. Apart from the job ID, this listing provides the status as well as creation date of the internal job. The user can access the results of a specific comparison by clicking on the respective identifier in the first column.

Finally, the results of the analysis performed by ppprint can be accessed via the detail result pages. The overview tab (Figure 4.5) displays feature-independent properties such as protein lengths and sample sizes of the selected proteomes.

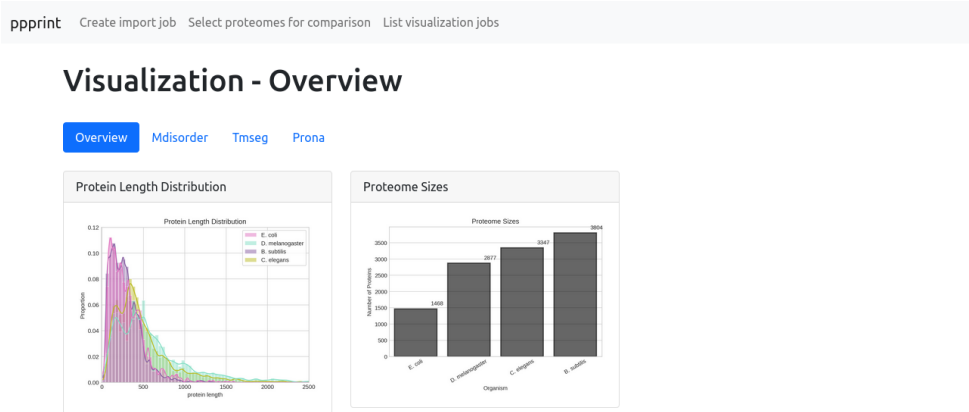

Figure 4.5.: Overview of feature-independent results as currently implemented in ppprint. This includes the protein length distribution as well as the sizes of the selected proteomes.

Using the tab bar, the user can navigate to the feature-related detail pages (Figure 4.6, Figure 4.7, Figure 4.8) in order to view the produced forms of visualization assembled as basic dashboards. Since ppprint is still in development, user-guiding indications such as plot-specific help texts for interpretation and the introduction of required concepts are yet to be added.

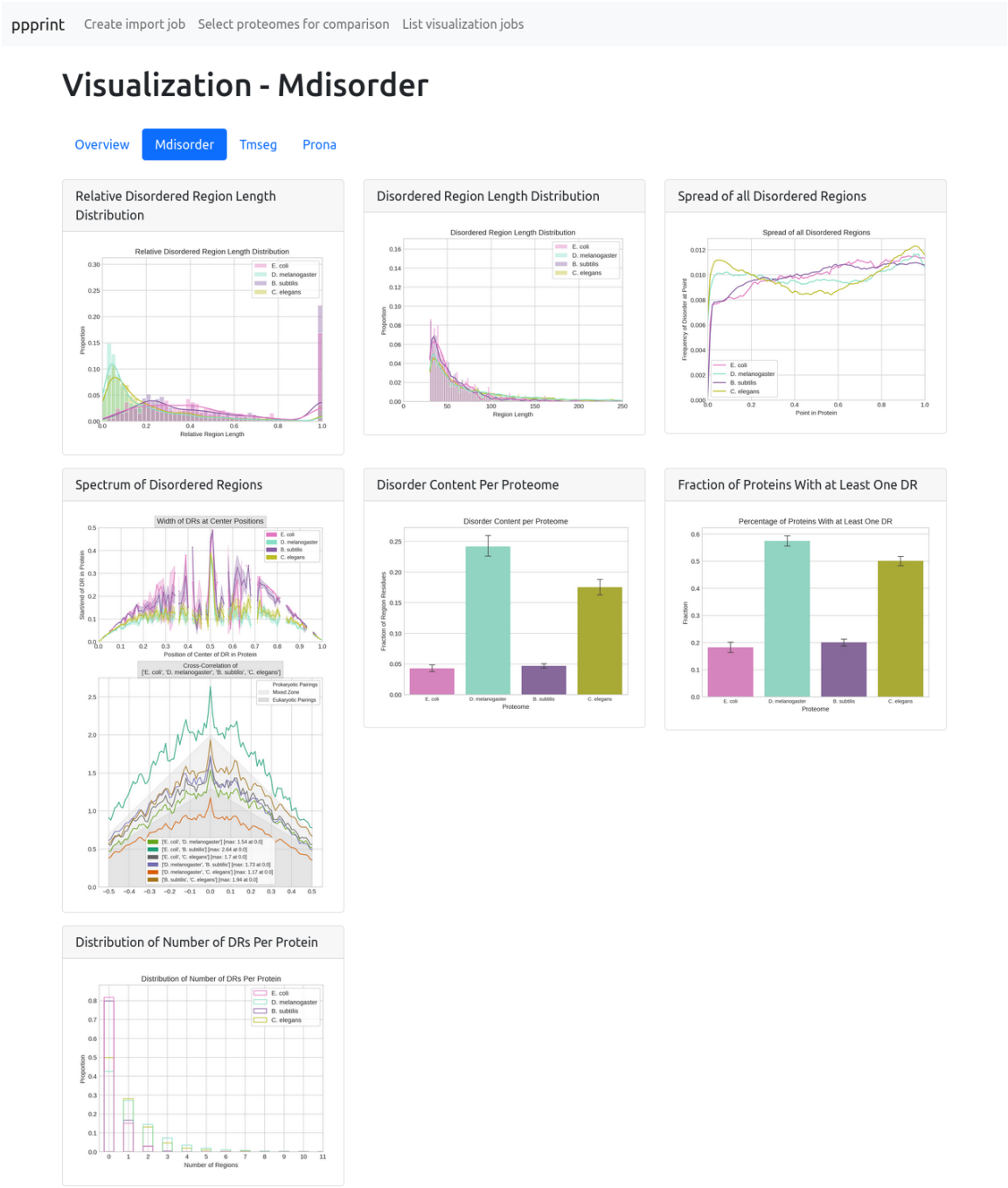

Figure 4.6.: Analysis results for protein disorder as currently implemented in `ppprint`. This includes the distribution of absolute and relative disordered region (DR) lengths, the spread of all DRs, the disorder spectrum with pairwise cross-correlation, the disorder content per proteome, the disorder composition per proteome as well as the distribution of the number of DRs per proteome.

Figure 4.7.: Analysis results for transmembrane helices (TMHs) as currently implemented in `ppprint`. This includes the distribution of absolute and relative TMH lengths, the spread of all TMHs, the proportions of transmembrane protein (TMP) orientations, the proportions of protein classes, the TMH content per proteome, the topological fractions as well as their per-proteome distributions, the distribution of the number of TMHs per protein, and distribution of the number of TMHs per TMP incorporating orientation information.

