Large-scale Characterization Of Intrinsic Disorder And High-throughput Prediction Of RNA, DNA and Protein Binding Mediated By Intrinsic Disorder

by

Zhenling Peng

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Intrinsically disordered proteins lack stable 3D structures *in vivo*, are functionally important, and are very common in nature. In the past three decades, many studies focused on prediction of intrinsic disorder from protein sequence, estimation of its abundance, and analyses of its functional roles. However, these studies were limited in their scope; for example, they focused only on one of many functional and structural aspects. We performed first-of-its-kind comprehensive and detailed analysis of abundance, functional roles, and cellular localizations of intrinsic disorder in complete proteomes. We show that intrinsic disorder is abundant across all kingdoms of life including viruses, is involved in crucial cellular processes, such as translation, transcription, metabolism, regulation, signaling, and so on, and is preferentially located in the ribosome and nucleus. We also mapped intrinsic disorder into eukaryotic, bacterial and archaean cells. These observations motivated us to further analyze two protein families – ribosomal proteins and proteins involved in the programmed cell death. We performed analysis across multiple species, which shows that intrinsic disorder is enriched and performs a variety of important cellular functions in ribosomal and cell death proteins. These two studies reveal that intrinsic disorder is involved in the interactions between proteins, RNAs, and DNAs. The prediction and characterization of these interactions for ordered proteins (i.e., proteins with stable 3D structures *in vivo*) recently attracted significant attention. However, there are no methods that target these functions/interactions mediated by the intrinsic disorder. Development of such methods is now possible by using the curated functional annotations of intrinsic disorder from the DisProt database. Utilizing these data we developed the first computational prediction method, DisoRDPbind, that predicts protein-protein, -RNA and -DNA interactions mediated by the intrinsic disorder. Our method utilizes logistic regression algorithm and a custom-designed and empirically selected set of descriptors of the input protein sequence. Empirical assessment using two benchmark datasets and large-scale predictions on four eukaryotic proteomes suggests that DisoRDPbind provides good predictive quality, differs from the methods focused on the predictions for the ordered proteins, and its computational efficiency allows for annotation of these interactions in whole proteomes.

Preface

This thesis is an original work conducted by Zhenling Peng. The research project, of which this thesis is a part, received funding from the Alberta Innovates Technology Future, Project Title "Computational characterization and improved prediction of protein disorder with application in protein-protein interactions", No. 201100048, May 1st, 2011 to April 30th, 2014.

Some of the research in Chapter 3 and Chapter 4 forms part of international research collaboration, led by Professor V.N. Uversky at the University of South Florida, with Professor L. Kurgan being the lead collaborator at the University of Alberta. The data collection and analysis and technical apparatus in these Chapters are my original work, with the supervision of Professor L. Kurgan.

Chapter 3 of this thesis has been accepted as Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Uversky VN and Kurgan L, "Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in a thousand proteomes from all domains of life", *Cellular and Molecular Life Science*, on Jun 18, 2014. I was responsible for the data collection and analysis, technical apparatus and manuscript composition.

Chapter 4 of this thesis has been published as two papers, including Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, Kurgan L and Uversky VN, "A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome", *Cellular and Molecular Life Science*, vol. 71, issue 8, 1477-504; and Peng Z, Xue B, Kurgan L and Uversky VN, "Resilience of death: intrinsic disorder in proteins involved in the programmed cell death", *Cell Death and Differentiation*, vol. 20, 1257-1267. I was responsible for the data collection and analysis, technical apparatus and manuscript composition.

Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisor Dr. Lukasz Kurgan for all his guidance, passion, motivation and most of all for the tremendous amount of time he spent to make me a better researcher. I could not have imagined having a better mentor for my PhD study.

I would like to especially thank my parents and my husband Jianyi for their love and support, their encouragement, and for their understanding of my choices. They helped me to become a better person.

I would like to thank my fellow lab members for their support and collaboration.

I would like to thank my friends for their support, motivation and time spent together.

I would like to extend my gratitude to Alberta Innovates for the financial assistance during my PhD studies.

Table of Contents

CHAPTER 1 INTRODUCTION 1

1.1 Thesis Statements and Aims 3

1.2 Thesis Outline 5

CHAPTER 2 BACKGROUND AND RELATED WORK 6

2.1 Proteins 6

2.2 Classic Sequence-to-Structure-to-Function Paradigm 9

2.3 Intrinsically Disordered Proteins 10

2.4 Data Sources for the Intrinsic Disorder 12

2.5 Prediction of Intrinsic Disorder from Protein Sequences 13

2.6 Characterization of Abundance and Functions of Intrinsic disorder 15

2.7 Molecular Recognition Features 17

2.8 Computational Methods 18

2.8.1 Prediction 18

2.8.2 Logistic Regression 21

2.8.3 Cross-validation 21

2.8.4 Evaluation Criteria 22

2.8.5 Statistical Tests 23

CHAPTER 3 SYSTEMATIC ANALYSIS OF INTRINSIC DISORDER IN COMPLETE PROTEOMES 27

3.1 Motivation 27

3.2 Materials and Methods 28

3.2.1 Proteomic Level Dataset 28

3.2.2 Disorder Prediction 28

3.2.3 Analysis of Disorder Enrichment in Functions and Cellular Components 29
3.2.4 Evolutionary Pace 30
3.2.5 Sequence Conservation 30

3.3 Results and Discussion 31
3.3.1 Abundance of Intrinsic Disorder across Kingdoms of Life 31
3.3.2 Functional Roles of Intrinsic Disorder across Kingdoms of Life 34
3.3.3 Cellular Localization of Intrinsic Disorder across Kingdoms of Life 37
3.3.4 Evolution and Disorder 38

3.4 Conclusions 42

CHAPTER 4 CHARACTERIZATION OF INTRINSIC DISORDER IN RIBOSOMAL AND CELL DEATH PROTEINS 44

4.1 Introduction and Motivation 44
4.1.1 Introduction and Motivation for Study on Ribosomal Proteins 44
4.1.2 Introduction and Motivation for Study on Cell Death Proteins 45

4.2 Materials and Methods 46
4.2.1 Benchmark Datasets 46
4.2.2 Amino Acid Composition Analysis 48
4.2.3 Computational Prediction and Characterization of Disorder 48
4.2.4 Search for Potential Globular Domains in Ribosomal Proteins 48
4.2.5 Search for Potential Functional Sites 49
4.2.6 Calculation of Sequence Conservation 50

4.3 Results for Study on Ribosomal Proteins 50
4.3.1 Peculiarities of the Amino Acid Compositions in Ribosomal Proteins 50
4.3.2 Overall Characterization of the Intrinsic Disorder in Ribosomal, RNA-, and DNA-Binding Proteins 53
4.3.3 Functional Analysis of Disordered Regions in Ribosomal Proteins 59
4.3.4 MoRF Regions in Ribosomal, RNA-, and DNA-Binding Proteins 60
4.3.5 Evolutionary Conservation of Disorder in Ribosomal Proteins 62

4.4 Discussion and Conclusions for Study on Ribosomal Proteins 65
4.4.1 Commonness and Peculiarities of Intrinsic Disorder in Ribosomal Proteins 65

4.4.2 Functional Viewpoint of Intrinsic Disorder in the Ribosomal Proteins 66

4.5 Results for Study on Programmed Cell Death Proteins 71

4.5.1 Characterization of the Intrinsic Disorder in Human Proteins Associated with the Programmed Cell Death 71

4.5.2 Overall Characteristics of the Intrinsic Disorder in Deathbase 73

4.5.3 Functional Analysis of Disordered Regions 76

4.5.4 MoRF Regions in Proteins Involved in Programmed Cell Death 78

4.5.5 Evolutionary Conservation of Disorder in PCD-related Proteins 80

4.6 Discussion and Conclusions for Study on Cell Death Proteins 81

CHAPTER 5 HIGH-THROUGHPUT PREDICTION OF RNA, DNA AND PROTEIN BINDING MEDIATED BY INTRINSIC DISORDER 84

5.1 Introduction and Motivation 84

5.2 Materials and Methods 86

5.2.1 Annotation of Disordered RNA-, DNA-, and Protein-binding 86

5.2.2 Benchmark Datasets 86

5.2.3 Evaluation Criteria and Statistical Significance 88

5.2.4 Assessment at the Whole Proteome Level 88

5.2.5 Architecture of DisoRDPbind 91

5.2.6 Sequence Representation 92

5.2.7 Feature Selection 93

5.2.8 Regression Model 94

5.2.9 Combining Regression with Sequence Alignment 95

5.3 Results and Discussion 96

5.3.1 Comparative Evaluation of DisoRDPbind 96

5.3.2 Evaluation of Runtime 100

5.3.3 Validation on the Whole Proteomes 102

5.4 Conclusions 107

CHAPTER 6 SUMMARY AND CONCLUSIONS 109

6.1 Major Contributions 110

6.2 Major Findings 112

BIBLIOGRAPHY 114

APPENDIX A LIST OF RIBOSOMAL PROTEINS WITH IDRS AND ON-RIBOSOME FUNCTIONS 133

APPENDIX B FUNCTIONAL ANNOTATIONS FOR INTRINSIC DISORDER 135

APPENDIX C BENCHMARK DATASETS FOR DISORDPBIND 136

APPENDIX D SELECTION OF PHYSICOCHEMICAL PROPERTIES OF AMINO ACIDS FOR DISORDPBIND 138

APPENDIX E SEQUENCE REPRESENTATION FOR DISORDPBIND 140

List of Tables

TABLE 2.1: THE LIST OF 20 STANDARD AMINO ACIDS ALONG WITH THEIR SELECTED PROPERTIES. 7

TABLE 2.2. CONFUSION MATRIX. 23

TABLE 2.3 RELEVANT FRAGMENT OF THE *T* TABLE. 25

TABLE 3.1. SUMMARY OF GO ANNOTATIONS. 35

TABLE 5.1. COMPARISON BETWEEN DISORDPBIND AND OTHER SIX CONSIDERED METHODS. 98

List of Figures

FIGURE 2.1. GENERAL STRUCTURE OF AN AMINO ACID. 6

FIGURE 2.2. MECHANISM OF EUKARYOTIC PROTEIN SYNTHESIS.......... 8

FIGURE 2.3. RAMACHANDRAN PLOT AND HIERARCHY OF PROTEIN STRUCTURES. 9

FIGURE 2.4. AN EXAMPLE OF 3D STRUCTURE OF IDP.......... 10

FIGURE 2.5. NUMBER OF ENTRIES IN DISPROT DATABASE FROM 2005 TO 2013. 13

FIGURE 3.1 OVERVIEW OF DISORDER CONTENT AND LONG IDRS ACROSS PHYLA AND KINGDOMS.......... 32

FIGURE 3.2. DISTRIBUTION OF DISORDER CONTENT, LENGTH OF IDRS, AND FULLY DISORDERED PROTEINS.......... 34

FIGURE 3.3. FUNCTIONS ENRICHED IN DISORDER. 36

FIGURE 3.4. CELLULAR LOCALIZATION ENRICHED IN DISORDER. 38

FIGURE 3.5. MAPPING INTRINSIC DISORDER INTO EUKARYOTIC, BACTERIAL, AND ARCHAEA CELLS. 38

FIGURE 3.6. RELATION BETWEEN DISORDER CONTENT AND EVOLUTIONARY CHARACTERISTICS INCLUDING EVOLUTIONARY SPEED, SEQUENCE CONSERVATION AND PROTEOME SIZE FOR THE BACTERIAL AND EUKARYOTIC SPECIES.......... 42

FIGURE 4.1. OVERVIEW OF PECULIARITIES OF AMINO ACID COMPOSITION IN RIBOSOMAL PROTEINS. 52

FIGURE 4.2. AMOUNT OF INTRINSIC DISORDER IN THE RIBOSOMAL, DNA-, AND RNA-BINDING PROTEINS. 56

FIGURE 4.3. DISTRIBUTION OF LONG DISORDERED REGIONS ACROSS CHAINS WITH DIFFERENCE LENGTH. 57

FIGURE 4.4. CHARACTERIZATION OF THE DOMAINS IN RIBOSOMAL PROTEINS. 58

FIGURE 4.5. FUNCTIONS OF SHORT AND LONG IDRS EXTRACTED FROM RIBOSOMAL PROTEINS, RESPECTIVELY. 60

FIGURE 4.6. MORFS ON RIBOSOMAL, RNA- AND DNA-BINDING PROTEINS, RESPECTIVELY.......... 61

FIGURE 4.7. DISTRIBUTION OF THE EVOLUTIONARY CONSERVATION FOR RIBOSOMAL, RNA- AND DNA-BINDING PROTEINS.......... 63

FIGURE 4.8. THE AVERAGE EVOLUTIONARY CONSERVATION FOR RIBOSOMAL, RNA- AND DNA-BINDING PROTEINS. 64

FIGURE 4.9. COMPUTATIONAL DISASSEMBLY OF THE EUKARYOTIC RIBOSOME A FROM THE YEAST *S. CEREVISIAE* (PDBID: 3U5C AND 3U5E). 65

FIGURE 4.10. PECULIARITIES OF INTRINSIC DISORDER DISTRIBUTION IN HUMAN PCD PROTEINS. 72

FIGURE 4.11. EVALUATION OF ABUNDANCE OF INTRINSIC DISORDER IN PCD-RELATED PROTEINS FROM THE DEATHBASE.......... 75

FIGURE 4.12. CORRELATION BETWEEN INTRINSIC DISORDER IN CELL DEATH PROTEINS AND THEIR FUNCTIONS. 77

FIGURE 4.13. MORF REGIONS IN PCD-RELATED PROTEINS.......... 79

FIGURE 4.14. EVOLUTIONARY CONSERVATION FOR PCD-RELATED PROTEINS. 80

FIGURE 5.1. ARCHITECTURE OF DISORDPBIND 91

FIGURE 5.2. ROC CURVES FOR DISORDPBIND AND OTHER SIX CONSIDERED METHODS. 99

FIGURE 5.3. RELATIONS BETWEEN PAIR OF METHODS. 100

FIGURE 5.4. EVALUATION OF RUNTIME OF DISORDPBIND. 101

FIGURE 5.5. CORRELATION BETWEEN THE PROMISCUITY OF A PROTEIN AND THE NUMBER OF ITS PREDICTED DISORDERED PROTEIN-BINDING REGIONS. 103

FIGURE 5.6. VENN DIAGRAMS OF THE OVERLAP BETWEEN THE GO ANNOTATIONS AND THE PREDICTION FOR RNA AND DNA BINDING PROTEINS. 105

FIGURE 5.7. SIGNIFICANCE OF OVERLAP BETWEEN THE GO ANNOTATIONS AND THE PREDICTION FOR RNA AND DNA BINDING PROTEINS. 106

FIGURE 5.8. SEQUENCE SIMILARITY BETWEEN THE NOVEL PUTATIVE RNA (DNA) BINDING PROTEINS AND PROTEINS FROM GO_RNA AND/OR RBPDB (GO_DNA AND ANIMALTFDB) DATABASES. 107

List of Abbreviations

3D – Three-dimensional

AA(s) – Amino Acid(s)

AUC – Area Under the ROC Curve

CASP – Critical Assessment of Techniques for Protein Structure Prediction

DisProt – Database of Protein Disorder

DisoRDPbind – Predictor of RNA, DNA and Protein binding mediated by disorder

DNA – Deoxyribonucleic acid

FN – False Negatives (positive annotations that were predicted as negatives)

FP – False Positives (negative annotations that were predicted as positives)

GO – Gene Ontology

IDP - Intrinsically disordered protein

IDR - Intrinsically disordered region

MoRF – molecular recognition feature

mRNA – messenger RNA

PBC – Point-biserial correlation coefficient

PCC – Pearson correlation coefficient

PCD – programmed cell death

PDB – Protein Data Bank

PPI – protein-protein interaction

RNA – Ribonucleic acid

ROC – Receiver operating characteristic

RPG – Ribosomal Protein Gene Database

rRNA - Ribosomal ribonucleic acid

TN – True Negatives (correctly predicted negative annotations)

TP – True Positives (correctly predicted positive annotations)

tRNA – transporting ribonucleic acid

Chapter 1

Introduction

Intrinsically **d**isordered **p**roteins (IDPs), also called natively denatured, natively unfolded, intrinsically unstructured, mostly unstructured, and natively disordered, etc. (Dunker, et al., 2013; Tompa, 2002; Uversky, et al., 2010), lack stable three-dimensional (3D) structure under physiological conditions *in vivo*.

The first examples of IDPs were found in the 1930s (Landsteiner, 1936). Since they do not fit the classic sequence-to-structure-to-function paradigm, which states that the protein sequence determines its three-dimensional (3D) structure and this specific structure in turn determines its functions, IDPs were taken as rare exceptions. However, later on IDPs were found to be relatively common in nature. For example, eukaryotic proteomes, such as *C. elegans*, *A. thaliana*, *S. cerevisiae*, and *D. melanogaster*, were estimated to have between 52% and 67% of their proteins with long **i**ntrinsically **d**isordered **r**egions (IDRs), which comprise of at least 40 consecutive disordered residues (Dunker, et al., 2000). Several experimental techniques including nuclear magnetic resonance, X-ray crystallography, Circular dichroism spectropolarimetry, etc., are used to detect or provide useful information about IDPs/IDRs. Based on the corresponding experimental data, IDPs and IDRs were implicated in various human diseases, such as cancer (Uversky, et al., 2008; Uversky, et al., 2009), and were suggested as important targets for drug discovery (Cheng, et al., 2006). Consequently, intrinsic disorder draws an increasing amount of attention.

However, the experimental annotations of intrinsic disorder lag behind the rapidly accumulating number of known proteins. Recent studies show that the intrinsic disorder is predictable from protein chains because IDPs/IDRs possess relatively unique characteristics in their sequences, such as a biased amino acid composition, relatively low sequence complexity, etc. (Dosztanyi, et al., 2010; Dyson, et al., 2005; Liu, et al., 2002; Romero, et al., 2001; Uversky, et al., 2000). Therefore, dozens of computational methods were developed for the prediction of intrinsic disorder from the sequences. Comprehensive summaries of these methods are given in recent reviews (Dosztanyi, et al., 2010; He, et al., 2009; Peng, et al., 2012; Uversky, et al., 2010). These predictors allow for high-throughput and accurate prediction of intrinsic disorder, and thus provide a viable solution to close the annotation gap.

In addition, these computational methods enable a systematic analysis of the prevalence and cellular functions of intrinsic disorder. Several efforts have been devoted to estimate the natural abundance of IDPs/IDRs, and some of them also included the investigation of the functional roles of IDPs/IDRs (Burra, et al., 2010; Dunker, et al., 2000; Feng, et al., 2006; Galea, et al., 2009; Tompa, et al., 2006; Ward, et al., 2004; Xue, et al., 2012). These analyses reveal that IDPs/IDRs are not only exceptionally common in nature, but also perform crucial biological functions, which complement those of ordered proteins that have unique 3D structure (Dunker, et al., 2008; Dunker, et al., 2008). All of these results led to the acceptance and appreciation of intrinsic disorder phenomenon in modern structural biology (Dunker, et al., 2005; Dunker, et al., 2001; Dunker, et al., 2008; Dyson, et al., 2005; Tompa, 2002; Uversky, et al., 2005; Wright, et al., 1999).

We have made several following observations that serve as the motivation for the work presented in this thesis:

1. Modern studies on the intrinsic disorder have some limitations, primarily in terms of the scope of the analyses (see section 3.1 for more details). This calls for a systematic and comprehensive analysis of intrinsic disorder that considers its abundance profiles, cellular functions, and cellular localizations across all kingdoms of life.
2. Ribosomal proteins are an important family of proteins that are known to be either completely disordered or to contain long IDRs in isolation (Ban, et al., 2000; Ben-Shem, et al., 2011; Harms, et al., 2001; Schuwirth, et al., 2005; Selmer, et al., 2006; Timsit, et al., 2009; Wimberly, et al., 2000; Yusupov, et al., 2001), but they were never the subject of a focused large-scale bioinformatics analysis.
3. Programmed cell death is an important cellular process that is regulated by different signaling pathways and guided by a series of protein-protein interactions (PPI)

(Bialik, et al., 2010; Tan, et al., 2009). Previous works demonstrated that IDPs act as important regulators of PPI networks (Dosztanyi, et al., 2006; Dunker, et al., 2005; Ekman, et al., 2006; Haynes, et al., 2006; Patil, et al., 2006). These observations indicate the intrinsic disorder could be associated with cell death processes, and this aspect was not studied so far.

4. Recent studies demonstrate that IDPs/IDRs are involved in the interactions between RNAs, DNAs and proteins (Chen, et al., 2006; Dunker, et al., 2005; Peng, et al., 2012), which leads to important roles of IDPs/IDRs in regulation, signaling, translation, transcription, and various other cellular processes. However, computational prediction of these interactions, except for PPIs, mediated by intrinsic disorder was to date neglected.

1.1 Thesis Statements and Aims

Motivated by the observations listed above, our aim is to perform comprehensive and detailed analysis of natural abundance and functional roles of intrinsic disorder at the proteome level, and to predict RNA, DNA and protein binding (i.e., interaction) mediated by the intrinsic disorder. We address the following thesis statements:

1. Each kingdom of life has its own unique disorder abundance profile, besides the known differences in the overall amount of disorder.
2. Besides the already known functional roles of intrinsic disorder, the disorder is involved in a larger repertoire of biological functions that may differ between kingdoms of life.
3. Intrinsic disorder is known to be preferentially located in certain parts of a cell. These preferences may be different across kingdoms of life.
4. Intrinsic disorder is enriched and plays crucial functional roles in ribosomal proteins.
5. Intrinsic disorder is abundant and performs important functions in proteins involved in the programmed cell death.
6. RNA, DNA and protein binding mediated by intrinsic disorder is predictable from amino acid sequence. These binding events can be accurately predicted on whole proteomes.

We define three aims to address the aforementioned thesis statements:

1. **To perform first-its-kind comprehensive and detailed analysis of the abundance and the cellular functions of IDPs/IDRs in all complete proteomes**. This aim addresses the thesis statements 1, 2 and 3. We characterized the abundance and the functional roles of intrinsic disorder across 965 complete proteomes, from all four kingdoms of life including eukaryota, bacterial, archaea and viruses. We annotated intrinsic disorder in these proteomes utilizing an accurate and time-efficient consensus-based prediction method. This putative disorder was used to estimate the abundance of intrinsic disorder in each proteome and each kingdom of life. We analyzed and discussed enrichment of the disorder in a broad range of functions and cellular localizations, and we mapped it into eukaryotic, bacterial and archaea cells. Based on a phylogenetic tree, we also investigated relation between the intrinsic disorder and evolutionary rate in eukaryotes and bacteria.
2. **To investigate the prevalence and the biological importance of intrinsic disorder in two protein families -- ribosomal proteins and proteins involved in the programmed cell death**. This aim addresses thesis statements 4 and 5. Using an accurate computational method, we performed prediction of the intrinsic disorder in multiple species, and analyzed its abundance in the ribosomal and cell death proteins. We also predicted cellular functions of IDRs that we found in these proteins. We computed and assessed their evolutionary conservation to characterize biological significance of their functional roles.
3. **To develop the first computational method that accurately and in high-throughput fashion predicts RNA, DNA and protein binding regions located in IDRs in protein sequences**. This aim stems from the thesis statement 6. Based on a multi-layered design, we have built a novel computational method, DisoRDPbind, to predict the RNA, DNA and protein binding residues located in IDRs. We performed empirical evaluation of DisoRDPbind on two benchmark test datasets to assess its predictive quality and compared it to other relevant methods. We also measured runtime of DisoRDPbind to test its computational efficiency. We applied DisoRDPbind to annotate and analyze these three binding events on four complete proteomes including *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*.

Our work provides insights into the characteristics of the intrinsic disorder in each kingdoms of life, and in specific protein families including ribosomal and cell death proteins. Our new computational method is the first to predict DNA and RNA binding functions of the intrinsic disorder from protein sequences.

1.2 Thesis Outline

In Chapter 2, we introduce the relevant background including introduction to proteins and IDPs, information concerning the corresponding IDP-related resources, and we summarize recent relevant studies including overview of disorder predictors and existing efforts concerning characterization of IDPs/IDRs. Since these characterization efforts are limited in scope and/or breadth, we have performed first-of-its-kind large-scale characterization of the abundance and the biological importance of intrinsic disorder on all complete proteomes and on two specific families of proteins – ribosomal proteins and proteins involved in programmed cell death. These studies are described in Chapter 3 and Chapter 4, respectively. Our analysis highlighted various functional roles of the intrinsic disorder, in particular related to the interactions between proteins and other large macromolecules, such as DNA and RNA. Given these findings and the fact that intrinsic disorder was shown to be important for PPIs, and since there are no existing computational methods that can find the IDRs that are involved in these binding events in protein chains, we developed the first computationally efficient method to predict the disorder-driven RNA, DNA and protein binding. The design of this method together with its empirical evaluation on benchmark datasets and on proteomic scale is discussed in Chapter 5. Chapter 6 summarizes this research and list major contributions.

Chapter 2

Background and Related Work

2.1 Proteins

Proteins are large biological molecules which are essential for every living organism and which are responsible for nearly every cellular level function including maintaining cell shape and routines, catalyzing biochemical reactions, neutralizing antigens during the immune response, serving as signal receptors during cellular signal transduction, and transporting molecules from one location to another, to name just a few (Gilman, 1987; Gutteridge, et al., 2005; Howard, et al., 2007).

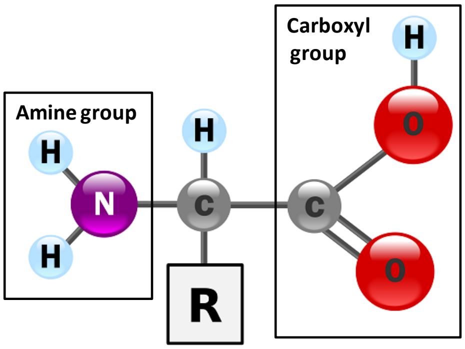

Figure 2.1. General structure of an amino acid. This structure is common to all amino acid, except proline.

Proteins are composed of one or more polypeptide chains. The polypeptide chain is a linear chain built from amino acids (AAs). There are 20 different standard AAs; see **Table 2.1**. The AAs are connected by peptide bonds. All AAs, except proline (a cyclic AA), possess a common structural feature which includes a central α-carbon atom that is covalently bonded to a carboxyl group (box on the right), a hydrogen (middle top), an amine group (box on the left) and a side-chain (i.e., R group in the middle bottom box); see **Figure 2.1**. The side-chains or R group attached to α-carbon atom are different for each AA. They are characterized by a variety of chemical structures and properties; see **Table 2.1**. The peptide bond is generated by the chemical reaction between the carboxyl group (-COOH) of a given AA and the amine group ($-NH_2$) of another AA, i.e., a peptide bond is the resulting -C(=O)NH- bond by releasing a molecule of water (H_2O) after the chemical reaction between two AAs.

Table 2.1: The list of 20 standard amino acids along with their selected properties.
The table gives abbreviated names of 20 standard AAs along with chemical composition of their side chains, the selected biochemical properties, occurrence in proteins ("Occ." column), and their codons in RNA. AA properties include annotation whether they have positive(+)/negative(-) charge, and whether they are polar(P) and/or hydrophobic(H) according to (Livingstone & Barton, 1993). C, O, H and N in side chain represent carbon, oxygen, hydrogen and nitrogen atoms, respectively. Codon is a nucleotide triplets. Since there are four types of nucleotides in RNA, there are 64 different codons. Thus several amino acids have more than one corresponding codon. A, C, G and U in codon represent adenine, cytosine, guanine and uracil, respectively.

Amino Acid	Abbr.	Side chain	Property			Occ.	Codon
Alanine	Ala, A	-CH_3			H	7.8 %	GCU, GCC, GCA, GCG
Arginine	Arg, R	-$(CH_2)_3NH-C(NH)NH_2$	+	P		5.1 %	CGU, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn, N	-CH_2CONH_2		P		4.3 %	AAU, AAC
Aspartate	Asp, D	-CH_2COOH	-	P		5.3 %	GAU, GAC
Cysteine	Cys, C	-CH_2SH		P	H	1.9 %	UGU, UGC
Glutamate	Glu, E	-CH_2CH_2COOH	-	P		6.3 %	GAA, GAG
Glutamine	Gln, Q	-$CH_2CH_2CONH_2$		P		4.2 %	CAA, CAG
Glycine	Gly, G	-H			H	7.2 %	GGU, GGC, GGA, GGG
Histidine	His, H	-CH_2-$C_3H_3N_2$	+	P	H	2.3 %	CAU, CAC
Isoleucine	Ile, I	-$CH(CH_3)CH_2CH_3$			H	5.3 %	AUU, AUC, AUA
Leucine	Leu, L	-$CH_2CH(CH_3)_2$			H	9.1 %	UUA, UUG, CUU, CUC, CUA, CUG
Lysine	Lys, K	-$(CH_2)_4NH_2$	+	P	H	5.9 %	AAA, AAG
Methionine	Met, M	-$CH_2CH_2SCH_3$			H	2.3 %	AUG
Phenylalanine	Phe, F	-$CH_2C_6H_5$			H	3.9 %	UUU, UUC
Proline	Pro, P	-$CH_2CH_2CH_2$-				5.2 %	CCU, CCC, CCA, CCG
Serine	Ser, S	-CH_2OH		P		6.8 %	UCU, UCC, UCA, UCG, AGU, AGC
Threonine	Thr, T	-$CH(OH)CH_3$		P	H	5.9 %	ACU, ACC, ACA, ACG
Tryptophan	Trp, W	-$CH_2C_8H_6N$		P	H	1.4 %	UGG
Tyrosine	Tyr, Y	-CH_2-C_6H_4OH		P	H	3.2 %	UAU, UAC
Valine	Val, V	-$CH(CH_3)_2$			H	6.6 %	GUU, GUC, GUA, GUG

Protein synthesis includes two steps: transcription and translation, which are visualized in **Figure 2.2**. During transcription, a messenger RNA (mRNA) is generated in the cell's nucleus. Specifically, enzyme helicase unzips the double helix strand of a DNA, by breaking the hydrogen bonds between the two strands. RNA polymerase then reads one of the strands from 3-prime (3') end to the 5-prime (5') end, to generate the mRNA from 5'-to-3' direction. Note that the nucleotide uracil (U) in RNA replaces the thymine (T) in DNA. This resulting single strand mRNA moves from nucleus to cytoplasm to undergo translation. During translation, this mRNA is loaded into ribosome, which is composed of a small and a large subunit that surround a given part of mRNA that is being translated. The small ribosomal subunit reads one nucleotide triplet (i.e., codon) of the mRNA at a time. According to the base pairing rules, this codon is matched with the anticodon located on the transfer RNA (tRNA), which carries the amino acid the codon recognizes (see **Table 2.1**). The large ribosomal subunit connects these amino acids sequentially by the peptide bond, forming a polypeptide chain. When the small ribosomal subunit reads a stop codon (i.e., UAA, UAG, or UGA), the translation is terminated, and the synthesized polypeptide chain is released into the cytoplasm.

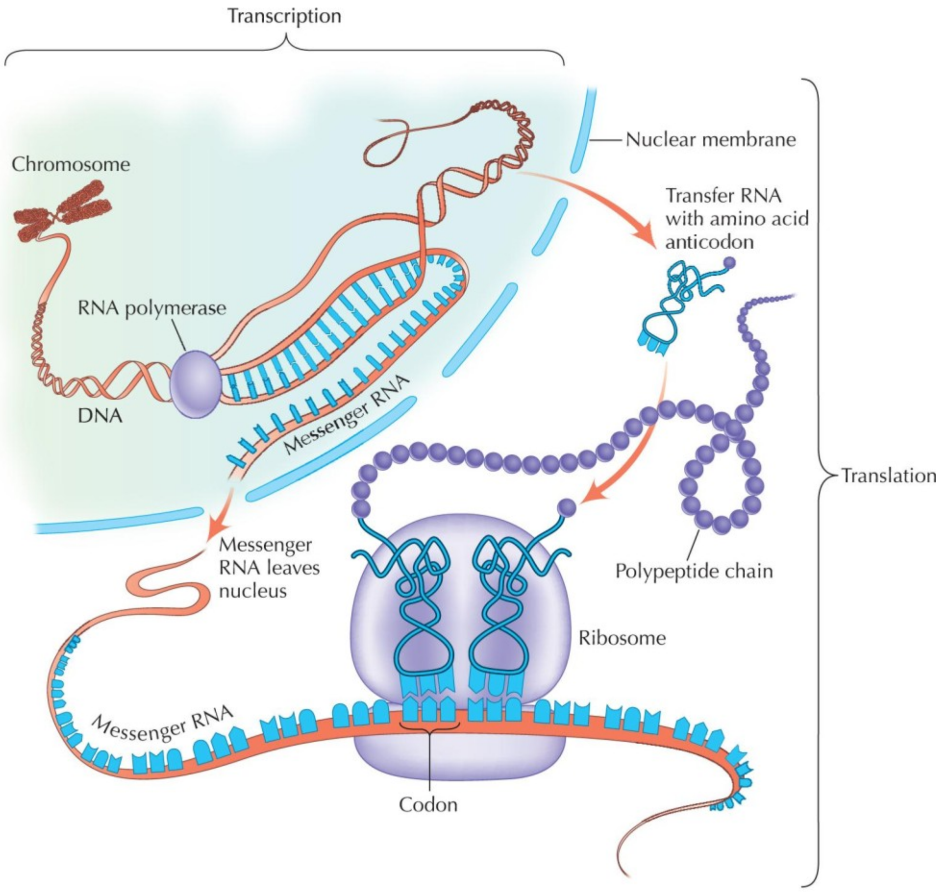

Figure 2.2. Mechanism of eukaryotic protein synthesis.
Protein synthesis includes two major steps: transcription of the DNA gene sequence to mRNA, and translation of the resulting mRNA into the amino acid sequence, i.e., the polypeptide chain. Source: *Evolution* © 2007 Cold Spring Harbor Laboratory Press.

During or after protein synthesis, the polypeptide chain folds into a unique 3D structure, which is defined by the coordinates of all the atoms of the protein at the equilibrium position. The protein structures are usually described at three hierarchical levels including primary, secondary and tertiary structure (i.e., 3D structure). The primary structure refers to the linear sequence of residues in a polypeptide chain; an example is shown in the upper panel on the right in **Figure 2.3**. The secondary structure is the spatially local organization of the polypeptide chain, which is driven by the hydrogen bonds between the main-chain peptide groups. The secondary structure includes three major types: α helix, β sheet and coil, where α helix contains a common right-handed helix and a relatively rare left-handed helix. In 1968, Ramachandran developed the Ramachandran plot that utilizes the dihedral angles around the α carbon atom of amino acids to describe the secondary structure; see **Figure 2.3** left panels. This plot reveals that α helix and β sheet are located at different place in Ramachandran plot, or have different pair of dihedral angles Φ and Ψ around the central α carbon. Consequently, α helix and β sheet have the geometry of a regularly twisted ribbon and a relatively flat sheet, respectively; see **Figure 2.3** right middle panel; while coils usually connect α helices and/or β strands with each other. The secondary structural elements of the polypeptide are folded into a compact globular structure – tertiary structure (i.e., 3D structure); see **Figure 2.3** right lower panel.

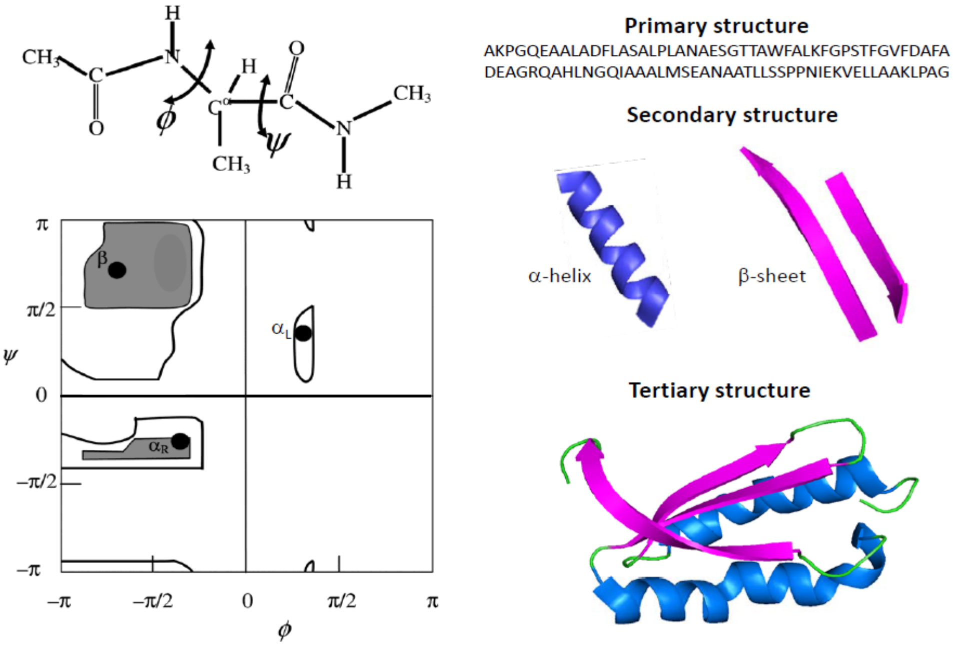

Figure 2.3. Ramachandran plot and hierarchy of protein structures.
(Left panels) Source: Hermans J PNAS 2011;108:3095-3096. The dihedral angles Φ and ψ (example for alanine dipeptide; left upper panel) and Ramachandran plot (left lower panel) of major preferred (dark gray) and allowed (white space rounded by black curves) Φ, ψ angle pairs in proteins, with the position of repetitive secondary structures marked. β denotes β-sheet. α_R and α_L represent right- and left-handed α-helix, respectively. Right panels show the primary, secondary and tertiary structure in the upper, middle and lower panels, respectively.