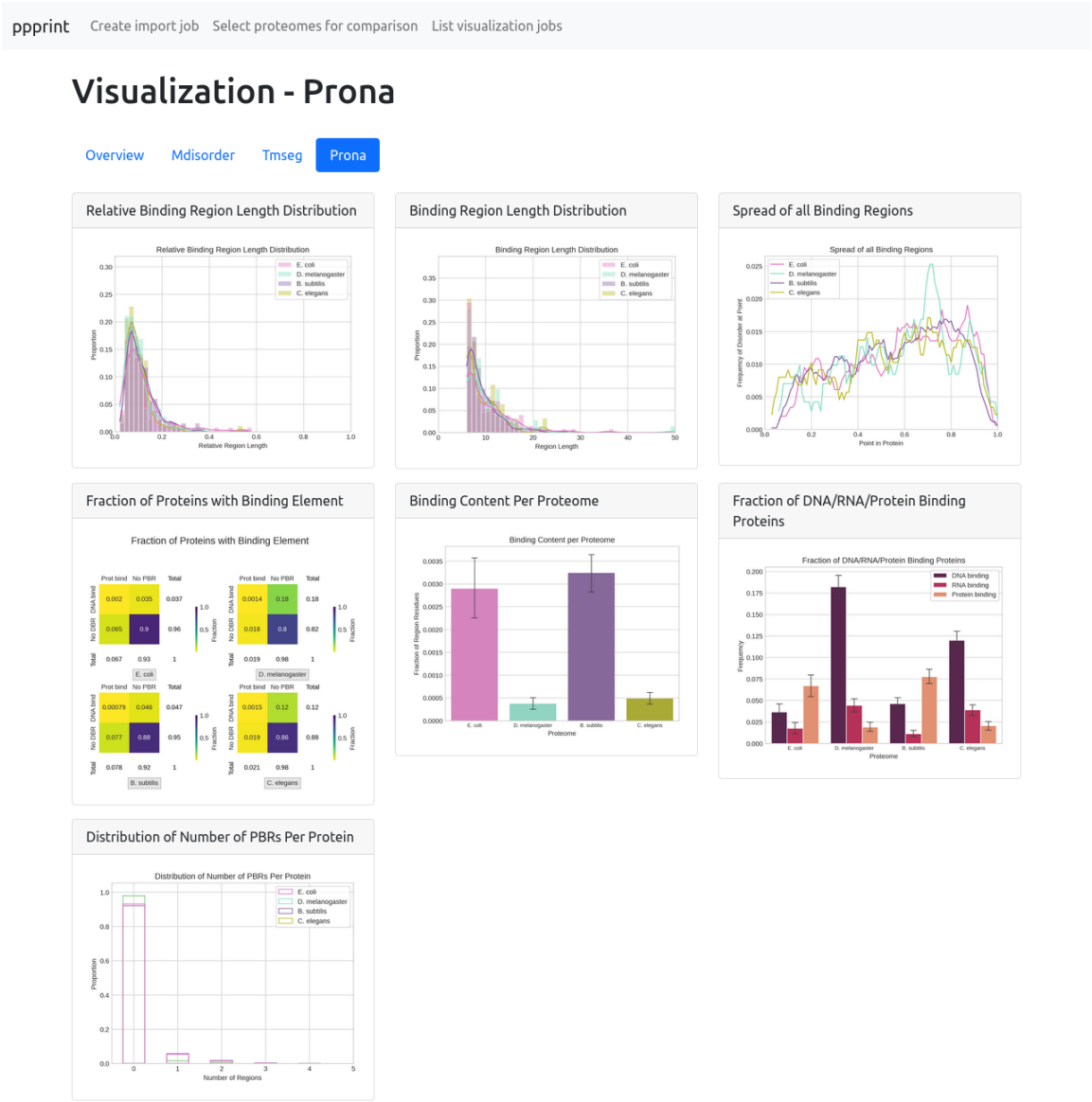

Figure 4.8.: Analysis results for protein binding sites (PBRs) as currently implemented in `ppprint`. This includes the distribution of absolute and relative PBR lengths, the spread of all PBRs, the fraction of proteins with binding elements in two different forms, the PBR content per proteome as well as the distribution of the number of PBRs per protein.

5. Conclusion and Future Work

This chapter will summarize the major points made in this thesis. In addition, it will provide an outlook on further investigations within the proteome comparison as well as future work on improvements for `ppprint`.

5.1. Concluding Remarks on the Analysis and Visualization of Sample Proteomes

Summary

Preceding the investigation of protein features, we performed analyses for bias identification. We established that among the nine selected sample organisms, data set size varies highly. That is why numbers and proportions in all produced plots are given relative to the total number of protein or regions in a protein, depending on what is applicable. Since we make use of several concepts that relate indices or lengths for regions to protein length, we also explored the protein length distribution. As prokaryotic proteins are on average significantly smaller in length, the underlying distribution contributes massively to the separation of prokaryotes and eukaryotes at several points of our analysis. Across all protein features selected for the comparison, we investigated a multitude of concepts that partially or fully characterize each kingdom. Concurrently, yeast deviated from the eukaryotic consensus on multiple occasions while displaying either unique distributions or even stronger similarities to the prokaryotic kingdom. We put these observations into the context of existing research regarding the compatibility of prokaryotes and *S. cerevisiae*.

Protein Disorder For protein disorder, we defined a variety of concepts. Using the disorder composition, we could confirm findings of previous research that relate composition to the level of complexity of an organism. The same trend was found for the disorder content per proteome, however we found contrary results when applying another form of normalization. Additional to disorder composition, the absolute length of disordered regions (DRs) increased with organism complexity. When using relative lengths, building on the previously defined concept of relative indices, calculating the distribution revealed that significantly more DRs span their protein in prokaryotes than among eukaryotes. This as well as greater width of prokaryotic DRs relative to protein length could also be observed with disorder spectra. We found that comparing the latter by calculating the pairwise cross-correlation (CC), the two-dimensional space taken up by the full-length CC can be partitioned into zones, depending on the kingdoms of the involved proteomes. Likewise, the spread of all DRs allowed the assignment of proteomes by distribution similarity to two eukaryotic groups and the prokaryotic duo. In terms of number of DRs per protein, the distributions were highly similar within kingdoms. As a measurement, the Kullback-Leibler (KL) divergence supported this separation of eukaryotes and prokaryotes by taking noticeably higher values for inter-kingdom proteome pairs.

Transmembrane Helices Using transmembrane helix prediction data, we found that the fraction of transmembrane proteins (TMPs) varies more among than between kingdoms and, on the contrary to several aspects found when investigating protein disorder, cannot be related to complexity. Yet, the higher fraction of single-pass transmembrane helices (TMHs) separates the eukaryotic kingdom from prokaryotes. In addition to classifying proteins, we reduced predictor biases arising from a classification according to segments by calculating the distribution of the transmembrane content per protein. Here, we found that prokaryotic and eukaryotic distributions noticeably differed from each other, in particular regarding the location of the maximum bin proportions. Since the TMH content of the average prokaryotic protein was almost double the number for its eukaryotic counterpart, the fraction of "membrane" residues in prokaryotes was significantly higher as well. In terms of topological fractions, prokaryotic organisms additionally had relatively less "outside" residues. We recognized this trend in the distribution of TMP orientations, which also differed vastly for the two kingdoms, with the exception of *S. cerevisiae*. Eukaryotes and prokaryotes could be safely discerned by their fraction of extracellularly oriented TMPs. When analyzing the distribution of the number of TMHs per protein, we could confirm hypotheses made in previous research. Generally, we found that the average prokaryotic TMP contained one helix more than the eukaryotic equivalent. Besides, the proportions of specific numbers of TMHs per protein discerned the two kingdoms particularly clearly. Upon the addition of topology information, we suggested linking our observations to biological realities. Analyzing the distribution of TMH lengths revealed the absence of clear differences between kingdoms or even single organisms regarding this property of transmembrane helices.

Secondary Structure As part of our analysis of secondary structure, we noted that the fraction of proteins containing at least one helix or beta strand were highly similar across all sample organisms. The observation of eukaryotes comprising less protein without strand was noticeably more distinct when performing the same analysis for cached prediction data. Using per-residue predictions instead of segment-dependent classification of proteins, we found that prokaryotes had a significantly higher mean helix content and contained less residues predicted as "other". The latter finding could be substantiated when relating helix and strand content per protein. Here, prokaryotic proteins were less likely to concurrently show low values for both helix and sheet content. Among eukaryotes, two slightly differing clusters formed. We could not find any such grouping for the spread of strand regions across the normalized protein. However, the spread of helices distinctly separates the two kingdoms, apart from yeast. By displaying a unique profile, the single-cell eukaryote was again the exception.

Binding Sites In terms of the presence of binding elements in proteins, we again found clear differences between the two kingdoms. In general, prokaryotes demonstrated distinctly less proteins containing sites for the interaction with ligands than eukaryotes. Additionally, the latter consisted of more DNA-binding proteins relative to proteome size, while prokaryotes showed a higher fraction of protein-binding proteins. Again, the profile of fractions for yeast was unique among all sample organisms. We also noted that adjusting the minimum length cutoff for recognizing a binding site as valid strongly affects the presence of protein-binding regions (PBRs) in prokaryotic proteins. Fittingly, the fraction of PBR residues per proteome was much higher for prokaryotes as well as yeast, distinctly varying from other eukaryotic levels.

Combining Features We supplemented our analysis of protein disorder, transmembrane helices, secondary structure and binding sites by analyzing the combination of certain features from our selection. This method of combining properties that capture divergent protein aspects enabled the demonstration of the small, yet existing overlap of disordered and transmembrane proteins. Moreover, it allowed the definition of the concept of the fraction of disordered residues used in binding. Since prokaryotes had noticeably higher values for this fractions, this new concept could distinctly separate kingdoms. Here, we showed that combining features has the potential to find ways of finding even clearer differences between organisms, or in this case, kingdoms. Although we applied the widest definition of an overlap of DRs and PBRs, our data displayed much less predicted disordered protein binding regions (DPBRs) than comparable research. Among kingdoms, binning proportions showed strikingly strong similarities. The distributions of DPBR lengths for eukaryotes and prokaryotes rather resembled the profiles stated in existing research. Present differences were linked to the supposedly varying strictness applied for accepting binding site predictions.

Outlook

The extensive analysis performed in the context of this thesis still harbours several possibilities for expansion. Using existing data, we will continue researching the effects of different cutoffs for the minimum length of regions and investigating novel combinations of features. Trivially, new relevant results will most probably be provided by adding additional organisms to the comparison, such as by expanding the set of prokaryotes. Alternatively, comparing existing results to the distributions of some form of random proteomes, sampled from all available proteins, could help estimating the significance of analysis results. Since PredictProtein offers the prediction of many further protein features, analyzing data of other structural or functional qualities remains possible. In the future, we hope to conduct further analyses using `ppprint`.

5.2. Current State and Future Possibilities for `ppprint`

Summary

Currently only running locally, `ppprint` is a [46] web application providing functionalities to analyze and compare PredictProtein [3] prediction data for multiple proteomes. The user can upload his or her own predictions and select up to four proteomes at once for a comparison. After successful completion, analysis results can be accessed via feature-related dashboards, which provide several forms of visualization displaying various aspects of the protein properties. As of April 2022, the majority of plots we presented in this thesis has been integrated into `ppprint` already. Thus, `ppprint` can extend the PredictProtein service and provide a way of gaining easy access to the analysis of predictions for collections of proteins.