2.2 Classic Sequence-to-Structure-to-Function Paradigm

The classic sequence-to-structure-to-function paradigm states that a proteins' primary structure determines the 3D structure, and the 3D structure then determines its biological functions. This view has been the cornerstone of structural biology over the past century. In 1894, Emil Fischer came up with the "lock and key" hypothesis based on the studies on different types of similar enzymes (Fischer, 1894), which suggests only the substrate (the key) with the correct size and shape would fit into the active site of the enzyme (the lock). In 1930s, globular proteins were found to denature (unfold) and lose their biological functions by altering the external environment, such as increasing temperature or adding solutes (Mirsky, et al., 1936). This denatured state could reverse back (fold) to its native state by adjusting the conditions of the external environment (Anson, et al., 1925). These observations suggested that the **native** and the **denatured** state are separate thermodynamic states for a protein, where the function is determined by the stable and specific structure of a given protein at its native (folded) state. However, some functionally important loops/coils were missing when their structure was solved under x-ray crystallography (Bloomer, et al., 1978; Bode, et al., 1978), and some proteins with known biological functions did not possess stable and specific structure in solution, as shown with nuclear magnetic resonance spectroscopy (Williams, 1978). These exceptions are called intrinsically disordered proteins or proteins with intrinsically disordered regions.

2.3 Intrinsically Disordered Proteins

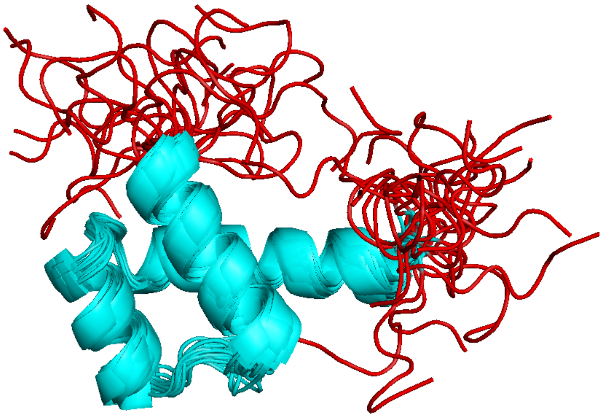

Figure 2.4. An example of 3D structure of IDP. Native structure of protein 1FTT that consists of an ensemble of superimposed conformations; IDRs at the N- and C-termini are shown in red.

The last 3 decades of research resulted in finding a new tribe of proteins which do not possess specific and stable 3D structures (in whole or in part), but which still perform biological functions under the physiologic conditions *in vivo* (Dunker, et al., 2001; Turoverov, et al., 2010; Uversky, 2003; Uversky, et al., 2000; Wright, et al., 1999; Xue, et al., 2010). The members of this novel tribe are known as the intrinsically disordered proteins (IDPs). Their structures take form of dynamic structural ensembles in whole or in part (intrinsically disordered regions (IDRs)), that undergo non-cooperative conformational changes. This means that the positions of their atoms and backbone angles have no specific equilibrium state and they vary (largely) over time. For example, the homeodomain of rat thyroid transcription factor 1 (PDB identifier: 1FTT; DisProt entry: DP00071) is a hybrid of two IDRs and an ordered region, where the disordered regions are located at the N- and C-termini; see the red strings in **Figure 2.4** (Esposito, et al., 1996). In fact, **Figure 2.4** shows that the ordered region highlighted in blue has a stable/ordered structure with three α-helices and the connecting coils; this part of the structure shows relatively small fluctuations between the conformations. At the same time, the conformations at the termini vary substantially in 3D space , i.e., they do not have an equilibrium state.

The existence of IDPs/IDRs does not fit the traditional sequence-structure-function paradigm (Cortese, et al., 2008; Dunker, et al., 2005; Dunker, et al., 2001; Dyson, et al., 2005; Tompa, 2002; Uversky, et al., 2005; Wright, et al., 1999; Xue, et al., 2010). Consequently, the IDPs were assumed to be relatively rare exceptions for a few decades, despite the fact that IDPs were described in scientific literature on multiple occasions. The first paper was published in 1936 (Landsteiner, 1936) and IDPs were rediscovered several times since then (Doolittle, 1973; James, et al., 2003; Jirgenesons, 1966; Karush, 1950; McMeekin, 1952; Pauling, 1940). However, this has changed recently. IDPs are now appreciated as an important class of proteins that is relatively common in nature. For example, Ward *et al.* in 2004 showed that eukaryotes have close to 20% of disordered residues (Ward, et al., 2004). In addition, the conformational plasticity associated with the intrinsic disorder was shown to provide IDPs/IDRs with a wide spectrum of exceptional functional advantages over the functional modes of well-structured proteins (Brown, et al., 2002; Cortese, et al., 2008; Dunker, et al., 2002; Dunker, et al., 2002; Dunker, et al., 2005; Dunker, et al., 1998; Dunker, et al., 2001; Dunker, et al., 1997; Dunker, et al., 2008; Dunker, et al., 2008; Dyson, et al., 2002; Dyson, et al., 2005; Oldfield, et al., 2008; Romero, et al., 2001; Uversky, et al., 2010; Uversky, et al., 2005; Wright, et al., 1999). For instance, this plasticity allows for easier accessibility of IDRs, which enhances their susceptibility to undergo post-translational modifications (e.g., phosphorylation, acetylation, lipidation, ubiquitination, sumoylation, etc.) allowing for “improved” modulation of their biological functions (Uversky, et al., 2010). Many IDRs contain specific identification regions via which they participate in various regulation, recognition, signaling and control pathways (Dunker, et al., 2005; Uversky, et al., 2005). As exemplified by a gene ontology-based analysis, IDPs were found to be involved in numerous biological processes, such as signaling, recognition, and regulation (Cortese, et al., 2008; Dunker, et al., 2002; Iakoucheva, et al., 2002; Vucetic, et al., 2007; Xie, et al., 2007; Xie, et al., 2007). To sum up, IDPs are common in nature and play important roles in all living organisms. Consequently, they have been recognized as important players in modern structural biology and proteomic studies (Dunker, et al., 2001; Tompa, 2002; Uversky, et al., 2000; Wright, et al., 1999).

2.4 Data Sources for the Intrinsic Disorder

Some techniques, such as nuclear magnetic resonance (Bracken, 2001; Dyson, et al., 1998), X-ray crystallography (Choy, et al., 2002; Huber, 1987; Huber, et al., 1983), circular dichroism spectropolarimetry (Adler, et al., 1973) etc., can detect or provide useful information about the intrinsic disorder in proteins. These methods are used either separately or in combination to characterized IDPs and IDRs. DisProt (Sickmeier, et al., 2007) and Protein Data Bank (PDB) (Berman, et al., 2000) are the most popular database that store experimentally characterized IDPs and IDRs. In PDB, a residue is identified as disordered if its atoms have missing coordinates, i.e., they do not establish an equilibrium state in the crystal. Previous studies suggested that 68% of chains in PDB have IDRs (Obradovic, et al., 2003). DisProt is a database that is entirely devoted to the intrinsic disorder. It includes the curated annotations of IDPs/IDRs, along with their experimentally verified cellular functions. Starting with the first release of DisProt in 2005, we summarized the total number IDPs, IDRs and the functionally annotated IDRs (i.e., the IDRs annotated with functions except for "Unknown" and "Disordered region is not essential for protein function") on the annual basis in **Figure 2.5**. This Figure shows a steady growth with two relatively large increases in the depositions in 2006 and 2011. Compared to the increase in the number of annotated IDRs, the number of new proteins added to DisProt is lower. This means that some have multiple IDRs and some previously deposited protein were subsequently annotated with new IDRs. Compared to 2005, about 8.5 times more IDRs are now annotated with functions. This implies that functional analysis of IDRs is gaining momentum. To the best of our knowledge, DisProt is the only database that provides curated functional annotations for IDRs. Until now, 38 functions (http://www.disprot.org/view_function_subclass.php) are found to be associated with IDPs/IDRs.

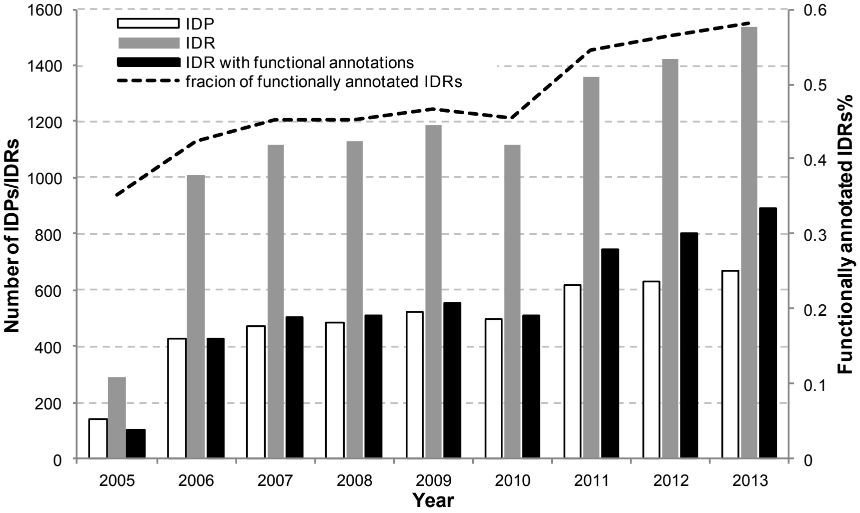

Figure 2.5. Number of entries in DisProt database from 2005 to 2013.
The summary includes the number of entries, such as IDPs, IDRs and functionally annotated IDRs.

2.5 Prediction of Intrinsic Disorder from Protein Sequences

The sequence-to-structure paradigm states that protein structure is encoded by the protein sequence. Assuming that the ensemble of 3D structures that is characteristic to IDPs is a (dynamic) type of structure, this paradigm could be extended to IDPs, i.e., intrinsic disorder could be also potentially determined by the protein sequence. To test this hypothesis, previous studies investigated the relationship between intrinsic disorder and the corresponding protein sequences. The results show that the amino acid composition of IDPs and IDRs is characterized by certain biases, such as low mean hydropathy combined with high mean net charge. These biases determine the highly unstructured and extended state of these proteins and regions, since high net charge leads to strong electrostatic repulsion, and low hydropathy prevents efficient compaction (Uversky, et al., 2000). In agreement with these observations, IDPs and IDRs were shown to be significantly depleted in so-called order-promoting amino acids: C, W, I, Y, F, L, H, V, and N; and substantially enriched in the disorder-promoting residues: A, G, R, T, S, K, Q, E, and P (Dunker, et al., 2001; Radivojac, et al., 2007; Romero, et al., 2001; Vacic, et al., 2007; Williams, et al., 2001). These findings support the hypothesis that intrinsic disorder is defined at the sequence level. In addition to the sequence-derived characteristics, intrinsic disorder is often associated with lack of secondary structures, i.e., enrichment in coils and depletion in helices and strands (Dosztanyi, et al., 2010; Dyson, et al., 2005; Liu, et al., 2002; Romero, et al., 2001; Uversky, et al., 2000).

The abovementioned observations suggest that intrinsic disorder is predictable from the protein sequence. Past two decades have witnessed strong efforts in the development of computational methods for the prediction of disordered regions from protein chains. Since 2002 disorder prediction was included in the biannual Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments (Noivirt-Brik, et al., 2009). By now there are dozens of methods that were developed and published and which are comprehensively summarized in a few recent reviews (Dosztanyi, et al., 2010; He, et al., 2009; Peng, et al., 2012; Uversky, et al., 2010). These disorder predictors allow for high-throughput annotations of disorder.

Based on their runtime, the computational predictors of disorder can be classified into low- or high-throughput. Most of accurate predictors, such as DISOPRED2 (Ward, et al., 2004), ProfBval (Schlessinger, et al., 2006), Norsnet (Schlessinger, et al., 2007), MD (Schlessinger, et al., 2009), PONDR-FIT (Xue, et al., 2010), MFDp (Mizianty, et al., 2010) and CSpritz (Walsh, et al., 2011) etc., are low-throughput and are estimated to take at least 5 minutes to make disorder prediction for an average sized protein with 300 residues on a single processor using a modern desktop computer (Peng, et al., 2014). Their low runtime efficiency is due to the fact that they compute evolutionary profiles extracted from multiple sequence alignment, which usually generated by PSI-BLAST method (Altschul, et al., 1997) against a large protein database, like the NCBI's non-redundant protein database (Pruitt, et al., 2007). Consequently, analysis of IDPs/IDRs with these methods on a proteomic/genomic scale is relatively time-consuming. The high--throughput methods include IUPred (Dosztanyi, et al., 2005), ESpritz (Walsh, et al., 2012) and VSL2 (Peng, et al., 2006). These predictors can make disorder prediction for a single protein in about a second, and thus are capable of predicting intrinsic disorder for entire proteomes.

When focusing on predictive quality, the consensus-based predictors, e.g., MFDp (Mizianty, et al., 2010) and metaPrDOS (Ishida, et al., 2008), are shown to outperform other types of methods (Mizianty, et al., 2010; Peng, et al., 2012; Peng, et al., 2012; Schlessinger, et al., 2009). This is because the consensus-based methods combine several individual predictors, which target prediction of different types of disorder (e.g., short or long disordered regions) and/or are built using data coming from different sources. The three main data sources that are used to generate experimental disorder annotations include protein structures generated through X-ray crystallography, structures generated utilizing nuclear magnetic resonance spectroscopy, and manually curated protein sequences that are stored in the DisProt database (Sickmeier, et al., 2007). Disorder annotations of a given protein could be inconsistent across these three sources (e.g., since disorder could be stabilized and become structured in the crystal structure), and disorder predictors that were developed using disordered regions from one source could be less accurate for the prediction of disorder from the other sources (Schlessinger, et al., 2007).

Overall, high-throughput methods generate disorder predictions for a protein in seconds while consensus-based methods, which are substantially slower, provide more accurate predictions.

2.6 Characterization of Abundance and Functions of Intrinsic disorder

The availability of computational disorder predictors enables large scale investigations of IDPs and IDRs. We summarize the recent progress in the investigations of the abundance and the functional roles of intrinsic disorder, which use the putative disorder generated by computational disorder predictors.

Several efforts have been devoted to estimating the abundance of the intrinsically disordered proteins in nature (Burra, et al., 2010; Dosztanyi, et al., 2006; Dunker, et al., 2000; Feng, et al., 2006; Galea, et al., 2009; Ward, et al., 2004; Xue, et al., 2012; Xue, et al., 2010). In these studies, various algorithms were used to estimate the content of intrinsic disorder in various proteomes or specific protein families. Although the estimated fractions of disordered residues for any given organism were slightly different in these studies (being dependent on the algorithms used to predict the disorder content), the general trend over the tree of life was quite consistent. The eukaryotes were systematically predicted to have much higher intrinsic disorder content than the prokaryotes. The numbers of species analyzed in these studies ranged from a few to a few hundreds. For example, the abundance of IDPs and IDRs in 53 archaean species was recently evaluated (Xue, et al., 2010). In another recent study, Burra et *al*. analyzed 332 prokaryotic proteomes (Burra, et al., 2010), and in still another recent work (which, to the best of our knowledge, is the largest scale intrinsic disorder analysis undertaken so far) the proteomes of 3484 species from three kingdoms of life (archaea, bacteria and eukaryotes) and from viruses were analyzed (Xue, et al., 2012).

In addition to studies of the abundance of the protein disorder in various proteomes, the cellular functions of IDPs and IDRs at the proteome/large protein database level were also scrutinized. For example, Ward *et al*. analyzed distribution of IDPs in six archaean, thirteen bacterial and five eukaryotic genomes and studied the function of proteins with long predicted regions of disorder using the gene ontology annotations in the *Saccharomyces* genome. They have shown that proteins that have disorder are often located in the cell nucleus and are involved in the regulation of transcription and cell signaling, and are commonly associated with kinase activity and nucleic acid binding (Ward, et al., 2004). Based on the bioinformatics analysis of the functional keywords associated with 20 or more proteins in Swiss-Prot, a few recent studies have shown that many cellular functions are associated with the increased propensity for intrinsic disorder. Out of 710 considered Swiss-Prot keywords, 310 functional keywords were found to be associated with ordered proteins, 238 were attributed to the disordered proteins, and the remaining 162 yield ambiguity in the function-structure associations (Vucetic, et al., 2007; Xie, et al., 2007; Xie, et al., 2007). Study of the occurrence of protein disorder in the human proteome and analysis of the ontology categories that are enriched in the disordered human proteins revealed that the IDP-specific functions are both length- and position-dependent and these observations were used to develop predictors of for human protein functions(Lobley, et al., 2007). Moreover, inclusion of the disorder information improved the prediction accuracies for 26 GO categories related to signaling and molecular recognition (Lobley, et al., 2007). Recently, analysis of human proteome revealed that disordered regions frequently act as independent functional units (Pentony, et al., 2010), and this functional modularity supported the earlier notion of an association between disorder and alternative splicing (Romero, et al., 2006).

The abovementioned studies demonstrate that IDPs/IDRs are common across the three kingdoms of life (i.e., eukaryota, bacteria and archaea), where eukaryotes have higher amounts of IDPs/IDRs. In addition, IDPs/IDRs were shown to be functionally important in many biological processes. For example, intrinsic disorder have been implicated to play an important role in molecular recognition (Dunker, et al., 2005; Dunker, et al., 2001; Dyson, et al., 2002; Iakoucheva, et al., 2002; Oldfield, et al., 2005; Uversky, et al., 2005).