Future Work on Extending `ppprint`

On the level of plotting details, we plan on enabling error bars indicating 95% confidence intervals in histograms as soon as possible to attest for the statistical significance of results. In addition, we want to include visualizations of concepts that combine features such as given in chapter 3, which has not been possible so far because of the restricted time frame

of this thesis. Potentially, `ppprint` will also enable the user to request the grouping of organisms per kingdom as specified for each proteome upon data upload. In the front-end, a more elaborate dashboard design with interactive functionalities will be implemented as well. Prior to an official suggestion of `ppprint` as an extension for PredictProtein, extensive testing with a high coverage of code will be required. Finally, we will supply `ppprint` as a web service by hosting it externally. Thus, we can omit local installments and enable a straightforward usage for researchers that lack programming experience.

Part II.

Appendix

Interfacing the PredictProtein Cluster

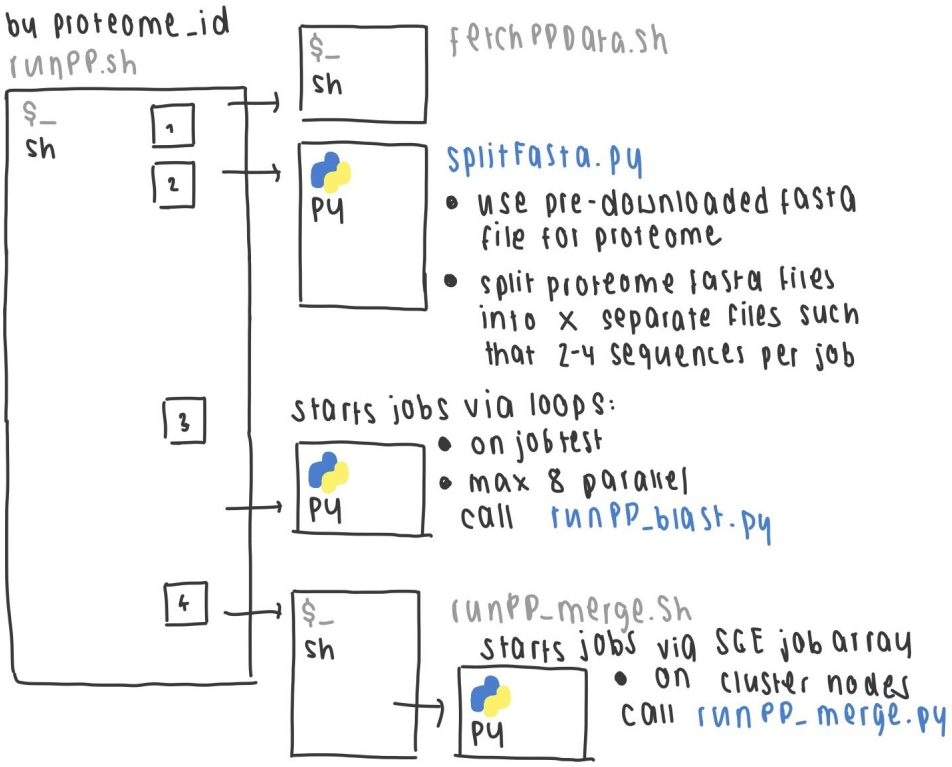

Figure II.1.: An overview of how we interfaced the PredictProtein cluster in order to generate predictions for one proteome. From the regulating script `runPP.sh`, we first fetch the local data base with `fetchPPData.sh`. Subsequently, we split the input `.fasta` file containing all protein sequences of the proteome into smaller `.fasta` files of two to five sequences to increase prediction speed by parallelization. Thus, up to eight of these files could be used simultaneously to generate multiple sequence alignment data via `runPP_blast.py`. Upon alignment completion, jobs executing the prediction for the two to five sequences per `.fasta` file get created. Depending on proteome size and number of sequences per file, the number of `runPP_merge.sh` jobs ranged from 1017 for *E. coli* to 5093 for *H. sapiens*. Within these jobs, new PredictProtein predictions get generated using `runPP_merge.py`. In order to save computation time for the largest proteomes, we extracted cached predictions for mouse and human, which we justified by the recency of the creation date of existing predictions.

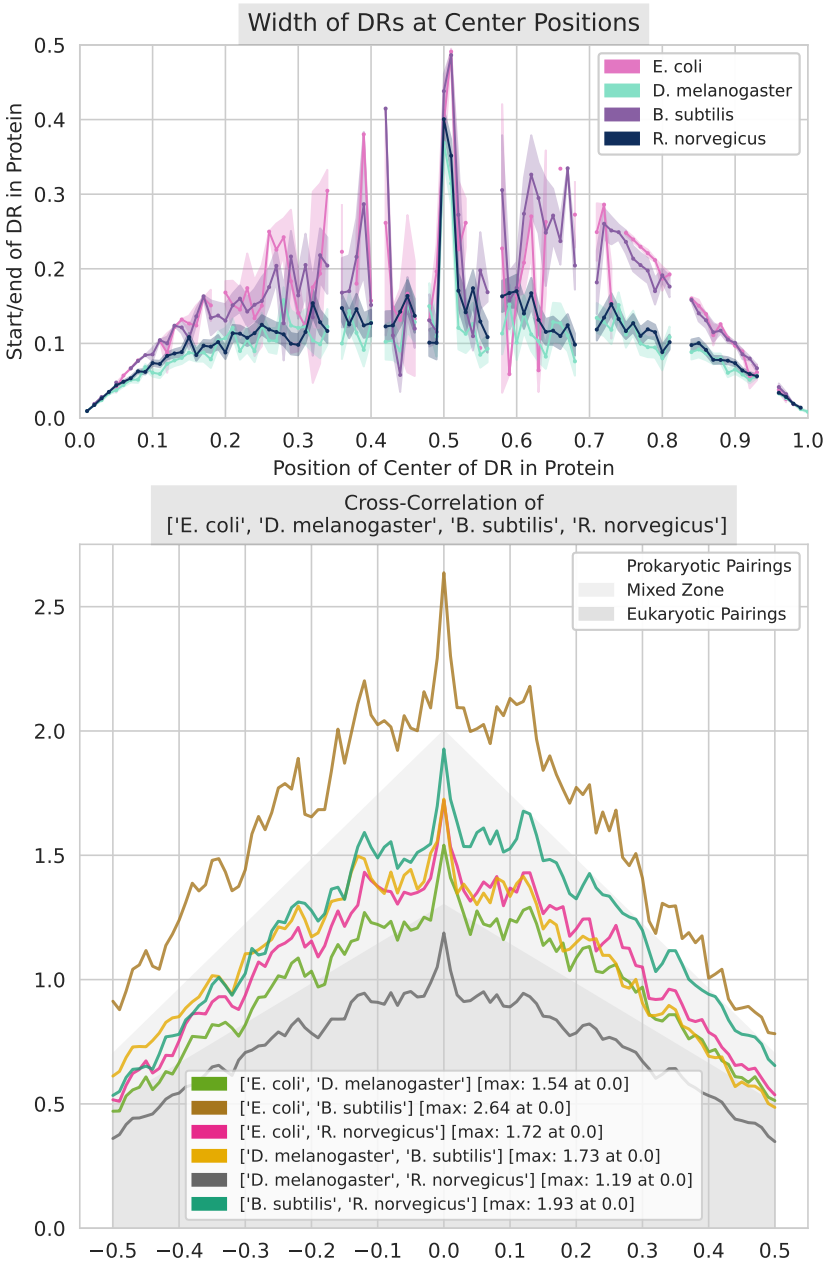

Figure II.2.: **Disorder spectrum plot** and **pairwise cross-correlation** for four sample proteomes. For a detailed description, see Figure 3.9, Figure 3.30. Spectrum plots of other combinations of four out of the nine sample proteomes can be referred to in the spectra directory at `https://git.rostlab.org/orlishausen/proteome-analysis`.

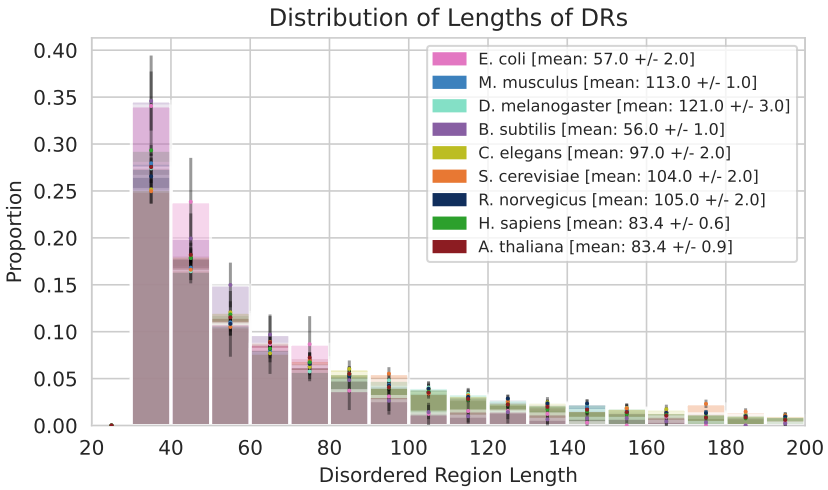

Figure II.3.: **Absolute disordered region lengths.** Binned values are relative regarding proteome size. Error bars indicate 95 % confidence intervals.