2.7 Molecular Recognition Features

An important cellular function of intrinsic disorder is molecular recognition (Dunker, et al., 2005; Dunker, et al., 2001; Dyson, et al., 2002; Iakoucheva, et al., 2002; Oldfield, et al., 2005; Uversky, et al., 2005). Molecular recognition is the biological process that refers to interactions between two or more molecules to form complexes. Intrinsic disorder in molecular recognition represents a specific type of IDRs, which was suggested to be common in proteomes (particularly in eukaryotes) and to be involved in signaling and regulatory functions (Mohan, et al., 2006; Oldfield, et al., 2005; Uversky, et al., 2010; Vacic, et al., 2007). These IDRs are known as molecular recognition features (MoRFs), and they are defined as short (5 to 25 amino acids) disordered regions and undergo disorder-to-order transition upon binding to protein partners (Mohan, et al., 2006; Oldfield, et al., 2005). MoRFs are divided into four subtypes: α-MoRFs (that fold into α-helices), β-MoRFs (that fold into β-strands), γ-MoRFs (coils) and complex-MoRFs (mixture of different secondary structure), based on the predominant content of helix, strand, or coil structures in the bound state (Mohan, et al., 2006; Oldfield, et al., 2005; Vacic, et al., 2007).

To our best knowledge, four computational methods that predict MoRFs are available. They include α-MoRF-PredI (Oldfield, et al., 2005), α-MoRF-PredII (Cheng, et al., 2007), ANCHOR (Dosztanyi, et al., 2009; Meszaros, et al., 2009), and MoRFpred (Disfani, et al., 2012). Both of α-MoRF-PredI and α-MoRF-PredII are limited to the prediction of α-MoRFs, where α-MoRF-PredII extends α-MoRF-PredI. ANCHOR predicts general protein-binding regions that located in IDRs. This means the prediction ANCHOR includes putative MoRFs and also other (i.e., longer) MoRF-like regions. MoRFpred is a leading method that is focused on the prediction of MoRFs, including all four subtypes of MoRFs.

2.8 Computational Methods

Prior studies indicated that intrinsic disorder is common in nature and plays crucial functions *in vivo*. However, the curated annotations and especially the functional annotations of intrinsic disorder lag behind the rate with which new protein sequences are accumulated. The experimental data, e.g., disorder annotations and/or functional annotations from DisProt, provides invaluable and well-annotated source of information that can be used to develop computational methods. Such methods would use the annotated with intrinsic disorder data to learn mapping from sequence to the annotation and then this mapping could be used to characterize intrinsic disorder on large scale in proteins with known sequences. In this thesis, we utilize and develop computational methods that include the abovementioned predictors (see Chapter 4 and Chapter 5) and also methods for statistical analysis (see Chapter 3 to Chapter 5); the latter are primarily used to assess predictive performance. Therefore, we introduce background related to prediction and statistical analysis.

2.8.1 Prediction

Prediction is a problem of identifying categories (also called labels or classes) for a new (test) sample based on a model that maps samples to categories which is learned/trained from a set of samples with known categories (training dataset). In our work, categories include intrinsically disordered vs. ordered residues, IDRs vs. ordered regions, or cellular functions of IDRs such as interactions between protein and RNA, DNA, and other proteins; the samples are amino acids (residues), protein chains, or protein regions. The design of the prediction model usually consists of the following four steps:

1. **Feature representation**. This step aims to represent each sample by a set of observed and quantifiable properties, known as features. Features can be categorical, ordinal, integer-valued and/or real-valued. The categorical feature takes on one of a limited and a fixed number of possible values, like "A", "B", "AB" or "O" for blood type. The ordinal feature is defined as an arbitrary numerical scale where the exact numerical quantity of a particular value has no significance beyond its ability to establish a ranking over a set of data points. Such as "$0-$10000", "$10000-$30000", "$30000-$50000" or ">$50000" to measure individuals' income. The integer- or real-valued feature is an integer and a real number, respectively. For example, the number of occurrences of a keyword in a paragraph of text is an integer-valued feature; body temperature of a patient is a real-valued feature. The feature representation (feature set) must be the same for different samples in the same dataset, which is particularly challenging when dealing with proteins or protein regions, as they have different length and must be converted into a fixed-size feature set.
2. **Feature selection**. Choosing discriminative/relevant and independent/non-redundant features is an important step to design an accurate prediction model. This can be accomplished with feature selection, which is a process of selecting an best-performing (with respect to predictive quality) subset of features that were developed in the first step. The selection includes removing irrelevant and/or redundant features. The irrelevant features provide no discriminative information to identify the categories. Redundant features duplicate information provided by some other features that are already selected to develop the prediction model. Feature selection may result in improved predictive performance and reduced runtime when compared with using all features. Feature selection techniques can be broadly divided into filter, embedded, and wrapper approaches (Saeys, et al., 2007). We describe wrapper-based feature selection as we used them in this thesis. The wrapper-based approaches search for the best-performing subset of features by embedding the prediction. Specifically, a search in the space of possible feature subsets is defined, and various possible subsets of features are evaluated by using them to perform prediction. The feature subset that provides the best predictive performance is selected. The motivating advantage of wrapper-based methods is

that they consider both the redundancy and the relevance of features and they maximize predictive performance of the underlying prediction task.

3. **Model construction**. This step derives the relationship (mapping) between selected features and the corresponding known categories using samples from the training dataset. An example mapping could be to predict a diagnosis by linearly combining the observed features, such as gender, blood pressure, presence or absence of certain symptoms, etc., where coefficients of the linear mapping are based on the importance of a given feature to the diagnostic outcome. Model construction is an essential step, which requires making vital choices with respect to selection of a suitable model type and ways to combine features. Many different types of models can be used, most of which have certain requirements with respect to the input features and number of samples. Some models can only be used with categorical or ordinal features, some other only with real-valued features, and some with all types of features. Some models take a long time to compute and long time to make predictions. We focused on models that can be computed and used for prediction in the high-throughput fashion. A popular type of model is phrased as a linear function that assigns a score to each possible category *k* by linearly combining the input features that represent a given sample. In that case, the predicted/identified category is the one with the highest score generated by such linear model. Examples of such models include support vector machines, logistic regression, and linear discriminant analysis. Since in this thesis we used logistic regression to construct predictive model in Chapter 5, we describe this model in greater detail in section 2.8.2.
4. **Model validation**. This step quantifies the predictive performance of a given predictive model. The performance is measured by comparing the results of the predictive model with the observed categories (i.e., native/true annotations) for a set of samples. This is a necessary step, unless the model is very well understood, for instance, we no longer would validate DNA base pairing rules. The validation must be performed out of sample (using data that was not used to construct the model) to assure that the model does not overfit (too closely mimic) the training dataset. Two types of out of sample protocols are usually used: cross-validation (see section 2.8.3) and test on an “independent” (using samples that do not duplicate

and are different from samples in the training dataset) test dataset. The predictive quality can be quantified using several evaluation criteria, which are discussed in section 2.8.4. The differences between predictive performances of different models are usually assessed using statistical tests of significance, which are described in section 2.8.5.

2.8.2 Logistic Regression

Logistic regression is a probabilistic model, which is based on linear mapping between features and categories , and which outputs a propensity score that quantifies probability of a sample to be predicted as a given category. The logistic regression utilizes logistic function

$$f(t) = \frac{1}{1+e^{-t}},$$

where t is a linear combination of features, i.e., $t = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m$, x_i is a feature, a_i is a coefficient estimated using training dataset for feature x_i ($i = 1, 2, \dots , m$) (to minimize prediction error on the training dataset), and $f(t)$ is the predicted propensity score (i.e., probability) that always takes the value between 0 and 1. Here we apply ridge estimator (Cessie, et al., 1992), which is implemented in Weka platform (Hall, et al., 2009), to estimate coefficients.

2.8.3 Cross-validation

Cross-validation is a procedure used to estimate predictive performance of a given model using the training dataset. In *k*-fold cross-validation, a given training dataset is randomly partitioned into *k* equal size subsets/folds. Predictive model is designed and trained on *k*–1 of these subsets/folds, and is validated on the remaining *k*$^{\text{th}}$ subset/fold. This process is repeated *k* times, each time choosing a different subset/fold to perform validation of the predictive model. We usually report an average score of a certain evaluation criteria (see section 2.8.4) over the *k* subsets. Cross-validation is usually used to design the model (e.g., perform feature selection, choose the best-performing type of model, etc.) since this procedure help in avoiding overfitting the model into the training dataset.

2.8.4 Evaluation Criteria

In this thesis, we perform and evaluate predictions per-residue (for every predicted amino acid) considering two types of predictions: binary and real-valued outcomes. Real-value outcome, i.e., propensity score, quantifies the probability of a sample to be a given category. Binary outcome is usually represented by '1' or '0', which indicates whether a given sample is predicted to be in a given category (e.g., is or is not an RNA--binding residue located in IDRs). The binary outcome is often derived from the real-valued outcome by thresholding, i.e., samples with real-valued outcomes above a given threshold are assumed to be predicted in one category (say, RNA-binding), while sample with values below the threshold in the other category (say, not binding to RNA). The assessment for binary outcomes evaluates the ability of a predictive model to correctly identify categories. The evaluation of the real-valued outcomes reveals how well the propensity scores quantify the relationship between input features and the output categories.

In binary evaluation, each sample is defined as positive or negative, based on its observed category. In our case, positives are curated/native binding residues, i.e., RNA-, DNA- or protein-binding residues located in IDRs, and negatives are all other residues. For example, if curated/native disordered RNA-biding residues are assumed as positives, then all other residues including the remaining annotated disordered and ordered residues are negatives. There are four possible outcomes of the binary prediction: true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), where TP/TN is the number of correctly predicted positives/negatives, and FP/FN is number of negatives/positives that were predicted as positives/negatives. These outputs are summarized with a confusion matrix; see **Table 2.2**. The predictive quality is estimated by a variety of scores computed from this matrix. The commonly used scores include sensitivity and specificity, especially when the number of positive and negative samples is unbalanced, i.e., one category is much more abundant in the dataset than the other(s).

Table 2.2. Confusion matrix.

		Predicted categories	
		Positive	Negative
Curated categories	Positive	TP	FN
	Negative	FP	TN

Sensitivity and specificity are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{Positive}$$
$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{Negative}$$

Sensitivity and specificity quantify the predictive performance for positives and negative samples, respectively. A higher value of sensitivity/specificity indicates that more positive/negative samples are predicted correctly.

The real-valued propensity scores are usually assessed with the receiver operating characteristic (ROC) curve. Specifically, using each propensity score *p* (between 0 and 1) as a threshold, all predictions with propensities ≥ *p* are set as predicted positives, and all other residues are set as predicted negatives. Next, the TP-rate = TP/(TP + FN) and the FP-rate = FP/(FP + TN) are calculated, and ROC curve is plotted by connecting all points that correspond to all pairs (over all values of *p*) of FP-rate and TP-rate. The area under the ROC curve (AUC) is used to quantify the predictive performance. Higher AUC value indicates better predictive performance. With high AUC values, the higher/lower propensity score are likely to correctly indicate that the predicted sample is positive/negative.

2.8.5 Statistical Tests

A statistical test is a method with a pre-defined null hypothesis H_0 that assumes a general or default pattern based on conventional wisdom. However, this null hypothesis H_0 might be wrong, e.g., the pattern is just random or follows other rules than assumed. The statistical test usually computes *p*-value, a probability that the null hypothesis is actually correct. If *p*-value turns out to be less than a certain significance level, often 0.05 or 0.01, the null hypothesis H_0 is rejected. Such result indicates that the observed result would be highly unlikely under the null hypothesis. There are different tests for different types of data. Here, we describe Student's t-test and Wilcoxon signed-rank test in detail, because we utilize them in Chapter 3 to investigate the difference of intrinsic disorder between two protein sets, and in Chapter 5 to compare predictive performance between pairs of considered computational prediction methods.

Student's t-test evaluates differences between means of two normally distributed variables (e.g., values of predictive performance for two methods). The null hypothesis is that the means are equal. The test is defined by the following equation:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where:

$S_{X_1X_2} = \sqrt{\frac{(n_1-1)S_{X_1}^2+(n_2-1)S_{X_2}^2}{n_1+n_2-2}}$ is the pooled standard deviation;

$S_{X_i}^2 = \frac{1}{n_i}\sum_1^{n_i}(x_k - \bar{X}_i)$ is the variance of variable i $(i = 1, 2)$;

$\bar{X}_i = \frac{1}{n_i}\sum_1^{n_i} x_k$ is the mean of variable i $(i = 1, 2)$;

x_k $(k = 1, 2, \dots, n_i)$ is the k^{th} value of variable i $(i = 1, 2)$;

n_i is the number of values in variable i, $(i = 1, 2)$; i.e., sample size.

Once the *t*-value is determined, the significance can be found using a table of *t*-values, i.e., *t*-table, from the Student's t-distribution; see **Table 2.1**. The *t*-value that is specific to a given pre-defined significance α (i.e., the chosen *p*-value) and a given value of degree of freedom (DF) (i.e., *n*-1, *n* is the sample size) is known as the critical value. If the calculated *t*-value is greater than or equal to this critical value, then the compared two variables are significantly different, i.e., the null hypothesis H_0 is rejected at the level of the pre-defined significance (i.e., the chosen *p*-value); otherwise they are not significant and H_0 is accepted.

Table 2.3 Relevant fragment of the *t*-table.
DF represents degrees of freedom. We show critical *t*-values for the two *p*-values considered in this thesis, 0.05 and 0.01.

p-value DF	0.05	0.01
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Wilcoxon signed-rank test first ranks all absolute differences between two variables (e.g., values of predictive performance for two methods) and then compares the mean ranks of these variables. This test is usually used as an alternative to the Student's t-test when the distribution is not normal. The hypothesis is that the median difference between a pair of variables is zero. The test statistic is defined as the absolute value of the sum of the signed ranks:

$$W = \left| \sum_{i=1}^{N_r} \left[sgn(x_{2,i} - x_{1,i}) \cdot R_i \right] \right|$$

Where:

R_i is the rank of pair (ranked by absolute difference, ties receive a rank equal to the average of the ranks they span);

$x_{1,i}$ and $x_{2,i}$ denote the i^{th} values of the two variables;

N_r is the number of pairs with non-zero difference.

As N_r increases, especially with N_r≥10, the sampling distribution of *W* converges to the normal distribution. In this case, *z*-score can be calculated by using the following equation:

$$z = \frac{W - 0.5}{\sigma_W}, \sigma_W = \sqrt{\frac{N_r(N_r + 1)(2N_r + 1)}{6}}$$

When N_r is smaller than 10, the calculation of *z*-score must be based on the exact sampling distribution of *W*. The critical value z_{critical} of *z*-score is determined, by using a reference table similar to the t-table (see Student's t-test). If *z*-score is greater than z_{critical}, then there is significant difference at the level of significance corresponding to z_{critical}; otherwise, there is no difference.

Anderson-Darling normality test Student’s t-test assumes that the tested variables follow normal distribution, which has to be checked. This can be performed with the Anderson-Darling normality test. In fact, Anderson-Darling test can be used to test whether variables come from a specified type of distribution, such as normal, uniform, exponential distribution, and so on. In this work we focus exclusively on the normal distribution. In addition, we assume that both the mean and the variance of a variable are unknown, and we calculate them from sample the data. Test statistic A^2 is then defined by the following equation:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\big(\ln\Phi(Y_i) + \ln\big(1 - \ \Phi(Y_{n+1-1i})\big)\big)\,, Y_i = \frac{X_i - \bar{X}}{\sigma}$$

Where:

X_i is the *i*th value of variable *X*;

$\bar{X}$ is the mean of variable *X*;

σ is the standard deviation of variable *X*;

$\Phi(Y_i)$ is a cumulative distribution function of Y_i for normal distribution;

n is number of values of variable *X*.

Chapter 3

Systematic Analysis of Intrinsic Disorder in Complete Proteomes

3.1 Motivation

IDPs and IDRs are devoid of stable 3D structures, but they possess crucial cellular functions. In last few decades there was a substantial progress in the field of the intrinsic disorder; see section 2.6. However, the modern studies on the natural abundance and cellular functions of IDPs/IDRs are limited in terms of the number of species analyzed and scope of the analysis, which often targets only one of a handful of aspects. For example, the largest-scale characterization of intrinsic disorder on 3484 proteomes only focused on the natural abundance of intrinsic disorder and the relations between intrinsic disorder and organism complexity (Xue, et al., 2012). By using 875 keywords Swiss-Prot, Xie et al., and Vucetic et al., performed the broadest investigation of the functional roles and cellular localization of intrinsic disorder, respectively (Vucetic, et al., 2007; Xie, et al., 2007). However, this study aggregated the results over all collected proteins, i.e., the authors did not take into account difference of the intrinsic disorder across species and kingdoms of life. As a result, we are yet to enjoy a comprehensive study that characterizes intrinsic disorder across kingdoms of life and species in terms of a detailed disorder profile and cellular functions and localizations of the intrinsic disorder. Therefore, we performed a large-scale comprehensive and detailed analysis of 6,438,736 proteins from 965 complete proteomes using arguably more accurate consensus-based disorder predictions, when compared with the prior studies. Since in addition to the analysis of 59 archaean, 471 bacterial and 110 eukaryotic proteomes we studied ~20,000 proteins from 325 viral proteomes, our work represents one of the first large scale analysis of abundance and functions of the intrinsic disorder in viruses. Here, viruses were considered as a fourth kingdom of life (besides eukaryota, bacteria, and archaea), although currently there is no common opinion on whether viruses are a form of life, or organic structures that interact with living organisms. We seamlessly combine proteome-level analysis that characterizes abundance and differences in profiles of disorder between the kingdoms of life with the analysis at the protein level that concerns a detailed, large-scale, and comprehensive characterization of functional roles and cellular localization of the intrinsic disorder. We are the first to investigate enrichment of disorder in a broad range of over 1200 functional annotations, compared to previous "small-scale" studies that investigated a narrower range of functional aspects based on at most a couple dozen of proteomes that excluded viruses. We include a similarly comprehensive characterization of enrichment of disorder in cellular components/compartment in archaea, bacteria, eukaryota and viruses and, for the first time, map intrinsic disorder into archaea, bacterial, eukaryotic cells. We reveal an interesting observation based on first-of-its-kind proteomic level analysis of a relation between intrinsic disorder and evolutionary pace.

3.2 Materials and Methods

3.2.1 Proteomic Level Dataset

We analyzed all 965 complete proteomes, which total to 6,438,736 proteins, from release 2011_08 of UniProt (Consortium, 2012). The proteomes were assigned to their taxonomic lineage based on NCBI (Geer, et al., 2010), where the lowest taxonomic level referred as species could be genus, family or species. The resulting UniProt Complete Proteome Dataset (UCPD) is composed of 3.6% proteins from 59 species in archaea, 66.6% proteins from 471 species in bacteria, 29.5% proteins from 110 species in eukaryota, and 0.3% proteins from 325 viral proteomes. The 965 proteomes are used to characterize disorder at the kingdom level, while 225 small proteomes (with less than 30 proteins) are excluded when performing analysis at the species level.