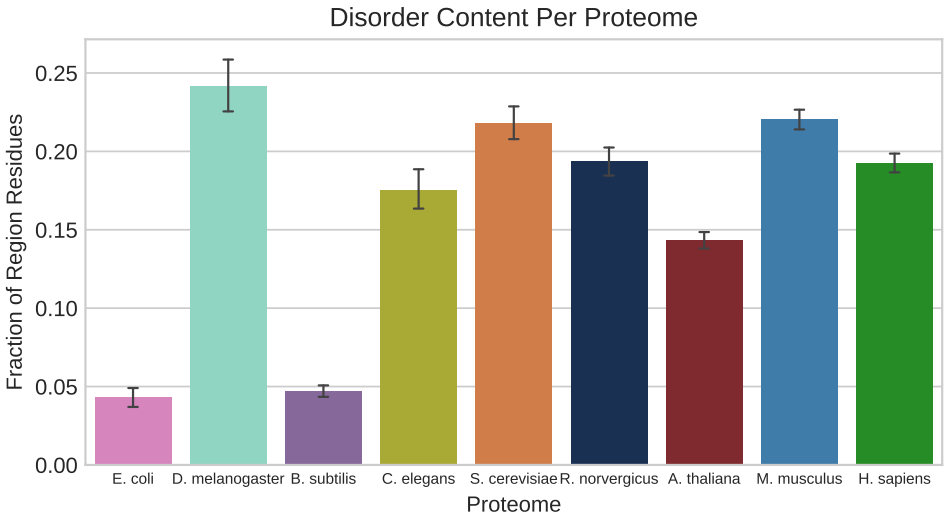

Figure II.4.: **Disorder content** per proteome. Here, we define the disorder content per proteome as the number of residues located in disordered regions (DRs), divided by the total number of residues of all proteins in the proteome. Error bars indicate 95 % confidence intervals.

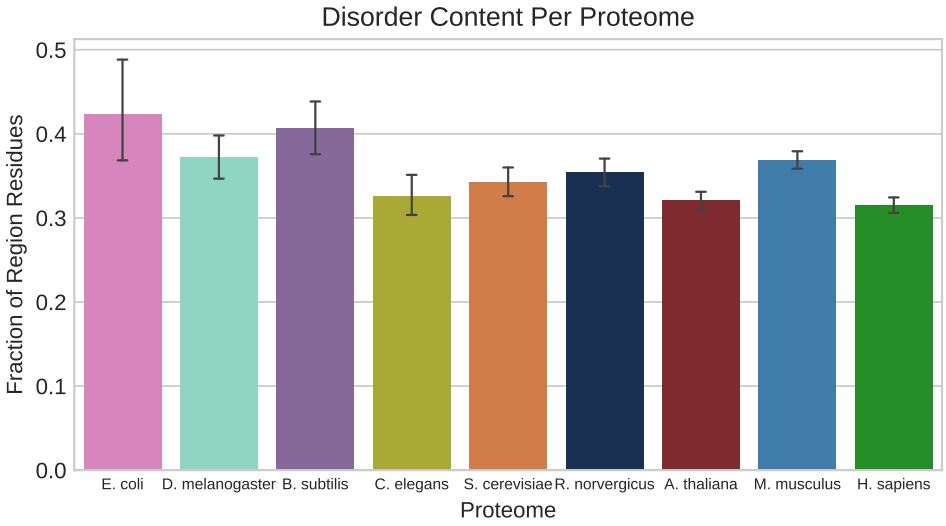

Figure II.5.: **Disorder content** per proteome. Here, we define the disorder content per proteome as the number of residues located in disordered regions (DRs), divided by the total number of residues in all proteins containing DRs in the proteome. Error bars indicate 95 % confidence intervals.

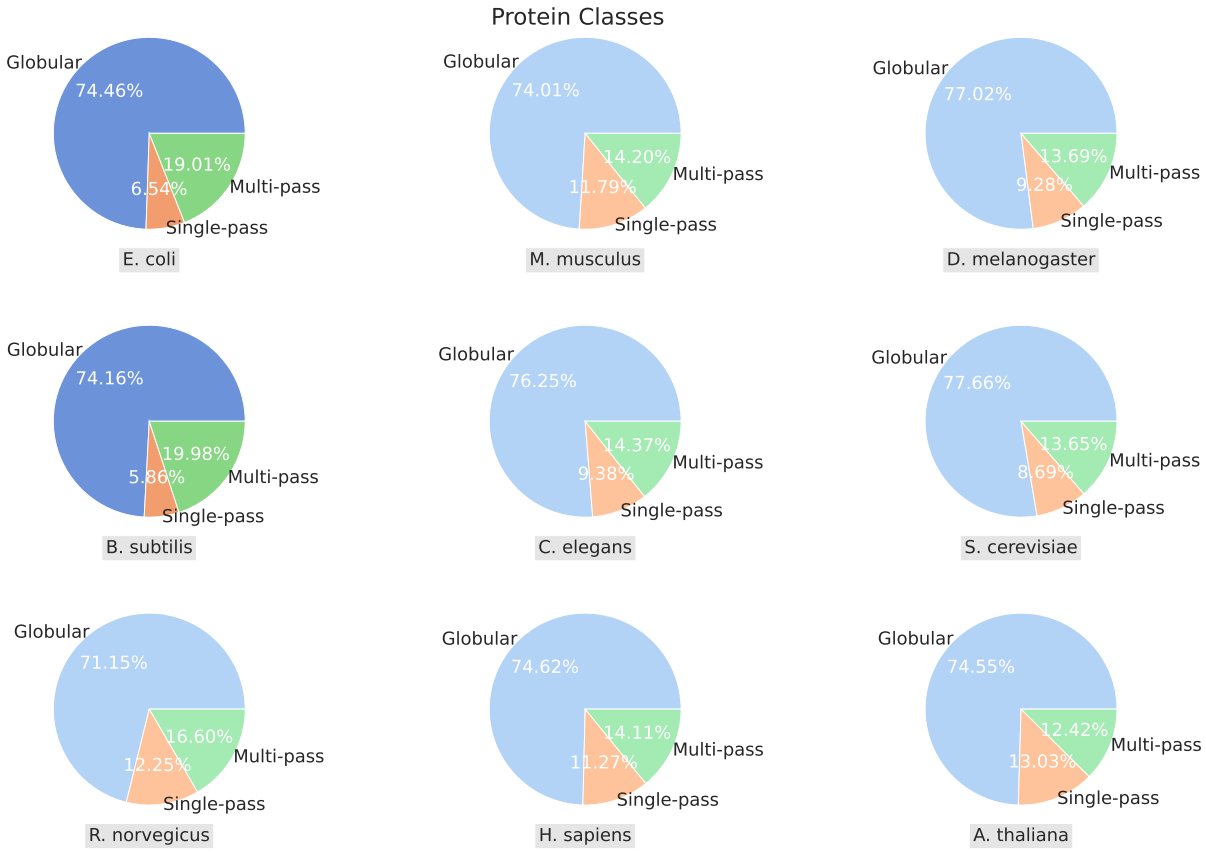

Figure II.6.: **Proportions of protein classes** among all sample proteomes. We used darker colors for prokaryotic data.

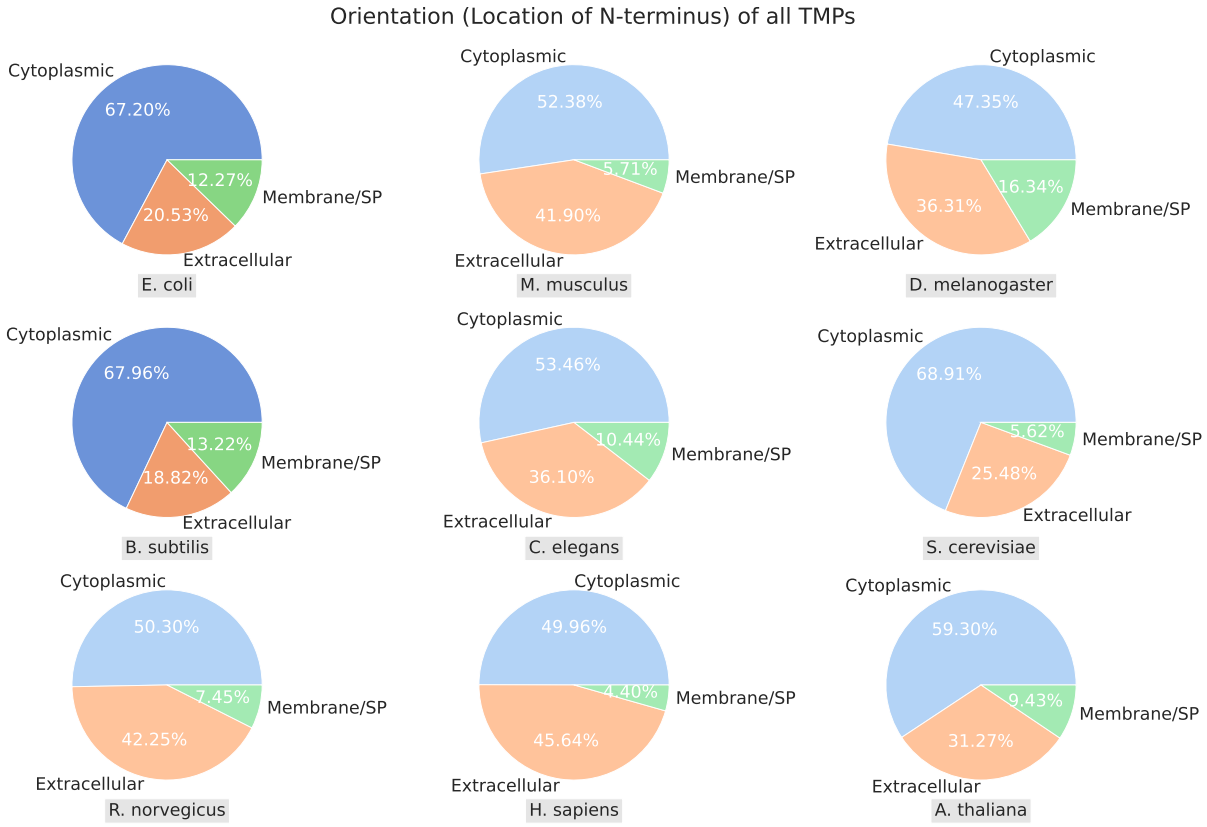

Figure II.7.: **Proportions of TMP orientations** among all sample proteomes. We used darker colors for prokaryotic data.