3.2.2 Disorder Prediction

To obtain putative disordered residues and regions, we applied two highly-efficient disorder predictors, IUPred (Dosztanyi, et al., 2005; Dosztanyi, et al., 2005) and ESpritz (Walsh, et al., 2012), which were shown to provide good predictive quality (Dosztanyi, et al., 2005; Peng, et al., 2012). IUPred has two versions that were designed for predictions of long and short disordered regions, respectively. ESpritz has three versions that considered disorder annotations based on the X-ray crystal structures, nuclear magnetic resonance structures, and the experimental annotations from DisProt (Sickmeier, et al., 2007). IUPred and ESpritz cover the main characteristics of disorder including the two types of IDRs (short vs. long) and the three sources of disorder annotations. The resulting five predictions were combined together using the majority vote consensus based on the fact that consensus-based approaches provide improved predictive quality (Peng, et al., 2012). The used of the consensus-based approach is a marked improvement over the previous studies that utilized only one (Ward, et al., 2004; Xue, et al., 2012) or two (Burra, et al., 2010; Xue, et al., 2010) predictors to characterize disorder. The putative disorder was used to calculate the disorder content (i.e., fraction of disordered residues in a given chain), the number and size of IDRs and long IDRS, where long IDRs consists of at least 30 consecutive disordered residues. The inclusion of long IDRs is motivated by the fact that they are implicated in protein-protein recognition (Tompa, et al., 2009) and serve as functional units (Pentony, et al., 2010). Consistent with previous studies (Monastyrskyy, et al., 2011; Noivirt-Brik, et al., 2009), we count the IDRs with at least four consecutive disordered residues. The count of IDRs was normalized by a unit of protein chain length (100 AAs) to accommodate for the bias due to differences in chains length between kingdoms.

3.2.3 Analysis of Enrichment of Disorder

We also investigate disorder across various cellular components (i.e., the localization of proteins) and relations between disorder and protein functions based on the Gene Ontology (GO) annotations (Ashburner, et al., 2000) that are linked in UniProt. We removed the annotations with insufficient number of samples in a given kingdom, i.e., the functions/components with less than 100 chains. In each kingdom, we empirically analyze whether disorder is significantly enriched/depleted in proteins with a given function and in a given cellular component. Similar to earlier analysis (Ward, et al., 2004), we evaluate statistical significance of enrichment/depletion by contrasting disorder content in a given functional or localization-based set of chains with the baseline disorder content in a given kingdom; this accommodates for differences in the abundance of disorder between the kingdoms. We randomly select half of the GO-annotated chains and compare them with the same number of chains drawn at random from the entire kingdom. This is repeated 10 times and we evaluate significance of the differences in the disorder content between these two vectors. If the measurements are normal, as evaluated with the Anderson-Darling test (Anderson, et al., 1952) at 0.05 significance, then we utilize the t-test; otherwise we use the non-parametric Wilcoxon rank sum test (Wilcoxon, 1945). As defined in section 2.8.5, the differences in the disorder content are assumed to be significant, when the *p*-value < 0.05. We consider only the Significant differences with sufficiently large magnitude, i.e., the average difference/enrichment must be larger than 50% of the average disorder content in a given kingdom.

3.2.4 Evolutionary Pace

Using the evolutionary tree reconstructed in (Ciccarelli, et al., 2006), we study relation between the intrinsic disorder and the evolutionary speed, which is quantified with the branch length, i.e., longer branches indicate faster pace of the sequence evolution. We mapped 112 bacterial, 14 eukaryotic and 2 archaea species into our dataset from among 191 species that were used in (Ciccarelli, et al., 2006), and compared their disorder content against the branch length. Consequently, we had to exclude viruses that were not considered in (Ciccarelli, et al., 2006) and archaea that had small sample size.

3.2.5 Sequence Conservation

The sequence conservation was quantified using relative entropy (Wang, et al., 2006), which was computed from the Weighted Observed Percentages (WOP) profiles produced by PSI-BLAST (Altschul, et al., 1997). PSI-BLAST was run with default parameters (-j 3, -h 0.001) against the nr database that was filtered using PFILT (Jones, et al., 2002) to remove low-complexity regions, trans-membrane regions and coiled-coil regions. Due to the high computational cost, we estimated conservation for a given proteome based on results for 100 randomly selected proteins from that proteome.

3.3 Results and Discussion

3.3.1 Abundance of Intrinsic Disorder across Kingdoms of Life

First, we analyzed the overall abundance of intrinsic disorder in 6,438,736 proteins from 965 complete proteomes. Results of this analysis for selected proteomes are shown in **Figure 3.1**, which represents the averaged disorder content (**Figure 3.1A**) and the normalized number of long (30 or more consecutive amino acids) IDRs (**Figure 3.1B**) across different phyla (second level of the taxonomic lineage) and kingdoms of life. This analysis revealed that intrinsic disorder is common in all the proteomes studied and that the eukaryotic proteomes are noticeably more disordered than proteomes from other kingdoms of life using different disorder measures. In fact, disorder content is at 20.5% for eukaryotes, 13.2% for viruses, 8.5% for bacteria, and 7.4% for archaea. Furthermore, the normalized number of long disordered regions per 100 amino acids is at 17.4% for eukaryotes, 10% for viruses, 4.2% for bacteria, and 3.6% for archaea. We note the relatively smaller proportions for the bacteria and archaea, which means that they have relatively fewer long IDRs. The results of our analysis are consistent (a bit higher but in the same order) with the results of earlier analysis performed for a smaller set of proteomes (6 archaean, 13 bacterial, and 5 eukaryotic proteomes) (Ward, et al., 2004), where the disorder content was estimated to be 18.9% in eukaryotes, 5.7% in bacteria, and 3.8% in archaea. **Figure 3.1** also shows that the disorder content in viral species varies in a widest extent, ranging between 3% and 55%, in eukaryotic species between 5% and 35%, and in bacterial and archaean species the disorder contents are below 20 and 21%, respectively (whiskers show the range). Also, the fraction of the long IDRs is proportional to the overall disorder content except for some viruses that contains relatively more of longer IDRs (whiskers are taller when compared to the content whiskers).

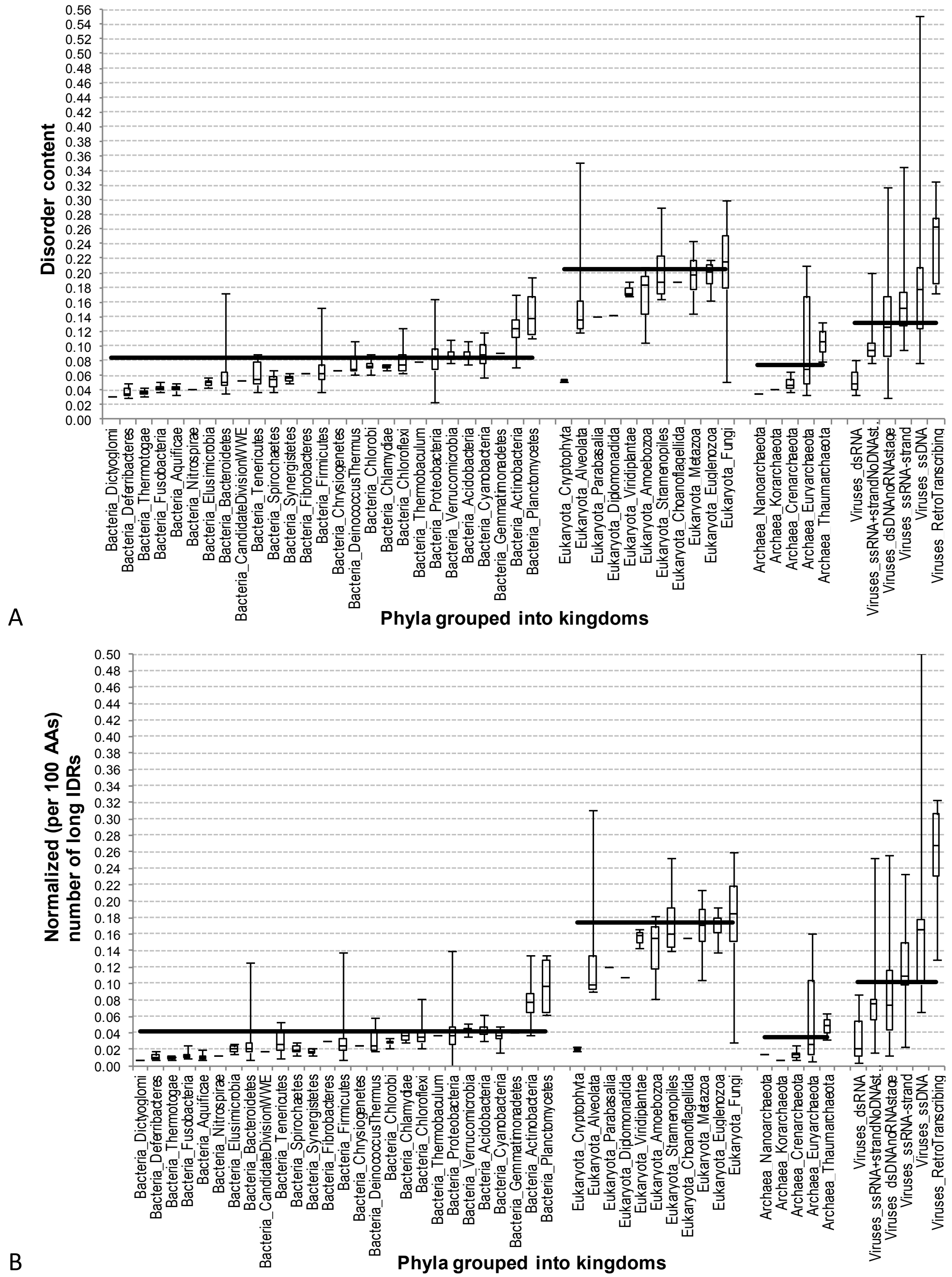

Figure 3.1 Overview of disorder content and long IDRs across phyla and kingdoms.
Disorder content (panel A) and normalized number of long (30 or more consecutive AAs) IDRs across different phyla (second level of the taxonomic lineage) and kingdoms. The phyla (*x*-axis) are grouped into kingdoms, including bacteria, eukaryota, archaea, and viruses. Solid horizontal lines denote average disorder content per kingdom. Box plots show the minimum, first quartile, second quartile (median), third quartile, and maximum disorder content (panel A) or normalized number of long disordered regions (panel B) across different species in a given phyla/kingdom; one line is shown for phyla with only one species (e.g., *Dictyoglomi*).

Next, we looked at the peculiarities of disorder distribution in the four kingdoms of life. **Figure 3.2A** shows that the majority of proteins in viral, bacterial, and archaean species have relatively small amounts of disorder. In fact, 79, 77, and 63% of chains in archaean, bacterial, and viral proteomes, respectively, have up to 10% disorder, compared to only 46% such proteins in eukaryotes. On the other hand, eukaryotic proteomes are characterized by a large fraction of chains with substantial amounts of disorder. Here, 36% of eukaryotic chains are characterized by >20% disorder and 12% of eukaryotic proteins possess >50% disorder.

Figure 3.2B illustrates another interesting fact, namely that in the bacterial and archaean species, the larger amounts of disorder are present only in short chains (shorter that 100 residues long). Specifically, 12% and 11% of proteins in archaea and bacteria, respectively, which are shorter than 100 residues, have on average 19 and 24% of disorder, which is almost three folds higher than their overall average. To compare, chains longer than 100 residues, which account for 88% of archaean and 89% of bacterial proteins, have on average below 6% of disorder. This is in contrast to viruses and eukaryotes, where the disorder is more evenly distributed across protein sizes. Specifically, chains longer than 100 residues, which account for 82 and 93% of proteins in viruses and eukaryotes, respectively, have the average amount of disorder at 12 and 20%, respectively, which is comparable with their overall disorder contents. Chains longer than 500 amino acids in eukaryotes, which total to 32% of eukaryotic proteins, have on average 22% of disorder, compared to 9% in viruses, 6% in bacteria, and 5% in archaea.

As it evident from **Figure 3.2C**, short (below 10 AAs) IDRs account for two-thirds of the IDRs in archaea and bacteria. This noticeably exceeds the corresponding values of 55 and 43% evaluated for viruses and eukaryotes, respectively. Only eukaryotes and viruses have relatively large fractions of longer IDRs, which results in the bimodal distribution in **Figure 3.2C**. More specifically, 25 and 16% of IDRs in eukaryotes and viruses, respectively, are longer than 30 residues, compared to just 7% in bacteria and archaea.

Our analysis revealed that between 0.9% of proteins in eukaryotes (close to 18 thousands) and 0.2% of proteins in archaea (around 500 chains) are fully disordered (i.e., all residues are disordered in entire chain). **Figure 3.2D** shows that the fully disordered proteins in archaea and bacteria are relatively short compared to those in eukaryotic and viral proteomes. In fact, in archaea and bacteria, 86 and 89% of fully disordered chains are shorter than 100 residues, compared to 53 and 52% in viruses and eukaryota, respectively. Interestingly, 20% of fully disordered viral proteins are longer than 300 AAs, compared to 8, 1, and 1% for eukaryotes, archaea, and bacteria, respectively.

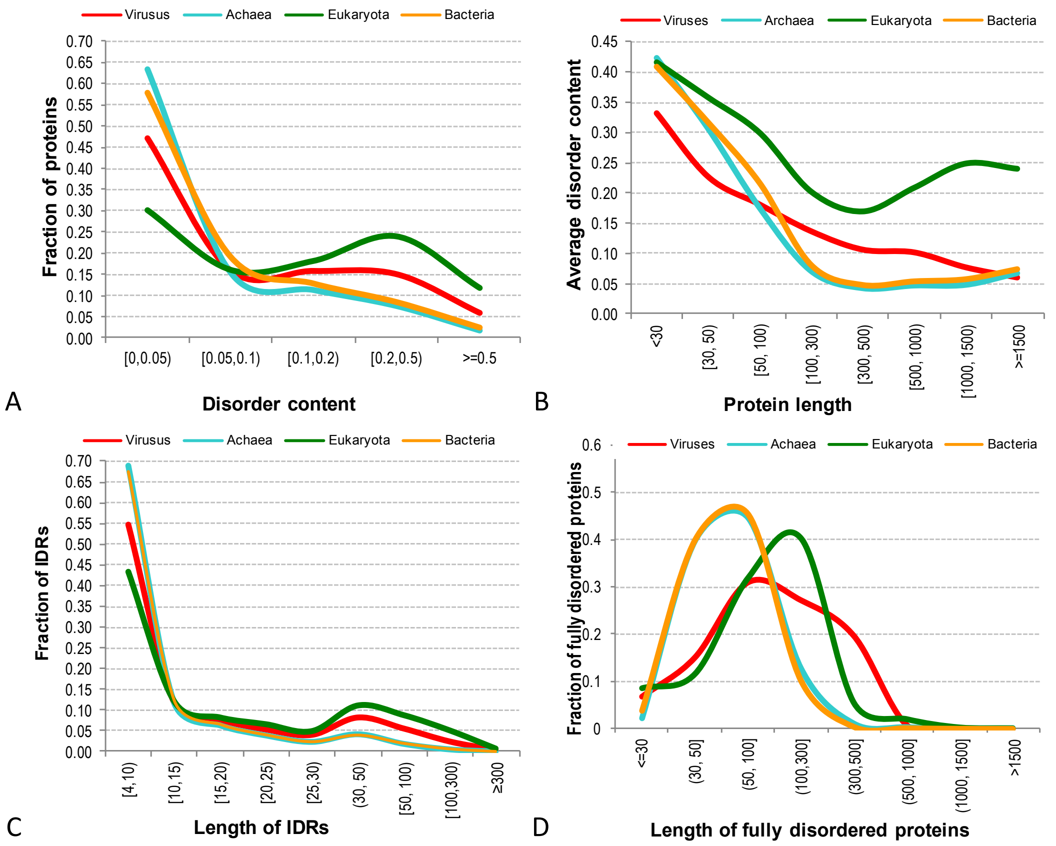

Figure 3.2. Distribution of disorder content, length of IDRs, and fully disordered proteins. Distribution of disorder content (panel A); disorder content against chain size (panel B); IDRs' length (panel C), and length of the fully disordered proteins (panel D) across the four kingdoms, including bacteria, eukaryota, archaea, and viruses.

3.3.2 Functional Roles of Intrinsic Disorder across Kingdoms of Life

The functional importance of IDPs/IDRs was investigated by considering correlations between the intrinsic disorder propensity and GO annotations of biological processes and molecular functions that are available in UniProt database for the considered complete eukaryotic, bacteria, archaean and viral proteomes. Results of these analyses are summarized in **Table 3.1** and **Figure 3.3**. **Table 3.1** suggests that the number of functional annotations does not reflect the complexity of a given kingdom, rather it is correlated with the completeness of its GO annotations. For example, there are 0.22 to 2.61 times more GO annotations in bacteria, compared to eukaryota. Moreover, in each kingdom of life, there are some GO annotations that are enriched in disorder and some other that are characterized by a significant depletion in disorder. For instance, 4 to 10% biological processes, molecular functions and cellular component in eukaryotes are significantly enriched in disorder, whereas in bacteria, about 20% of GO annotated cellular components are enriched in disorder.

Table 3.1. Summary of GO annotations.
Summary of the biological processes, molecular functions, and cellular components, which were annotated based on GO, across the four kingdoms. The numbers in bold indicate the total number of significant sub-functions in a given kingdom that are used to investigate potential depletion or enrichment of the disorder.

Annotation	Types of annotations	Archaea	Bacteria	Eukaryota	Viruses
Biological processes	**Total # of processes**	**12**	**318**	**104**	**2**
	# of processes with significant depletion in disorder	0	76	31	0
	# of processes with significant enrichment in disorder	1	14	10	1
Molecular functions	**Total # of functions**	**34**	**581**	**161**	**4**
	# of functions with significant depletion in disorder	1	184	63	0
	# of functions with significant enrichment in disorder	2	20	6	1
Cellular components	**Total # of components**	**6**	**61**	**50**	**5**
	# of components with significant depletion in disorder	0	12	6	0
	# of components with significant enrichment in disorder	1	13	3	2

Figure 3.3 provides a more detailed representation of correlation between intrinsic disorder and functions in the four kingdoms of life. Disorder-enriched biological processes in eukaryotes include transcription, regulation of GTPase, nucleosome assembly (Peng, et al., 2012), and RNA splicing. Overall, disorder in eukaryotes seems to be important for the protein-RNA, protein-DNA, and protein–nucleotide interactions. In addition to sharing similarities to eukaryotes with respect to disorder-based protein-DNA interactions, bacteria utilize a wider array of biological processes with enriched

disorder, with most illustrative examples being sporulation, protein polymerization, translation, catabolic and metabolic processes, pathogenesis, and chromosome condensation. Disorder in archaea and viruses is suggested to be involved in translation and the interspecies interaction between organism, respectively (see **Figure 3.3A)**.

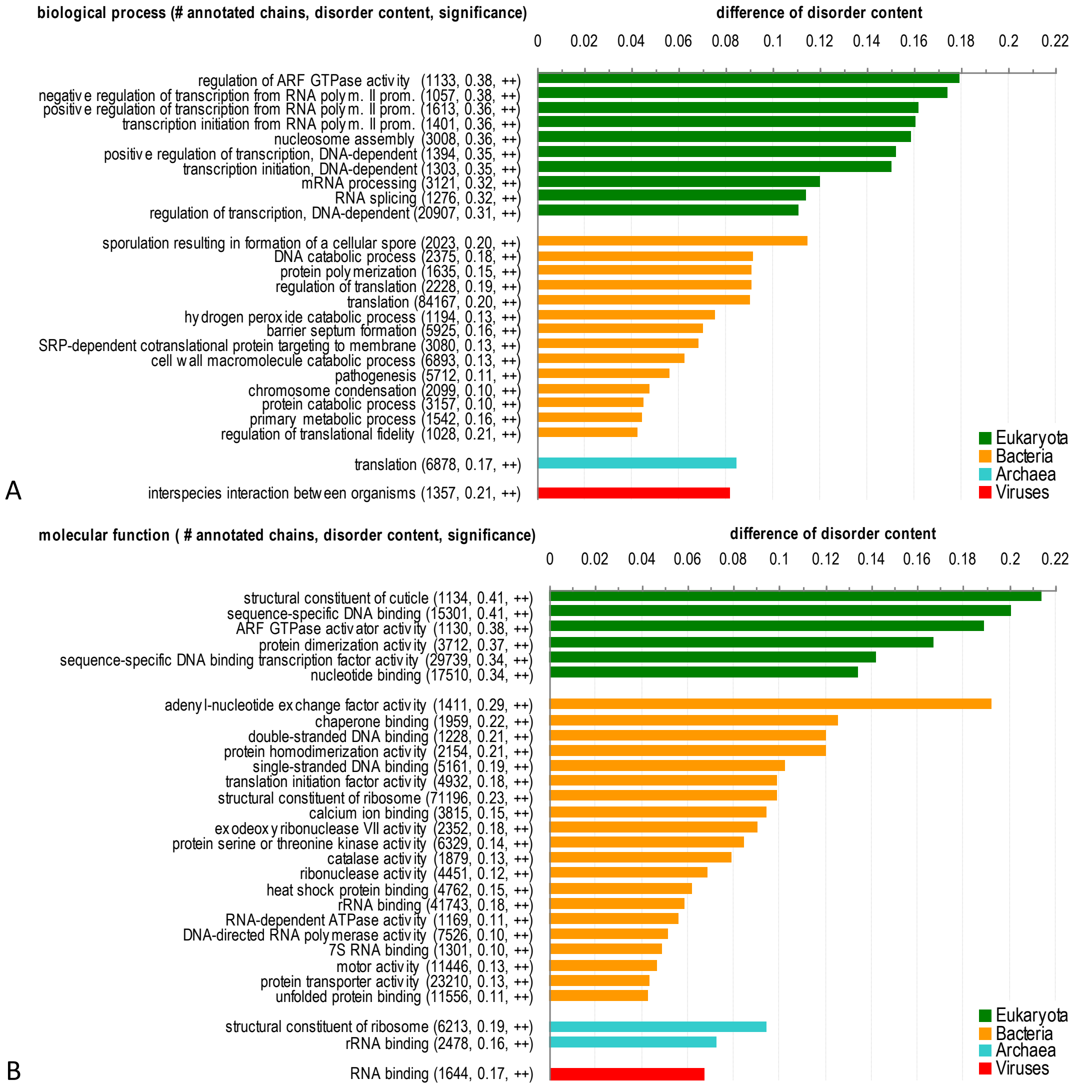

Figure 3.3. Functions enriched in disorder.
Biological processes (**A**) and molecular functions (**B**) which are significantly enriched in disorder across eukaryotic, bacterial, archaea, and viral species. The y-axis gives all significant functions including the number of corresponding proteins, the average disorder content, and significance of the enrichment. The x-axis shows the difference in average disorder content between proteins with a given functions and the baseline disorder content in a given kingdom. The functions are sorted, within each kingdom, by the values of the difference.

Figure 3.3B shows that intrinsic disorder is important for several molecular functions, such as DNA and nucleotide binding, protein dimerization, and transcription in eukaryotes and DNA and RNA binding, protein dimerization, translation, etc. in bacteria. This is consistent with the corresponding biological processes that are enriched in disorder across the four kingdoms of life.

3.3.3 Cellular Localization of Intrinsic Disorder across Kingdoms of Life

We also investigated the correlations between the intrinsic disorder propensity and cellular components, based on GO annotations for the considered complete proteomes. Results are summarized **Figure 3.4**. Considering eukaryotic cellular components, nucleosome, spliceosome, and transcription factor complexes are substantially enriched in the disorder. Bacteria also contain a large number of components associated with disorder, such as ribosome, cell wall, and flagellum, to name a few. We also show a substantial number of components in eukaryotic cells that are enriched in disorder when compared with bacterial cells; see inset in **Figure 3.4**. In contrast, proteins in Archaea use disorder primarily only for translation (see **Figure 3.3A**), which is why archaean IDPs are commonly involved in RNA binding and are located in ribosome. In addition to using disorder for the RNA binding, viruses commonly utilize IDPs to implement interactions with other organisms (see **Figure 3.3**), and their IDPs are often located in the cytoplasm and nucleus.

By mapping the components enriched in disorder from **Figure 3.4** into their cellular compartments, we observe that disorder is preferentially localized across the three kingdoms of life in the ribosome; see **Figure 3.5**. The organelles/compartments colored in red include at least one component that is significantly enriched in disorder, with dark red denoting the fact that eukaryotic nucleus includes the largest absolute disorder content, which is consistent with the observation that “the yeast proteins containing disorder are often located in the cell nucleus” (Ward, et al., 2004). **Figure 3.5** shows that disorder is relatively abundant across majority of elements in bacterial cell and in several eukaryotic organelles/compartments including nucleus, mitochondrion, cytoskeleton, peroxisome, and cell membrane and junction. However, some other compartments, such as most of intra-cellular membranes, golgi apparatus, endoplasmic reticulum, endosome, lysosome, centrosome, chloroplast, and vacuole, include mostly structured proteins.

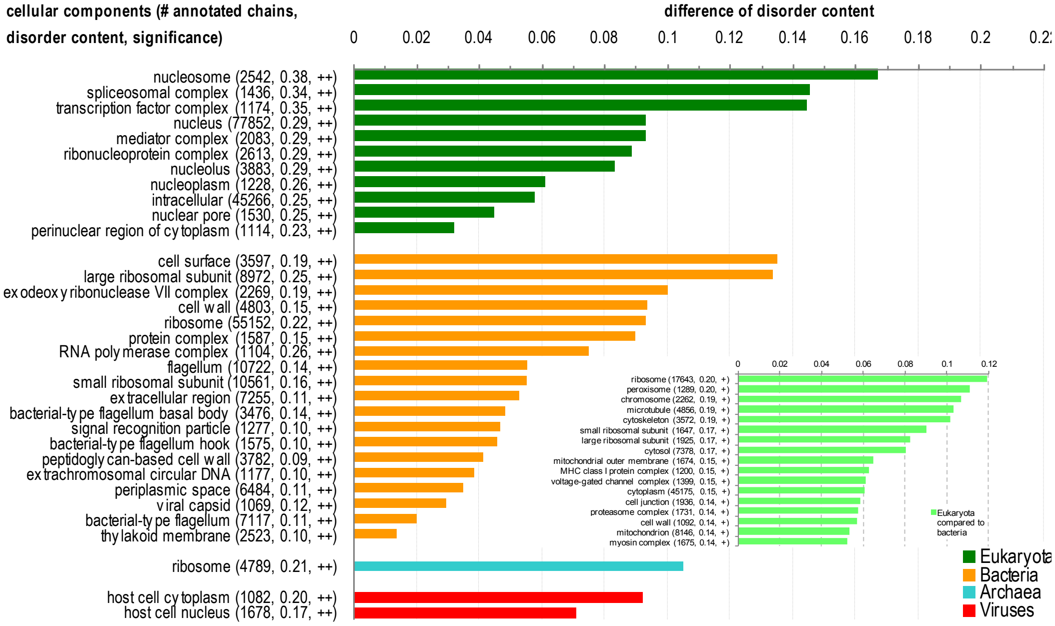

Figure 3.4. Cellular localization enriched in disorder.
Cellular components significantly enriched in disorder across eukaryotic, bacterial, archaea, and viral species. The y-axis gives all significant components including the number of corresponding proteins, the average disorder content, and significance of the enrichment. The x-axis shows the difference in average disorder content between proteins with a given functions and the baseline disorder content in a given kingdom. The components are sorted, within each kingdom, by the values of the difference.

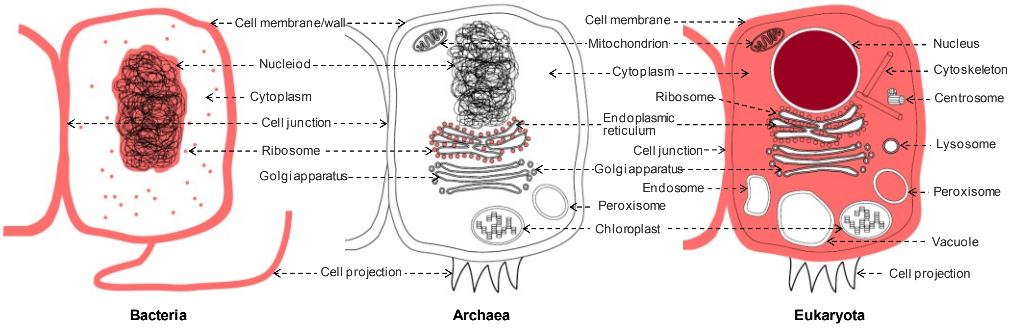

Figure 3.5. Mapping intrinsic disorder into eukaryotic, bacterial, and archaea cells.
The disorder-enriched cellular components from **Figure 3.4** were mapped into the corresponding organelles/compartments. The light red color identifies components that include at least one annotation that is enriched by at least 5% in bacteria or archaea; in eukaryota this color denotes components with annotations enriched by at least 5% compared to disorder in bacteria (based on inset in **Figure 3.4**). The dark red shows components that include at least one annotation enriched by at least 5% in eukaryota.

3.3.4 Evolution and Disorder

In order to put our observations into the evolutionary perspective, we built a phylogenetic tree by including 14 eukaryotic and 112 bacterial complete proteomes. This analysis is based on the evolutionary tree presented in (Ciccarelli, et al., 2006), which was reconstructed using a supermatrix of 31 concatenated, universally occurring genes with indisputable orthology in 191 species with completely annotated genomes in eukaryota, bacteria and archaea. In the original tree, the evolutionary pace of a given genome was proportional to the cumulative branch length from the tip to the root, with faster evolving genomes being characterized by longer branch lengths (Ciccarelli, et al., 2006). We performed detailed analysis of this evolutionary data to show the correlation between the disorder content in various phyla, the proteome size, and the evolutionary pace measured as branch length in the evolutionary tree. The correlation is quantified by Pearson correlation coefficient (PCC). Unfortunately, many phyla have too few species (less than 8 species) to have conclusive results, except for one eukaryotic phylum and three bacterial phyla that have sufficient data to be used in this analysis.

Figure 3.6 provides analysis of these evolutionary data combined with the analysis of sequence conservation for these four phyla. **Figure 3.6A** shows negative correlations between the disorder content and the evolutionary speed (measured as the branch length) within the selected four phyla. The PCC values are consistently negative and range between -0.3 and -0.86 suggesting that proteomes with more disorder evolve slower than proteomes with less disorder. Importantly, this trend holds true only within a given phylum; the correlation across proteomes from the four phyla is low and equals -0.11. **Figure 3.6B** shows that proteomes with higher disorder content are less conserved and that this trend is true even across phyla from bacteria and eukaryota, with the PCC value over all considered proteomes of -0.70. This agrees with prior observations that disordered regions are more likely to undergo non-conservative changes that lead to the lower sequence conservation compared to the structured regions (Brown, et al., 2010). Our analysis where we aggregate the conservation at the proteome level corroborates this finding. Furthermore, **Figure 3.6C** reveals that disordered regions have lower sequence conservation than ordered regions for majority of the considered proteomes, irrespective of the overall conservation in a given proteome. For instance, the lower overall conservation of the considered eukaryotes when compared with bacteria (**Figure 3.6B**) is combined with proportionally lower conservation of the corresponding disordered regions (**Figure 3.6C**). The relatively low conservation of the disordered regions does not explain the negative correlation between disorder content and evolutionary speed in specific phylum. A possible explanation for the latter trend is that disordered regions tend to be enriched in proteins with high connectivity (i.e., hubs) of protein-protein interactions networks (Bertolazzi, et al., 2013) and the connectivity of these networks was shown to be negatively correlated with their rate of evolution (Fraser, et al., 2002). Thus, enrichment in disorder could lead to higher connectivity (relative to a group of taxonomically related species in a given phylum), which, in turn, would lead to the reduced evolutionary speed. Another plausible explanation is related to the observation that smaller genomes evolved faster, which was explained by their limited ability to remove mutations by means of recombination or DNA repair (Ciccarelli, et al., 2006). **Figure 3.6D** shows a positive correlation between genome size (approximated by the number of proteins expressed by a given genome) and the disorder content within each of the four phyla. This figure taken together with Figure **Figure 3.6B** reported by Ciccarelli et al that represented negative correlation between the evolutionary speed and genome size (Ciccarelli, et al., 2006), suggests that lower evolutionary speed could be a consequence of the enlarged proteome size which is associated with enrichment in disorder. Based on our empirical results, we hypothesize that there is a correlation between the speed of evolution and the degree of disorderedness, where larger proteomes in the same phyla contain more disorder and evolve slower.

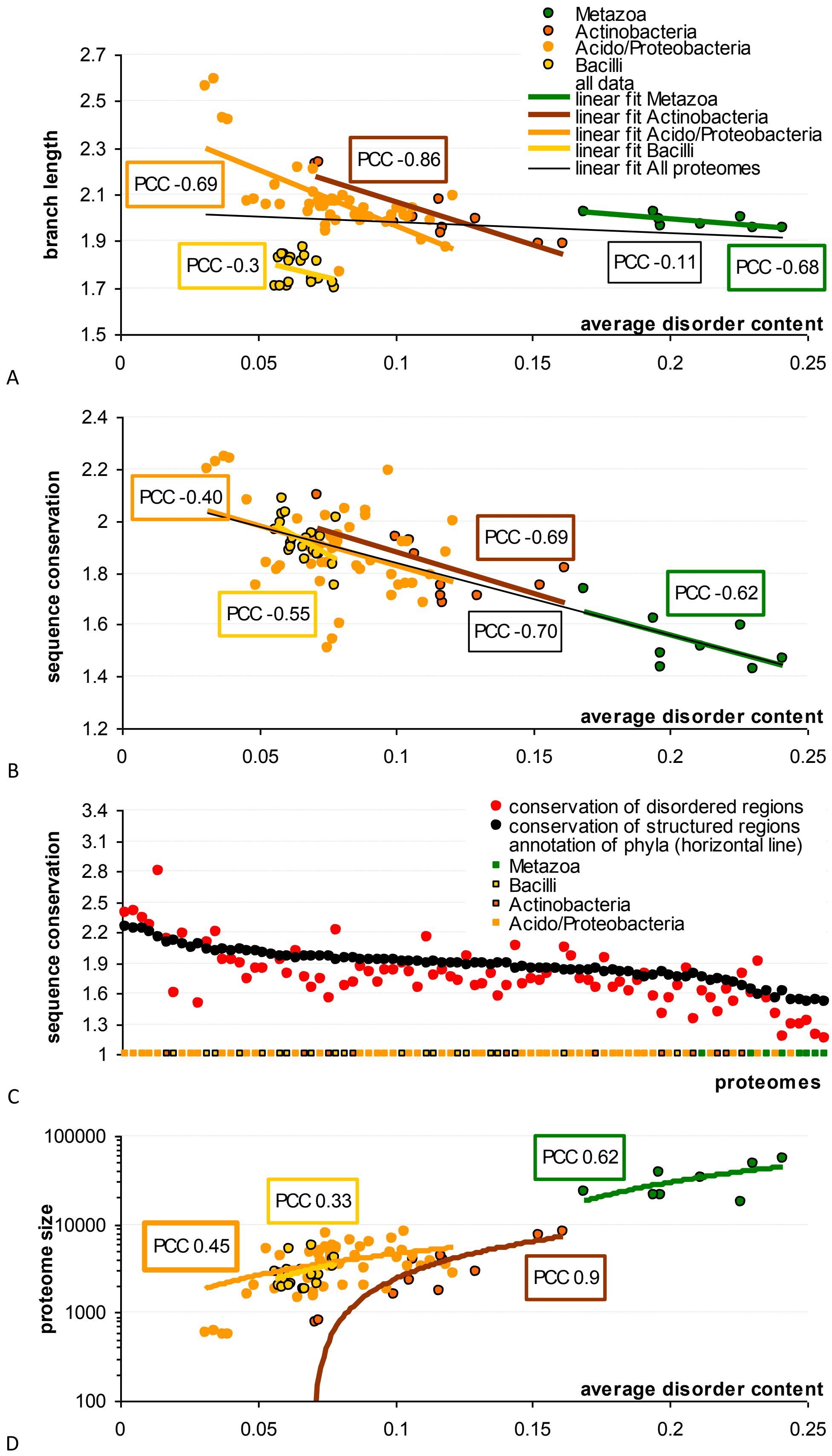
Metazoa
Actinobacteria
Acido/Proteobacteria
Bacilli
all data
linear fit Metazoa
linear fit Actinobacteria
linear fit Acido/Proteobacteria
linear fit Bacilli
linear fit All proteomes
branch length
2.7
2.5
2.3
2.1
1.9
1.7
1.5
PCC -0.69
PCC -0.86
PCC -0.3
PCC -0.11
PCC -0.68
average disorder content
0
0.05
0.1
0.15
0.2
0.25
A
sequence conservation
2.4
2.2
2
1.8
1.6
1.4
1.2
PCC -0.40
PCC -0.69
PCC -0.55
PCC -0.70
PCC -0.62
average disorder content
0
0.05
0.1
0.15
0.2
0.25
B
conservation of disordered regions
conservation of structured regions
annotation of phyla (horizontal line)
Metazoa
Bacilli
Actinobacteria
Acido/Proteobacteria
sequence conservation
3.4
3.1
2.8
2.5
2.2
1.9
1.6
1.3
1
proteomes
C
proteome size
100000
10000
1000
100
PCC 0.62
PCC 0.33
PCC 0.45
PCC 0.9
average disorder content
0
0.05
0.1
0.15
0.2
0.25
D

Figure 3.6. Relation between disorder content and evolutionary characteristics including evolutionary speed, sequence conservation and proteome size for the bacterial and eukaryotic species.
Relation of the disordered content with the pace of evolution quantified using branch length in an evolutionary tree (**A**), and with the sequence conservation (**B**), respectively. Panel **C** compares sequence conservation of disordered (red markers) and structured (black markers) regions across the species grouped by phyla that are denoted using the horizontal line at the bottom; species are sorted by the conservation of their structured regions. **D**. Relation between disorder content and proteome size. Solid lines in panels **A**, **B**, and **D** show linear fits together with the corresponding value of the Pearson correlation coefficient (PCC); y-axis in panel **D** is in logarithmic scale.