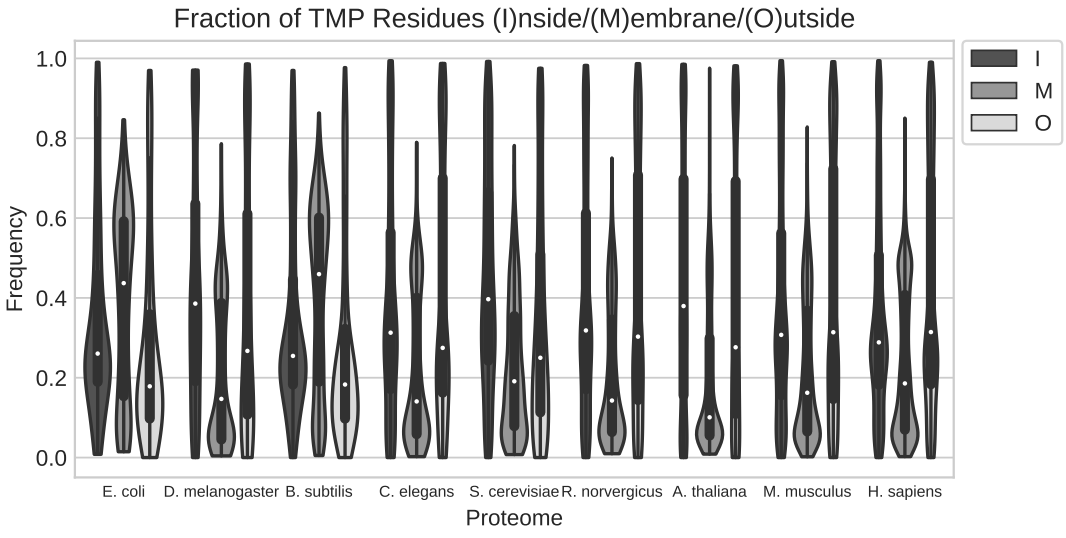

Figure II.8.: Distribution of the **fractions of residues** corresponding to TMP topology. These are defined as the number of residues per protein predicted to be part of the cytoplasmic (inside), membrane, or extracellular (outside) component, divided by the protein length.

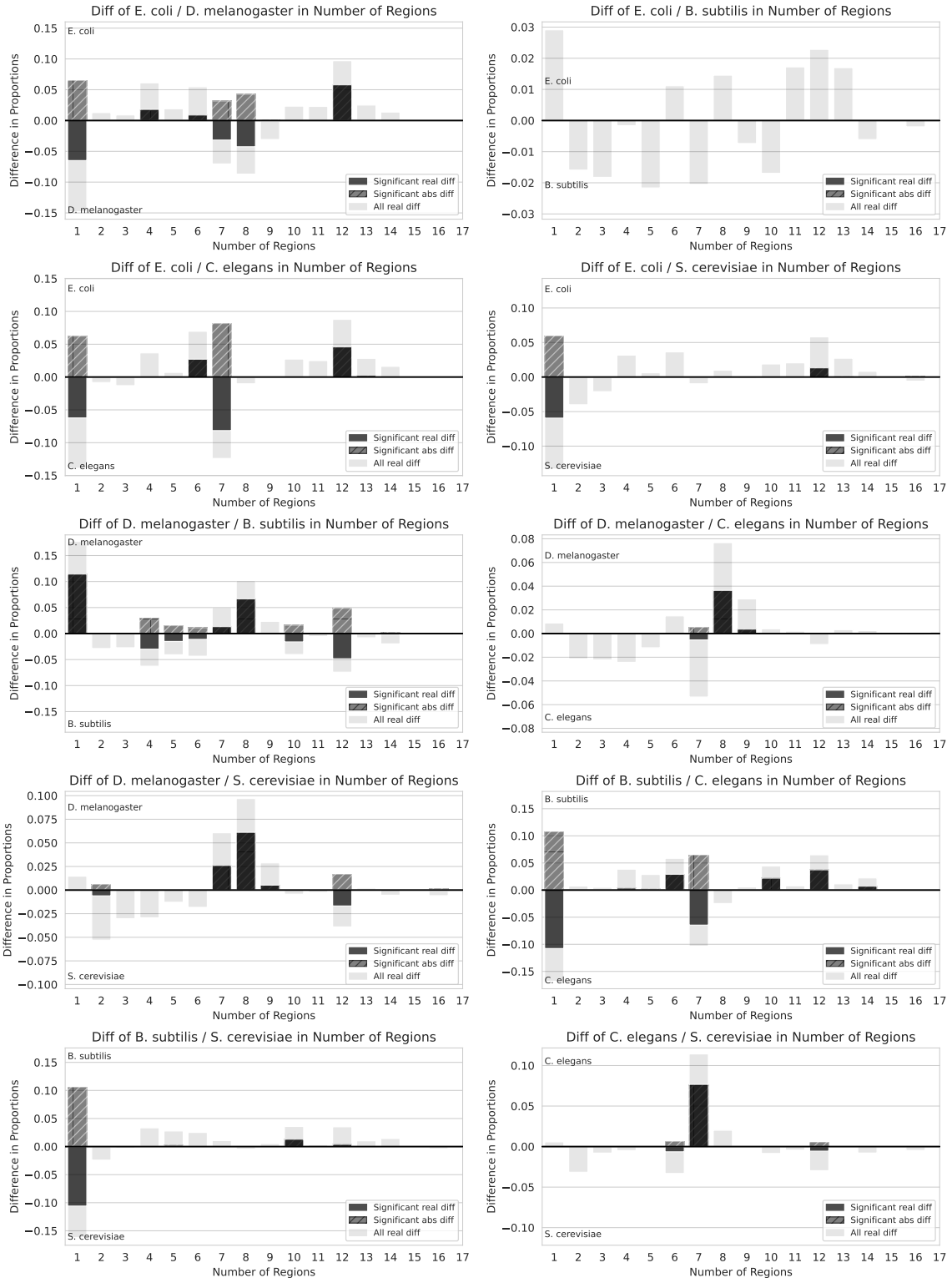

Figure II.9.: Difference in number of TMHs per protein for pairings of sample proteomes. Dark bars indicate significant differences, striped shading represents the absolute value of significant differences. All other differences in proportions within the 95 % CIs of both proteomes for the respective bin are given in light grey. Plots visualizing the pairwise differences for other pairings of sample proteomes are located in the diffs directory at `https://git.rostlab.org/orlishausen/proteome-analysis`.