3.4 Conclusions

In agreement with a number of earlier studies, we show that IDPs/IDRs should not be considered as rare and obscure exceptions. Instead, these proteins are very common in all kingdoms of life, including viruses, and clearly possess specific set of molecular functions. Our analysis revealed that the eukaryotic species have a unique disorder profile compared to the corresponding profiles of archaean, bacterial and viral species. Here, eukaryotic proteomes were overall substantially more (about 20%) disordered, contained more disorder in longer/larger proteins, and were characterized by the larger fraction of proteins with larger amounts of disorder. Eukaryotes and viruses have larger number of longer fully disordered proteins and longer disordered regions, compared to bacteria and archaea; particularly, viruses have relatively large number of long (over 300 AAs) fully disordered chains.

Abundance of intrinsic disorder in eukaryotes and some of the viruses can be connected to the requirement of more profound signaling and regulation of these species. Analysis of the length-dependence of the average disorder content produced rather unexpected outcomes. One can expect that short proteins would contain less disorder than long proteins, and therefore the disorder content would increase with the protein length. However, dependence of the average disorder content on the protein length obtained in our study possessed an intriguing shape; see **Figure 3.2**B. For example, in eukaryotes, short proteins were predicted to have significant amount of disorder. The amount of predicted disorder decreased as protein length increased and reached minimum at ~ 15% for proteins with the length of 300-500 residues. Then, the amount of intrinsic disorder started to increase, reached a plateau at the level of 25% for proteins with length of ~1,000-2,000 residues, and then again started to decrease for longer proteins. Since the number of very long proteins is relatively small, that part of the plot corresponding to proteins longer than 5,000 residues is possibly noisy. Importantly, some long proteins contained very significant amount of predicted disorder, up to 90-95%. Similarly, short proteins from other kingdoms of life were typically more disordered than longer proteins. The fact that short proteins contained highest amount of predicted disorder and the fact that long disordered proteins in eukaryotes seem to have some "optimal" length (1,500-2,000 residues) with relatively high disorder content (25%) may potentially have some functional explanations.

Functional correlation study showed that disorder is enriched in many key processes including transcription, translation, nucleosome assembly/chromosome condensation, RNA splicing, protein polymerization and dimerization, catabolic and metabolic processes, and pathogenesis in bacteria. Moreover, disordered proteins are preferentially located in certain cellular compartments including nucleosome, spliceosome, transcription factor complexes, ribosome, and cell wall and flagellum in bacteria. Archaean proteins use disorder for translation, whereas viruses use disorder for RNA binding and for interactions with other organisms. We also provided a convenient mapping of disorder into archaea, bacterial and eukaryotic cells. Finally, we expanded the prior observation that linked proteome/genome size with the evolutionary speed by inclusion of the degree of disorderedness. We observe that among closely related species from the same eukaryotic or bacterial phyla, species with smaller proteomes that evolved faster have less disorder.

Chapter 4

Characterization of Intrinsic Disorder in Ribosomal and Cell Death Proteins

4.1 Introduction and Motivation

4.1.1 Introduction and Motivation for Study on Ribosomal Proteins

Our analysis on the complete proteomes shows that intrinsic disorder is significantly enriched in protein-RNA and protein-DNA interactions (see **Figure 3.3**), and is preferentially located in ribosome (see **Figure 3.4** and **Figure 3.5**). This is consistent with prior observation that intrinsic disorder is very common in RNA- and DNA-binding proteins (Dunker, et al., 2002; Dunker, et al., 2001; Tompa, 2002; Uversky, et al., 2000). In fact, many RNA-binding proteins possess a multitude of intrinsic disorder-dependent functions, such as acting as specific RNA chaperones (Tompa, et al., 2004), being involved in RNA metabolism and alternative splicing (Haynes, et al., 2006), and regulating viral gene expression and replication (Shojania, et al., 2011; Xue, et al., 2012), forming ribonucleoprotein core (Chang, et al., 2006). Intrinsic disorder was also shown to be enriched in the cell's nucleus (Peng, et al., 2012; Ward, et al., 2004), and to be prevalent in transcription factors (Bhalla, et al., 2006; Liu, et al., 2006; Minezaki, et al., 2006).

Ribosomal proteins represent an interesting and important category of RNA-binding IDPs due to their unique functional and structural properties. In addition to be a crucial part of a ribosome, many ribosomal proteins are involved in translational regulation via binding to operator sites located on their own messenger RNA (Zengel, et al., 1994). Based on the analysis of the crystal structures of the ribosome subunits it was discovered that almost half of the ribosomal proteins have globular domains with long extensions that penetrate deeply into the ribosome particle's core (Ban, et al., 2000; Ben-Shem, et al., 2011; Harms, et al., 2001; Schuwirth, et al., 2005; Selmer, et al., 2006; Timsit, et al., 2009; Wimberly, et al., 2000; Yusupov, et al., 2001). These extensions were found to be disordered in solution still playing a key role in the ribosomal assembly (Brodersen, et al., 2002; Garrett, 1983; Klein, et al., 2004; Timsit, et al., 2009). In fact, the hypothesis is that the long basic extensions of ribosomal proteins (e.g., L3, L4, L13, L20, L22 and L24) can penetrate deeply into the ribosome subunit cores, undergo disorder-order transition individually or co-fold with their ribosomal RNA (rRNA), therefore facilitating the proper rRNA folding (Timsit, et al., 2009). The same article also suggested that different extensions may play different roles in the assembly of the ribosome subunits *in vivo* and might have some other functions (Timsit, et al., 2009).

Although the fact that many ribosomal proteins are either completely disordered or contain long disordered regions is known for some time (e.g., ribosomal proteins were included in the early bioinformatics studies dedicated to the sequence peculiarities (Uversky, et al., 2000) and functional repertoire of IDPs (Dunker, et al., 2002)), the abundance and the functional roles of intrinsic disorder in these proteins never were the subject of focused large-scale bioinformatics analysis. Our study fills this gap by reporting the results of a bioinformatics analysis of 3,411 ribosomal proteins from 32 species. Our results demonstrate that intrinsic disorder is very common among all the analyzed ribosomal proteins, that it has unique characteristics which differentiate it from the disorder in other RNA- and DNA-binding proteins, and that it plays important roles in the various functions of these important RNA-binding proteins.

4.1.2 Introduction and Motivation for Study on Cell Death Proteins

In a multicellular organism, exposure of a cell to a set of environmental factors may start specific intracellular programs that trigger a chain of biochemical events that could lead to the characteristic changes in cellular morphology and ultimately to the cell death. These cell-killing intracellular events constitute programmed cell death (PCD) phenomenon, which includes at least three different mechanisms: apoptosis, autophagy and necroptosis (programmed necrosis) (Bialik, et al., 2010; Galluzzi, et al., 2011; Ouyang, et al., 2012).

These three PCD modules are integrated into a common PCD network, where pathways of these functional modules are interconnected and where many death regulatory proteins are common to more than one module (Bialik, et al., 2010). This tight control of the various PCD processes and strong connectivity of the involved proteins suggest that PCD-related proteins possibly possess specific (and potentially common) structural characteristics. These characteristics would allow them to be uniquely and effectively modulated via multiple specific interactions with various partners and to control the regulation and execution of different PCD modules.

Our analysis in Chapter 3 suggests that intrinsic disorder possesses a wide spectrum of biological functions that are typically related to regulation, signaling, and control pathways (Dunker, et al., 2005; Iakoucheva, et al., 2002; Uversky, et al., 2005). In fact, IDPs could be considered as "control freaks", and they commonly act as important regulators of protein-protein interaction networks. For example, previous studies show that intrinsic disorder is intimately associated with "hubness" of proteins; i.e., the ability to be involved in multiple interactions with unrelated with each other partners via one-to-many and many-to-one binding mechanisms (Dosztanyi, et al., 2006; Dunker, et al., 2005; Ekman, et al., 2006; Haynes, et al., 2006; Patil, et al., 2006; Singh, et al., 2006). This commonness of the intrinsic disorder in proteins involved in the control and regulation related functions suggests that the PCD could belong to the crucial biological processes that are controlled and regulated by IDPs/IDRs. To test this hypothesis, we applied a broad spectrum of computational techniques to analyze abundance and functional roles of intrinsic disorder in proteins related to the different types of PCD, focusing on proteins involved in apoptosis, autophagy, and necroptosis.

4.2 Materials and Methods

4.2.1 Benchmark Datasets

Datasets of ribosomal and RNA-/DNA-binding proteins. We collected 3438 proteins from the Ribosomal Protein Gene Database (RPG) (Nakao, et al., 2004) on Nov 7th,

2011. This set includes proteins from 24 eukaryotic, 4 archaean and 4 bacterial species. We excluded 27 small peptides with less than 30 amino acids because they could not be predicted by MFDp (Mizianty, et al., 2010), which was used to predict disorder. For convenience, the final dataset is named as RPG_3411. We also collected a representative subset of RNA- and DNA-binding proteins from release 2012_07 of UniProt (Consortium, 2012) for the same set of species as in the RPG_3411 dataset. Next, for each species we selected at random a subset of RNA- and DNA-binding proteins to match the number of ribosomal chains. This allowed us to represent a wide spectrum of the nucleic acids binding chains, while keeping the dataset sizes at a level that allows completing computational analysis. The combined set of RNA- and DNA-binding chains includes 3084 proteins. This number is slightly lower than the size of RPG_3411 since some proteins interact/bind with both RNA and DNA and a couple of species (*Fusarium Graminearum* and *Rhizopus Oryzae*) had fewer DNA/RNA-binding proteins annotated in UniProt than the corresponding number of ribosomal chains in the RPG_3411. These three datasets were utilized to investigate disordered profile in the RNA-/DNA- binding and ribosomal proteins, and to analyze the functions of IDRs in the ribosomal proteins.

Datasets of proteins involved in the programmed cell death. First, 1138 and 137 curated human proteins associated with apoptosis and autophagy, respectively, were collected from UniProt (Consortium, 2012) on November 14, 2012. These proteins were selected using "reviewed: yes apoptosis human" and "reviewed: yes autophagy human" keywords and were grouped into human_apoptosis and human_autophagy sets, respectively. Since similar search for the human proteins associated with necroptosis gave only five hits, 35 human necroptosis-related proteins were manually picked based on the analysis of literature data (Bialik, et al., 2010; Declercq, et al., 2009; Galluzzi, et al., 2011; Vandenabeele, et al., 2010). These proteins were also collected from the UniProt database and assembled into human_necroptosis set. In addition to the larger scale analysis of human PCD-related proteins, we performed a more focused analysis of fewer PCD-related proteins from several proteomes. These proteins were collected from Deathbase, a specialized resource dedicated to describe proteins involved in programmed cell death, which includes high-quality manually curated annotations. We extracted 3,458 proteins from Deathbase (Diez, et al., 2010) in November 9th, 2011.

This set includes proteins from 5 manually curated species: human, mouse, zebrafish, fly and worm, and 23 reference species that were annotated based on the similarity to the manually curated proteins. The proteins from the curated species include annotations of the corresponding PCD processes. Specifically, they comprise 154, 11, 25, and 26 proteins that are annotated to participate in apoptosis, necroptosis, immune response, and other PCD processes, respectively. These datasets were utilized to analyze the abundance and the functional roles of intrinsic disorder in all PCD-related proteins, and in subsets of proteins that participate in each PCD process.

4.2.2 Amino Acid Composition Analysis

Amino acid compositional analysis was carried out using Composition Profiler (Vacic, et al., 2007) (http://www.cprofiler.org) using the PDB Select 25 (Berman, et al., 2000) and the DisProt (Sickmeier, et al., 2007) datasets as reference for ordered and disordered proteins, respectively. Enrichment or depletion in each amino acid type was expressed as $(C_x\text{-}C_{order})/C_{order}$, i.e., the normalized excess of a given residue's content in a query dataset (C_x) relative to the corresponding value in the dataset of ordered proteins (C_{order}).

4.2.3 Computational Prediction and Characterization of Disorder

Intrinsic disorder was predicted with MFDp method (Mizianty, et al., 2010), which is a consensus-based predictor that was recently shown to provide strong and competitive predictive quality (Monastyrskyy, et al., 2011; Peng, et al., 2012). The putative disorder was next used to calculate the disorder content (fraction of disordered residues), the number of intrinsically disordered regions (IDRs), and the number of long IDRs that consists of at least 30 consecutive disordered amino acids; such long regions were found to be implicated in protein-protein recognition (Tompa, et al., 2009). We only counted the IDRs with at least four consecutive disordered residues. This is consistent with other prior reports (Monastyrskyy, et al., 2011; Noivirt-Brik, et al., 2009).

4.2.4 Search for Potential Globular Domains in Ribosomal Proteins

Potential globular domains in ribosomal proteins were identified using the GlobPlot server (http://globplot.embl.de/), which is a popular predictor based on a running sum of the propensity for amino acids to be in an ordered or disordered state (Linding, et al., 2003). GlobPlot is a computationally efficient web service that allows the user to plot the tendency within the query protein for order/globularity and disorder (Linding, et al., 2003) and was recently evaluated to provide competitive predictive performance (Li, et al., 2012). We defined a predicted globular domain in the ribosomal proteins as disordered if it contains at least one IDR, and as significantly disordered if at least half of its residues are disordered.

4.2.5 Search for Potential Functional Sites

Function of IDRs was predicted based on local pairwise alignment against functionally annotated IDRs collected from release 5.9 of DisProt (Sickmeier, et al., 2007). Specifically, we aligned each of the 7548 and the 10952 IDRs from the RPG_3411 and the Deathbase, respectively, to a set of 775 IDRs collected from DisProt that have functional annotations. We performed alignment using the Smith-Waterman algorithm (Smith, et al., 1981) based on the EMBOSS implementation with default parameters (gap_open=10, gap_extend=0.5, and blosum62 matrix). We defined sequence similarity as the number of identical residues in the local alignment divided by the length of the local alignment or the length of the shorter of the two being aligned regions, whichever is larger. We transferred the annotation if the similarity is greater than 0.8. The value of the threshold was chosen to assume high similarity even in cases of alignment of a short region, i.e., for the shortest regions of five residues at least four amino acids have to be matched. We note that the same IDR could be annotated with multiple functions using this protocol. Consequently, we successfully annotated 911 and 2108 IDRs from RPG_3411 and Deathbase, respectively, with 26 functions that are listed **Table S1** in Appendix B. These annotations were used to investigate difference in the functional roles between short and long IDRs extracted from the RPG_3411 and Deathbase, respectively. We also discussed whether the IDRs involved in different PCD processes are associated with different functions, by considering all 128 annotated disordered regions from the manually curated protein species.

In addition, we performed detailed analysis for a specific functional type of IDRs: molecular recognition features (MoRFs; see section 2.7), where MoRFs were annotated by the MoRFpred method (Disfani, et al., 2012). Following Mohan et al. (Mohan, et al., 2006), we divided MoRF regions into four subtypes including α-MoRFs, β-MoRFs, γ-MoRFs and complex-MoRFs, based on the secondary structure predicted with PSI-PRED (Jones, 1999).

4.2.6 Calculation of Sequence Conservation

We also report sequence conservation for the ordered residues, the disordered residues and the residues in long IDRs (with at least 30 consecutive disordered amino acids). The conservation was quantified with relative entropy (Wang, et al., 2006) that was computed in the same way as described in section 3.2.5.

4.3 Results for Study on Ribosomal Proteins

4.3.1 Peculiarities of the Amino Acid Compositions in Ribosomal Proteins

Amino acid compositions of the full-length ribosomal proteins. Analysis of the amino acid composition biases can provide interesting information about the underlying protein. For example, IDPs were shown to be significantly depleted in order-promoting amino acids, C, W, I, Y, F, L, H, V, and N, and substantially enriched in disorder-promoting residues, A, G, R, T, S, K, Q, E, and P (Dunker, et al., 2001; Radivojac, et al., 2007; Romero, et al., 2001; Vacic, et al., 2007; Williams, et al., 2001). We use a computational tool, Composition Profiler (Vacic, et al., 2007), to investigate the compositional biases in ribosomal proteins. This approach is based on the calculation of a normalized composition of a given protein or protein dataset in the $(C_s – C_{order})/C_{order}$ form, where C_s is a content of a given residue in a query (ribosomal) protein or dataset, and C_{order} is the corresponding value for a representative set of ordered proteins collected from the PDB Select25 dataset (Berman, et al., 2000). **Figure 4.1A** shows that, in comparison with the representative ordered proteins, ribosomal proteins from three kingdoms of life (i.e., eukaryota, bacteria and archaea) are depleted in the major order-promoting amino acids, C, W, F, Y, L, H and N, and are enriched in some disorder-promoting residues, particularly R, K, G (except for the eukaryotic ribosomal proteins), A (except for the archaean ribosomal proteins), and E (except for the eukaryotic ribosomal proteins). The enrichment in positively charged R and K residues is determined by the functional need of the ribosomal proteins to interact with the negatively charged rRNA. This high lysine-arginine content also defines the unusually high pI values reported for the majority of the ribosomal proteins (average pI ~10.1). Overall, the pronounced depletion in the bulky hydrophobic and aromatic amino acids and enrichment in the polar and charge residues is possibly related to the low propensity of ribosomal proteins for autonomous (or partner-independent) folding. On the other hand, there are several interesting compositional biases in the ribosomal proteins that differentiate them from the typical IDPs. These biases include enrichment in the order-promoting amino acids I and V, and a noticeable depletion in the content of disorder-promoting residues T, D, Q and S.