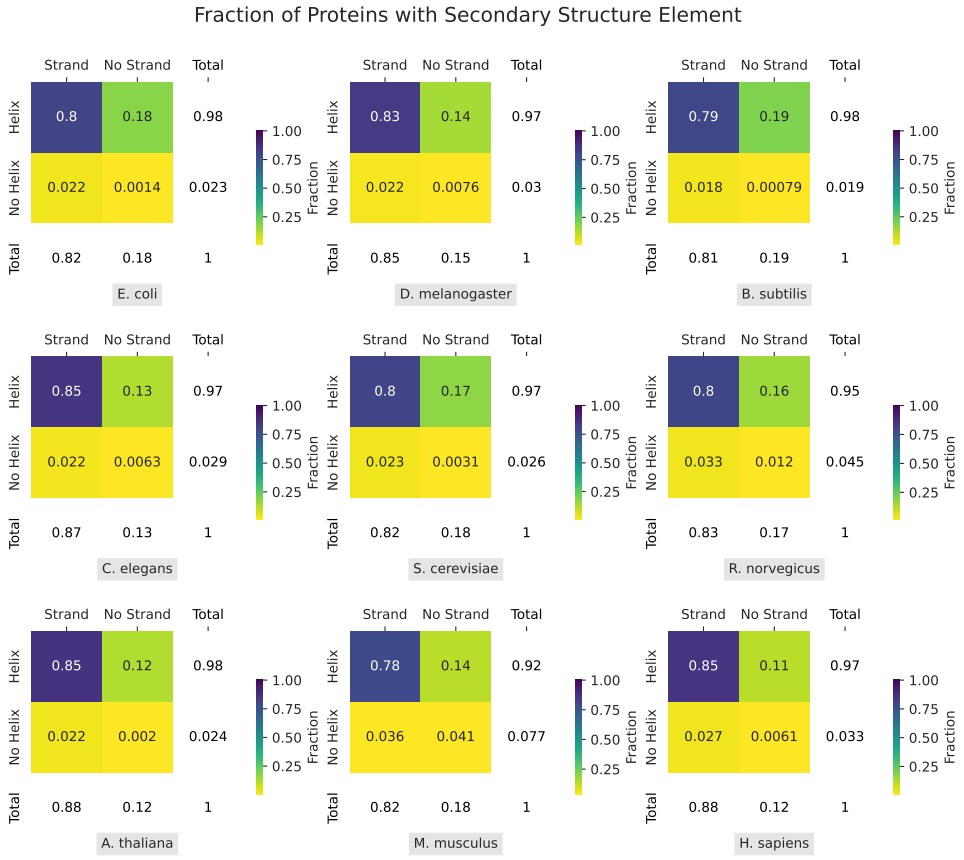

Figure II.10.: **Presence of secondary structure elements in proteins** among all sample proteomes. The decimal numbers displayed in the three by three square indicate the number of proteins, which consist of at least one sheet or helix region, respectively. Numbers are given relative to proteome size.

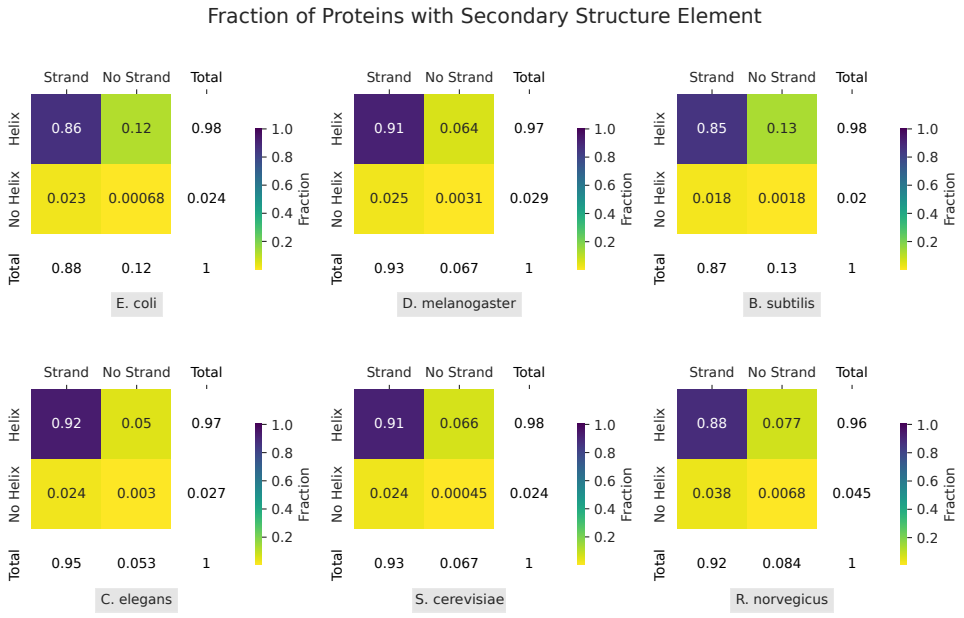

Figure II.11.: **Presence of secondary structure elements in proteins** for cached results of six sample proteomes (as also given in Figure II.10).

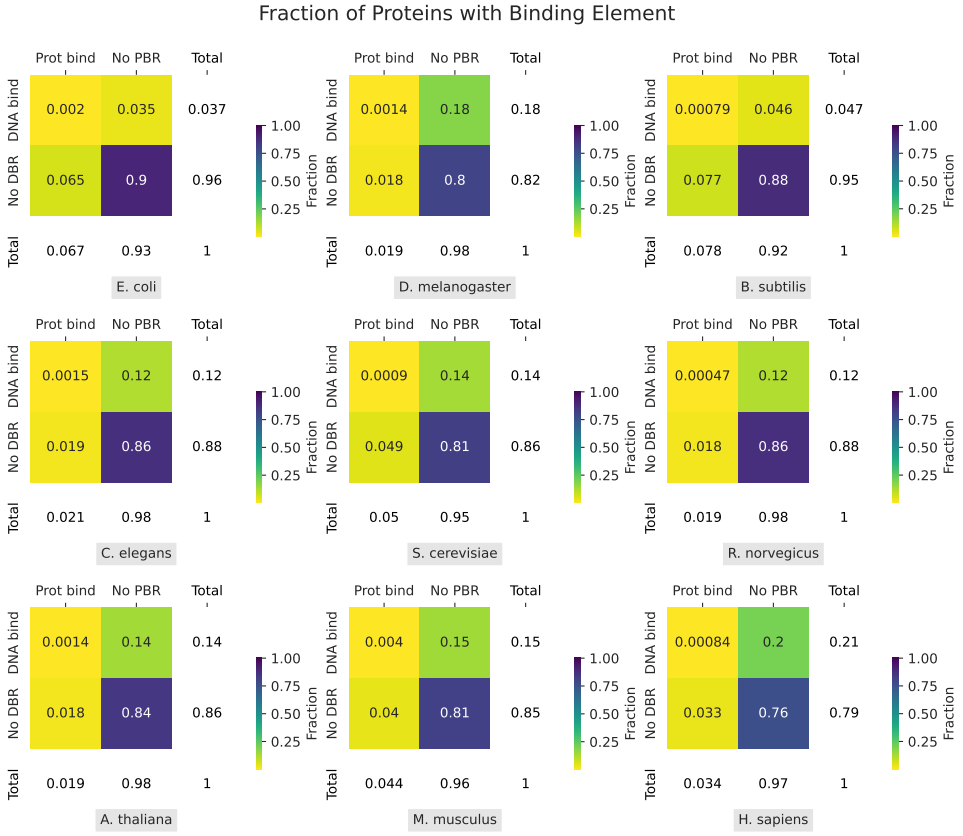

Figure II.12.: **Presence of DNA and protein binding elements in proteins** among all sample proteomes. The decimal numbers displayed in the three by three square indicate the number of proteins, which consist of at least one binding region. Numbers are given relative to proteome size.

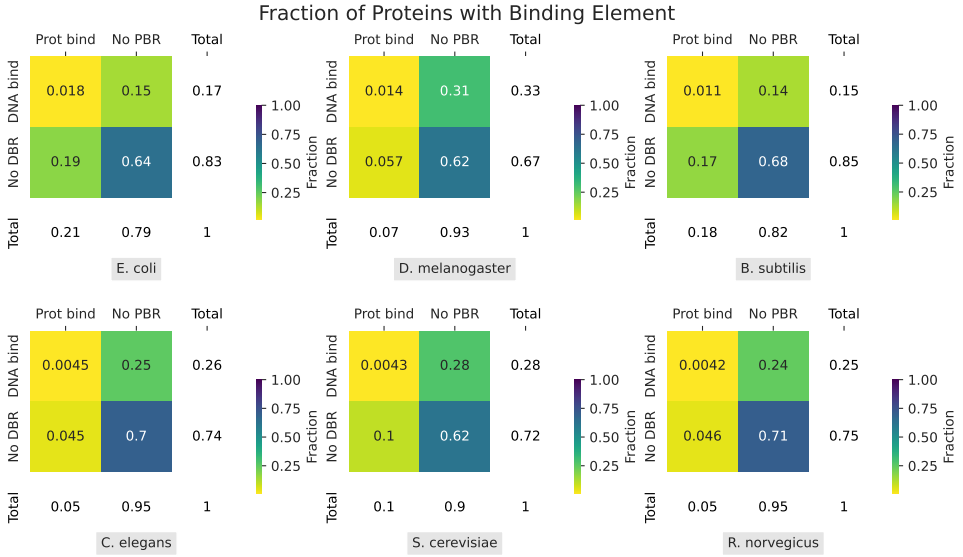

Figure II.13.: **Presence of DNA and protein binding elements in proteins** among six sample proteomes, without filtering to binding regions with a minimum length of six amino acids as performed for Figure II.12.

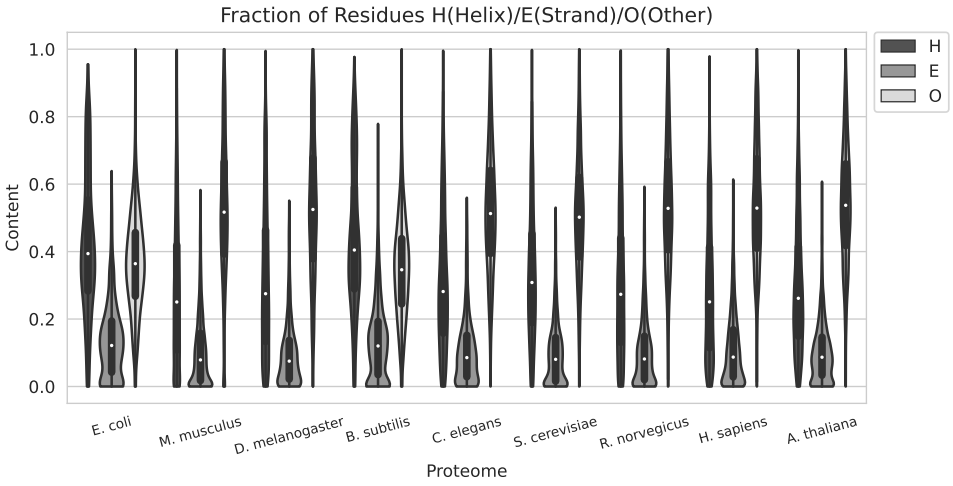

Figure II.14.: Distribution of the **fractions of residues** corresponding to secondary structure. These are defined as the number of residues per protein predicted to be located in a helical (H), strand (E) or other (O) segment, divided by the protein length.

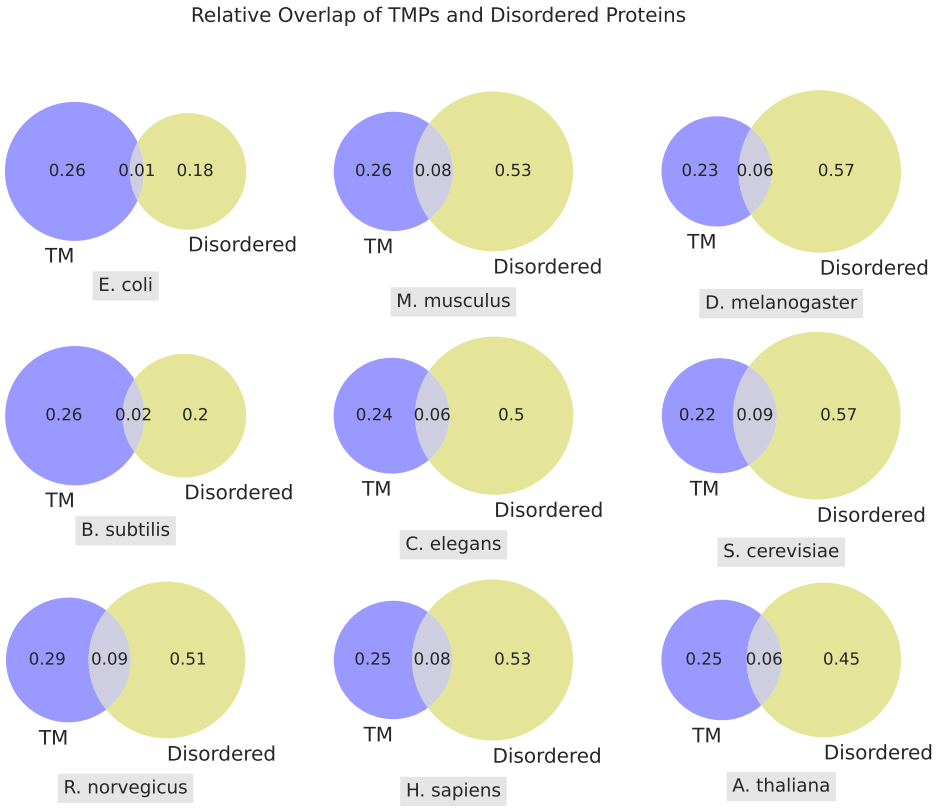

Figure II.15.: **Presence of transmembrane helices and protein disorder within proteins** for all sample proteomes. The decimal numbers displayed in each segment indicate the number of proteins containing at least one region of disorder (yellow), transmembrane helix (blue) or both (intersection). Numbers are given relative to proteome size.

Plotting in `ppprint`: Code Example

```python
from abc import ABC, abstractmethod
from pathlib import Path
from typing import Dict, Tuple

import matplotlib.pyplot as plt
import pandas as pd


class Plot(ABC):
    """
    General abstract class for all plots providing basic functionality for
    plot creation and storing.

    Calling run() on subclass performs all steps of preparing a plot for
    display in ppprint.
    Internally, run() calls the abstract method _run(), which must be
    implemented in all subclasses.
    The specific plotting procedure is handled by _run().
    """

    # Name of source DataFrame depending on feature and base (regions/proteins)
    SOURCE_TYPE: str
    # Plot title
    PLOT_NAME: str
    # Name of the resulting image file
    FILE_NAME: str

    # Storage place for plot
    base_folder: Path
    # All source DataFrames
    dataframes: Dict[str, pd.DataFrame]
    # Mapping of internal proteome identifier (primary key)
    # and proteome name with associated color
    proteome_mapping: Dict[int, Tuple[str, Tuple[float, float, float]]]

    def __init__(
        self,
        dataframes: Dict[str, pd.DataFrame],
        proteome_mapping: Dict[int, Tuple[str, Tuple[float, float, float]]],
        base_folder: Path,
    ):
        self.dataframes = dataframes
        self.proteome_mapping = proteome_mapping
        self.base_folder = base_folder

    @abstractmethod
    def _run(self, df: pd.DataFrame):
        """Handles the plotting procedure for the specific plot type."""

        pass

```

```
    def get_df(self):
        """Returns correct dataframe based on source type."""

        return self.dataframes[self.SOURCE_TYPE]

    def run(self):
        """Performs all procedures required for producing a plot."""

        # Set seaborn plotting style
        plt.style.use("seaborn-whitegrid")

        # Clear plt from previous plotting settings
        plt.clf()
        plt.figure().clf()
        plt.cla()

        # Plotting
        self._run(self.get_df())

        # Finalize plot
        self.set_title()
        self.store_plot()

    def set_title(self):
        """Sets the plot title."""

        plt.title(self.PLOT_NAME)

    def store_plot(self):
        """Stores the plot as a PNG image at the correct place."""

        out_path = self.base_folder / f"{self.FILE_NAME}.png"
        plt.savefig(out_path, dpi=200, bbox_inches="tight")

    def get_color_scheme(self):
        """Collects the pre-defined color for all selected proteomes."""

        return {key: value[1] for key, value in self.proteome_mapping.items()}

    def get_proteome_names(self):
        """Collects the pre-defined name for all selected proteomes."""

        return {key: value[0] for key, value in self.proteome_mapping.items()}
```

Listing 1: The abstract Plot class as currently implemented in `ppprint`. As indicated in documentation strings and comments, the Plot class provides basic functionalities required by all types of plots.

```
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt

from exemplary_code import Plot
```

```python


class PCompositionPerProteomePlotMdisorder(Plot):
    """Exemplary class corresponding to the Disorder Composition plot."""

    PLOT_NAME = "Fraction of Proteins With at Least One DR"
    SOURCE_TYPE = "mdisorder pbased"
    FILE_NAME = "mdisorder_p_composition"

    def _run(self, df: pd.DataFrame):
        """This method performs the plotting procedure for a simple barplot."""

        # Set new subplot
        ax1 = plt.subplot()
        # Get proteome names and associated colors for display
        names = self.get_proteome_names()
        colors = self.get_color_scheme()

        # Perform data handling
        df["composition"] = (df["number of regions"] >= 1).replace({True: 1, False: 0})
        df_sizes = pd.DataFrame(df.groupby("proteome").size()).rename(
            columns={0: "proteome size"}
        )
        df = df.join(df_sizes, how="left", on=["proteome"])
        df["composition"] = df["composition"] / df["proteome size"]

        # Compose seaborn plot
        sns.barplot(
            x="proteome",
            y="composition",
            data=df,
            palette=colors,
            ax=ax1,
            estimator=sum,
            ci=95,
            n_boot=1000,
            errwidth=1,
            capsize=0.1,
        )

        # Set plotting details
        ax1.set_xlabel("Proteome")
        ax1.set_ylabel("Fraction")
        ax1.set_xticklabels(labels=names.values(), fontsize=8, rotation=20)
```

Listing 2: An exemplary subclass of the Plot class as currently implemented in `ppprint`. This class provides all required functionalities for plotting the disorder composition in a simple bar plot as displayed in Figure 3.3.

List of Figures

Bibliography

1. Burley, S. K., Joachimiak, A., Montelione, G. T. & Wilson, I. A. Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. *Structure (London, England : 1993)* **16,** 5–11. doi:10.1016/j.str.2007.12.002 (Jan. 2008).
2. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **596,** 590–596. doi:10.1038/s41586-021-03828-1 (Aug. 2021).
3. Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., Olenyi, T., Qiu, J., Schütze, K., Yachdav, G., Ashkenazy, H., Ben-Tal, N., Bromberg, Y., Goldberg, T., Kajan, L., O'Donoghue, S., Sander, C., Schafferhans, A., Schlessinger, A., Vriend, G., Mirdita, M., Gawron, P., Gu, W., Jarosz, Y., Trefois, C., Steinegger, M., Schneider, R. & Rost, B. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Research* **49,** W535–W540. doi:10.1093/nar/gkab354 (May 2021).
4. Marot-Lassauzaie, V., Goldberg, T., Armenteros, J. J. A., Nielsen, H. & Rost, B. Spectrum of Protein Location in Proteomes Captures Evolutionary Relationship Between Species. *Journal of Molecular Evolution* **89,** 544–553. doi:10.1007/s00239-021-10022-4 (Oct. 2021).
5. Goldberg, T., Hecht, M., Hamp, T., Karl, T., Yachdav, G., Ahmed, N., Altermann, U., Angerer, P., Ansorge, S., Balasz, K., Bernhofer, M., Betz, A., Cizmadija, L., Do, K. T., Gerke, J., Greil, R., Joerdens, V., Hastreiter, M., Hembach, K., Herzog, M., Kalemanov, M., Kluge, M., Meier, A., Nasir, H., Neumaier, U., Prade, V., Reeb, J., Sorokoumov, A., Troshani, I., Vorberg, S., Waldraff, S., Zierer, J., Nielsen, H. & Rost, B. LocTree3 prediction of localization. *Nucleic Acids Research* **42,** W350–W355. doi:10.1093/nar/gku396 (May 2014).
6. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L. & Rost, B. Improved Disorder Prediction by Combination of Orthogonal Approaches. *PLOS ONE* **4,** 1–10. doi:10.1371/journal.pone.0004433 (Feb. 2009).
7. Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M. & Rost, B. Protein disorder—a breakthrough invention of evolution? *Current Opinion in Structural Biology* **21,** 412–418. doi:https://doi.org/10.1016/j.sbi.2011.03.014 (2011).
8. Schad, E., Tompa, P. & Hegyi, H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology* **12,** R120. doi:10.1186/gb-2011-12-12-r120 (Dec. 2011).

9. Hahn, M. W. & Wray, G. A. The g-value paradox. *Evolution & Development* **4,** 73–75. doi:https://doi.org/10.1046/j.1525-142X.2002.01069.x (2002).
10. Mészáros, B., Simon, I. & Dosztányi, Z. Prediction of Protein Binding Regions in Disordered Proteins. *PLOS Computational Biology* **5,** 1–18. doi:10.1371/journal.pcbi.1000376 (May 2009).
11. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434. doi:10.1093/bioinformatics/bti541 (June 2005).
12. Singh, S. & Mittal, A. Transmembrane Domain Lengths Serve as Signatures of Organismal Complexity and Viral Transport Mechanisms. *Scientific Reports* **6,** 22352. doi:10.1038/srep22352 (Mar. 2016).
13. Storey, A. J., Naceanceno, K. S., Lan, R. S., Washam, C. L., Orr, L. M., Mackintosh, S. G., Tackett, A. J., Edmondson, R. D., Wang, Z., Li, H.-y., Frett, B., Kendrick, S. & Byrum, S. D. ProteoViz: a tool for the analysis and interactive visualization of phosphoproteomics data. *Mol. Omics* **16,** 316–326. doi:10.1039/C9MO00149B (4 2020).
14. Jehl, P., Manguy, J., Shields, D. C., Higgins, D. G. & Davey, N. E. ProViz—a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Research* **44,** W11–W15. doi:10.1093/nar/gkw265 (Apr. 2016).
15. Razban, R. M., Gilson, A. I., Durfee, N., Strobelt, H., Dinkla, K., Choi, J.-M., Pfister, H. & Shakhnovich, E. I. ProteomeVis: a web app for exploration of protein properties from structure to sequence evolution across organisms' proteomes. *Bioinformatics* **34,** 3557–3565. doi:10.1093/bioinformatics/bty370 (May 2018).
16. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **28.** rcsb.org, 235–242. doi:10.1093/nar/28.1.235 (Jan. 2000).
17. TheUniProtConsortium. *Swiss-Prot* Accessed: 2022-04-20. https://www.uniprot.org/uniprot/?query=reviewed:yes (2022).
18. TheUniProtConsortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49,** D480–D489. doi:10.1093/nar/gkaa1100 (Nov. 2020).
19. TheUniProtConsortium. *TrEMBL* Accessed: 2022-04-20. https://www.uniprot.org/uniprot/?query=reviewed:no (2022).
20. Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M., Richter, L., Ashkenazy, H., Punta, M., Schlessinger, A., Bromberg, Y., Schneider, R., Vriend, G., Sander, C., Ben-Tal, N. & Rost, B. PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research* **42,** W337–W343. doi:10.1093/nar/gku366 (May 2014).
21. Liu, J. & Rost, B. Comparing function and structure between entire proteomes. *Protein Science* **10,** 1970–1979. doi:https://doi.org/10.1110/ps.10101 (2001).
22. Bakheet, T. M. & Doig, A. J. Properties and identification of human protein drug targets. *Bioinformatics* **25,** 451–457. doi:10.1093/bioinformatics/btp002 (Jan. 2009).

23. Bernhofer, M., Kloppmann, E., Reeb, J. & Rost, B. TMSEG: Novel prediction of transmembrane helices. *Proteins: Structure, Function, and Bioinformatics* **84,** 1706–1716. doi:https://doi.org/10.1002/prot.25155 (2016).

24. Baeza-Delgado, C., von Heijne, G., Marti-Renom, M. A. & Mingarro, I. Biological insertion of computationally designed short transmembrane segments. *Scientific Reports* **6,** 23397. doi:10.1038/srep23397 (Mar. 2016).

25. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181,** 223–230. doi:10.1126/science.181.4096.223 (1973).

26. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637. doi:https://doi.org/10.1002/bip.360221211 (1983).

27. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D. & Rost, B. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *bioRxiv.* doi:10.1101/2020.07.12.199554 (2021).

28. Rost, B. & Sander, C. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology* **232,** 584–599. doi:https://doi.org/10.1006/jmbi.1993.1413 (1993).

29. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS* **7,** 2833–2842. doi:https://doi.org/10.1002/pmic.200700131 (2007).

30. Radivojac, P., Clark, W. T., Oron, T. R., *et al.* A large-scale evaluation of computational protein function prediction. *Nature Methods* **10,** 221–227. doi:10.1038/nmeth.2340 (Mar. 2013).

31. Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F. & Rost, B. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *Journal of Molecular Biology* **432,** 2428–2443. doi:https://doi.org/10.1016/j.jmb.2020.02.026 (2020).

32. Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K. & Rost, B. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports* **11,** 23916. doi:10.1038/s41598-021-03431-4 (Dec. 2021).

33. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **9,** 90–95. doi:10.1109/MCSE.2007.55 (2007).

34. Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6,** 3021. doi:10.21105/joss.03021 (2021).

35. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272. doi:10.1038/s41592-019-0686-2 (2020).

36. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362. doi:10.1038/s41586-020-2649-2 (Sept. 2020).

37. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete Saccharomyces cerevisiae. *Genome Research* **8,** 464–478. doi:10.1101/gr.8.5.464 (1998).

38. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33,** 3390–3400. doi:10.1093/nar/gki615 (Jan. 2005).

39. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in transcription factors. *Biochemistry* **45,** 6873–6888. doi:10.1021/bi0602718 (June 2006).

40. Henderson, P. J. The 12-transmembrane helix transporters. *Current Opinion in Cell Biology* **5,** 708–721. doi:https://doi.org/10.1016/0955-0674(93)90144-F (1993).

41. Zhang, X., He, X., Baker, J., Tama, F., Chang, G. & Wright, S. H. Twelve Transmembrane Helices Form the Functional Core of Mammalian MATE1 (Multidrug and Toxin Extruder 1) Protein. *Journal of Biological Chemistry* **287,** 27971–27982. doi:10.1074/jbc.M112.386979 (Aug. 2012).

42. Denard, B., Han, S., Kim, J., Ross, E. M. & Ye, J. Regulating G protein-coupled receptors by topological inversion. *eLife* **8** (eds Dustin, M. L., Aldrich, R. & Dustin, M. L.) e40234. doi:10.7554/eLife.40234 (Mar. 2019).

43. Karlin, S., Brocchieri, L., Campbell, A., Cyert, M. & Mrázek, J. Genomic and proteomic comparisons between bacterial and archaeal genomes and related comparisons with the yeast and fly genomes. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 7309–7314. doi:10.1073/pnas.0502314102 (May 2005).

44. Preisler, S. S., Wiuf, A. D., Friis, M., Kjaergaard, L., Hurd, M., Becares, E. R., Nurup, C. N., Bjoerkskov, F. B., Szathmáry, Z., Gourdon, P. E., Calloe, K., Klaerke, D. A., Gotfryd, K. & Pedersen, P. A. Saccharomyces cerevisiae as a superior host for overproduction of prokaryotic integral membrane proteins. *Current Research in Structural Biology* **3,** 51–71. doi:https://doi.org/10.1016/j.crstbi.2021.02.001 (2021).

45. Ambri, F., Snoek, T., Skjoedt, M. L., Jensen, M. K. & Keasling, J. D. in *Synthetic Metabolic Pathways: Methods and Protocols* (eds Jensen, M. K. & Keasling, J. D.) 269–290 (Springer New York, New York, NY, 2018). doi:10.1007/978-1-4939-7295-1_17.

46. DjangoSoftwareFoundation. *Django* Accessed: 2022-04-21. https://www.djangoproject.com/ (2022).

47. TheBootstrapTeam. *Bootstrap* Accessed: 2022-04-21. https://getbootstrap.com/ (2022).