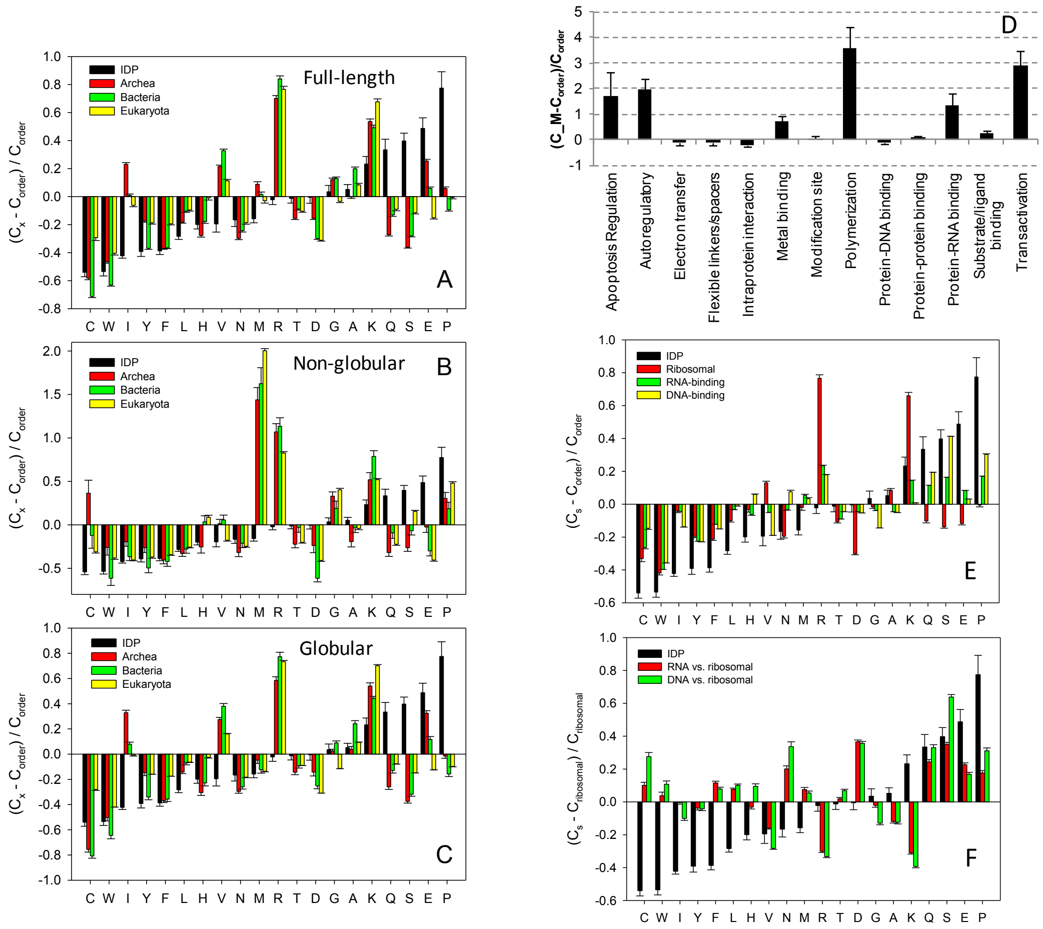

Figure 4.1. Overview of peculiarities of amino acid composition in ribosomal proteins. Fractional difference in the amino acid composition between the different members of the family of ribosomal proteins from bacteria (green bars), archaea (red bars), and eukaryota (yellow bars) and a set of ordered proteins calculated for each amino acid residue (compositional profiles). The fractional differences were evaluated for the full-length ribosomal proteins (**A**) and for non-globular (**B**) and globular domains (**C**) predicted by GlobPlot server. Compositional profile of IDPs from the DisProt database is shown for comparison (black bars). Positive bars correspond to residues found more abundantly in ribosomal proteins, whereas negative bars show residues, in which ribosomal proteins are depleted. Amino acid types were ranked according to their increasing disorder-promoting potential (Radivojac, et al., 2007). Panel **D** shows enrichment of amino acid M in the functions assumed by IDRs that are considered in this work. To assure statistically sound results, we include 13 functions that have at least 20 annotated IDRs. The fractional difference was calculated for M for the 13 functions that are sorted alphabetically on the *x*-axis. Positive bars correspond to functions (disordered regions annotated with a given function) found with high counts of M while negative bars show functions where M is depleted. Panels **E** and **F** compare the amino acid compositions of the ribosomal, RNA- and DNA-binding proteins. In **E**, the fractional difference was calculated as $(C_x\text{-}C_{order})/C_{order}$, where C_x is the content of a given amino acid in a query set, and C_{order} is the corresponding content in the dataset of fully ordered proteins. In **F**, the compositions of the RNA- and DNA-binding proteins are compared with the general amino acid composition of the ribosomal proteins. Here, the normalized compositions of the RNA- and DNA-binding proteins are evaluated in the $(C_s - C_{ribosomal})/_{Cribosomal}$ form, with C_s being a content of a given residue in a dataset of the RNA- or DNA-binding proteins, and $C_{ribosomal}$ being the corresponding value for ribosomal proteins. In both plots, composition profiles of typical IDPs from the DisProt database are shown for comparison (black bars).

Compositions of globular domains and non-globular regions. We analyzed peculiarities of the amino acid compositions of globular and non-globular domains predicted using the GlobPlot server. **Figure 4.1B** shows that all non-globular domains of the ribosomal proteins clearly possess compositions typical for the IDPs/IDRs, being enriched in major disorder-promoting residues and depleted in order-promoting residues. On the other hand, **Figure 4.1C** illustrates that predicted globular domains possess amino acid biases consistent with the idea that they might contain significant amount of disorder. In fact, in many respects, the composition profile of globular domain resembles profiles calculated for the full-length ribosomal proteins. These domains are depleted in all order-promoting residues except for the isoleucine and are enriched in some disorder-promoting residues (e.g., G, A, K, and E). **Figure 4.1D** provides further analysis of amino acid methionine that we found to be substantially enriched in the non-globular domains (**Figure 4.1B**) while being moderately depleted in the globular domains (**Figure 4.1C**). We study the enrichment/depletion of this residue type over IDRs with functional annotations (as explained in section 4.3.3); we consider 13 functions that are possessed by at least 20 annotated sequences. We show that the enrichment in methionine is associated with several functions carried out by the intrinsic disorder, such as polymerization, transactivation, autoregulation, regulation of apoptosis, and interactions with RNA and metals.

4.3.2 Overall Characterization of the Intrinsic Disorder in Ribosomal, RNA-, and DNA-Binding Proteins

Ribosomal proteins are important parts of ribonucleoprotein machine, the ribosome, where they specifically interact with rRNA and other ribosomal proteins. Therefore, we compared various characteristics of the ribosomal protein group (RPG) with those of general RNA- and DNA-binding proteins. To this end, representative sample sets of RNA- and DNA-binding proteins were assembled as described in section 4.2.1 and these three datasets were used in the subsequent studies.

The panels **E** and **F** in **Figure 4.1** represent the comparison of amino acid compositions of the ribosomal proteins, RNA- and DNA-binding proteins. In **Figure 4.1E**, the normalized amino acid compositions of these three classes of nucleic acid-binding proteins are shown, where the normalized compositions were calculated as described in section 4.2.2. This figure shows that all nucleic acid binding proteins are characterized by comparable depletion in the order-promoting residues. As far as the disorder-promoting residues are concerned, while the RNA- and DNA-binding proteins generally follow the trend typical for the IDPs, being moderately enriched in the major disorder-promoting residues, the ribosomal proteins are quite different. Two major features strike the eye – a substantial enrichment of the ribosomal proteins in R and K compensated by a noticeable depletion in D, Q, S, and E residues. To get better understanding of the amino acid composition biases of the RNA- and DNA-binding proteins relative the ribosomal proteins, we evaluated their normalized compositions in the $(C_s – C_{ribosomal})/C_{ribosomal}$ form, with C_s being a content of a given residue in a dataset of the RNA- or DNA-binding proteins), and $C_{ribosomal}$ being the corresponding value for ribosomal proteins. Results of this analysis are shown in **Figure 4.1F**, which reemphasizes the relative enrichment of the RNA- and DNA-binding proteins in N, D, Q, S, E and P and their depletion in V, R, A, and K. Generally, data shown in **Figure 4.1E** and **F** suggest that the RNA- and DNA-binding proteins are closer to each other than to the ribosomal proteins.

The average disorder content (i.e., the fraction of disordered residues) in the ribosomal protein group (RPG) ranges between 36% and 37.4% across the three kingdoms of life, see **Figure 4.2**. This is substantially higher than the overall disorder content in various proteomes, which was estimated to be 18.9%, 5.7%, and 3.8% for eukaryota, bacteria, and archaea, respectively (Ward, et al., 2004). Our results indicate similar levels of disorder in the three kingdoms of life and across the 32 considered species, with the lowest content over 28%. **Figure 4.2** also shows that between 2.5 and 23.2% of ribosomal proteins across the 32 species are fully disordered, with the largest average fraction (11.7%) of fully disordered chains being found in the bacterial species.

This characteristic of the ribosomal proteins is different from that of the DNA- and RNA-binding proteins. In fact, disorder in the DNA- and RNA-binding proteins is unevenly distributed among the three kingdoms of life, with proteins from eukaryotes being substantially more disordered than corresponding proteins from archaea and bacteria. Interestingly, the overall disorder contents of eukaryotic ribosomal and RNA-binding proteins are rather similar (~37% and 41%, respectively) whereas eukaryotic DNA-binding proteins possess more disorder (~60%). However, in archaea and bacteria, situation is reversed and ribosomal proteins are more disordered than the RNA- and DNA-binding proteins (see **Figure 4.2**). Fully disordered eukaryotic ribosomal proteins are somewhat more abundant than the fully disordered RNA-binding proteins and noticeably less abundant than the fully disordered DNA-binding proteins. In archaea and bacteria, fully disordered chains are essentially more abundant among the ribosomal proteins than among the corresponding RNA- and DNA-binding proteins.

On average, ribosomal proteins have between 1.4 (in eukaryota) and 1.5 (in bacteria and archaea) IDRs per 100 residues (we normalize by unit of length to allow direct comparison to longer DNA- and RNA-binding chains), including 0.3 to 0.4 long IDRs (>30 amino acids) per 100 residues. Therefore, according to all these parameters, ribosomal proteins are substantially more disordered than the RNA- or DNA-binding proteins. This is an interesting observation since ribosomal proteins are typically significantly shorter (at least 1.7 times shorter) than the RNA- and DNA-binding proteins.

We further analyze the distribution of IDRs across chains with different length; see **Figure 4.3**. While in archaea the number of long IDRs in the ribosomal proteins increases linearly with the length of the protein chain, we observe increased number of IDRs for short chains in eukaryota and bacteria (see **Figure 4.3**A). Furthermore, short (<100 AAs) fully disordered ribosomal proteins are relatively common in eukaryota and bacteria, where about 1/3 of short chains are fully disordered. In contrast, archaea has some longer fully disordered chains. This is due to the inclusion of *Halobacterium Salinarum* (HAL) that has the highest disorder content (59.3%), which stems from the fact that it has the largest fraction (23.2%) of fully disordered proteins among all considered species; see **Figure 4.2**. Overall, our analysis implies that small ribosomal proteins in eukaryota and bacteria are enriched in disorder, when compared with the ribosomal proteins in archaea. These characteristics are different from the trends observed for the DNA- and RNA-binding proteins, which typically possess less disorder-related features than the ribosomal proteins, except for the eukaryotic DNA-binding proteins, and whose disorder attributes decrease with the protein length (see **Figure 4.3B** and **C**).

We also characterized the intrinsic disorder in domains of ribosomal proteins, by applying GlobPlot and MFDp tools to the set of 3,438 ribosomal proteins. The results revealed that 412 proteins (12.0%) were predicted without globular domains, 502 proteins (14.6%) were predicted not to have disordered regions, whereas remaining proteins were predicted to be hybrid proteins that contained both globular and disordered domains. **Figure 4.4**A shows that in the three kingdoms of life, most ribosomal proteins with globular domains are single domain proteins (in ~60% proteins, >95% residues are included in a GlobPlot predicted domain). However, more detailed analysis of globular domains using the MFDp tool showed that many of them contained IDRs and some are predicted to be entirely disordered (see **Figure 4.4**B). **Figure 4.4**C shows that almost all globular domains contain at least one disordered region with more than three consecutive disordered residues, and ~20% of domains were significantly disordered, containing at least half disordered residues. These observations clearly show that intrinsic disorder is very common in ribosomal proteins from the three kingdoms of life.

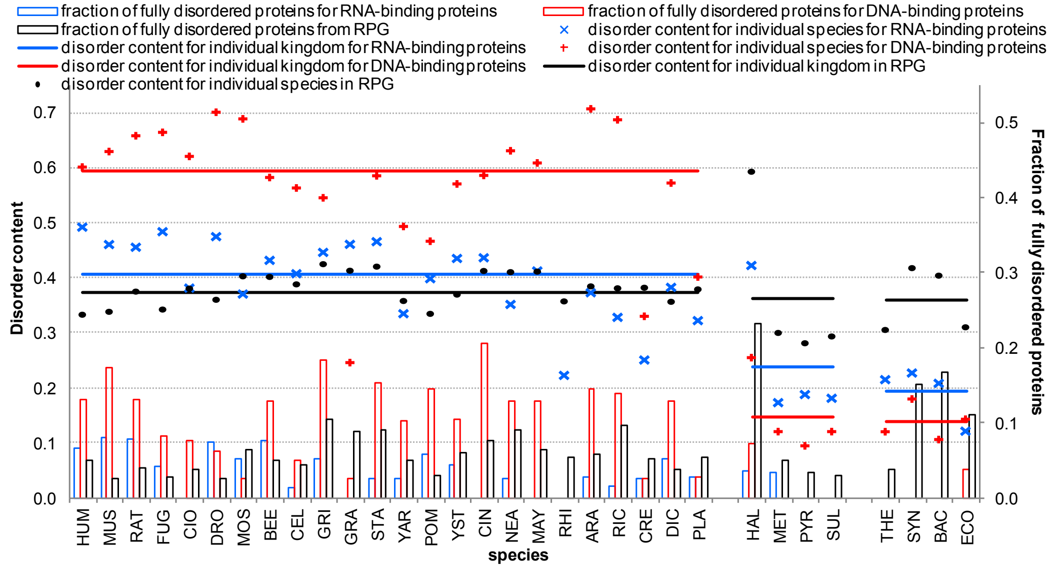

Figure 4.2. Amount of intrinsic disorder in the ribosomal, DNA-, and RNA-binding proteins. Disorder content (crosses and lines) and fraction of fully disordered proteins (bars) in different species and kingdoms of life for the ribosomal, DNA-, and RNA-binding proteins. The species, which are shown on the *x*-axis, are grouped into the kingdom of eukaryota, archaea and bacteria.

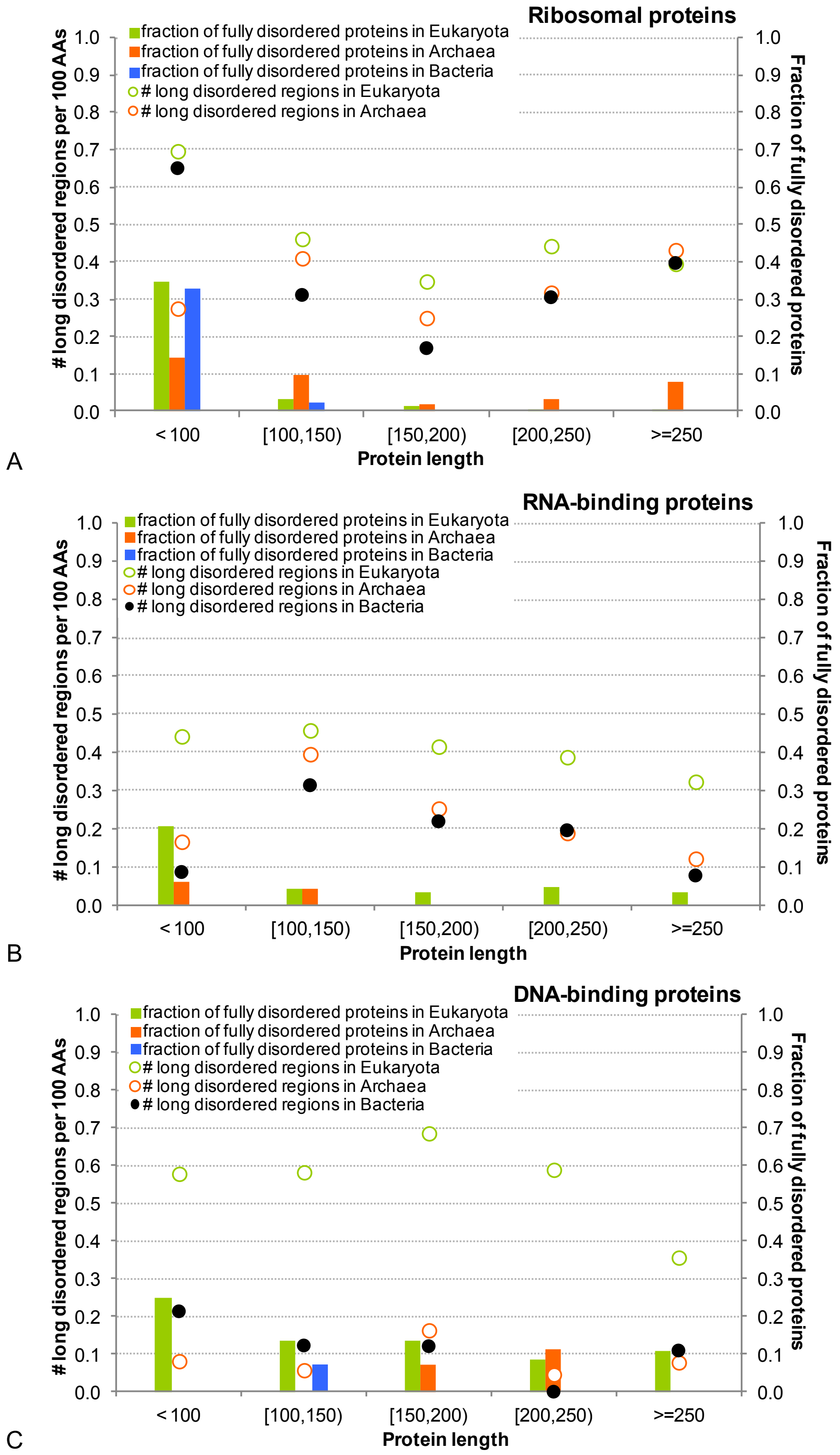

Figure 4.3. Distribution of long disordered regions across chains with difference length. The number of long IDRs (≥30 AAs) per protein (*y*-axis on the left; hollow points) and the fraction of fully disordered protein (*y*-axis on the right; solid bars) against protein length (*x*-axis) across the three kingdom of life in ribosomal (**A**), RNA- (**B**) and DNA-binding proteins (**C**).

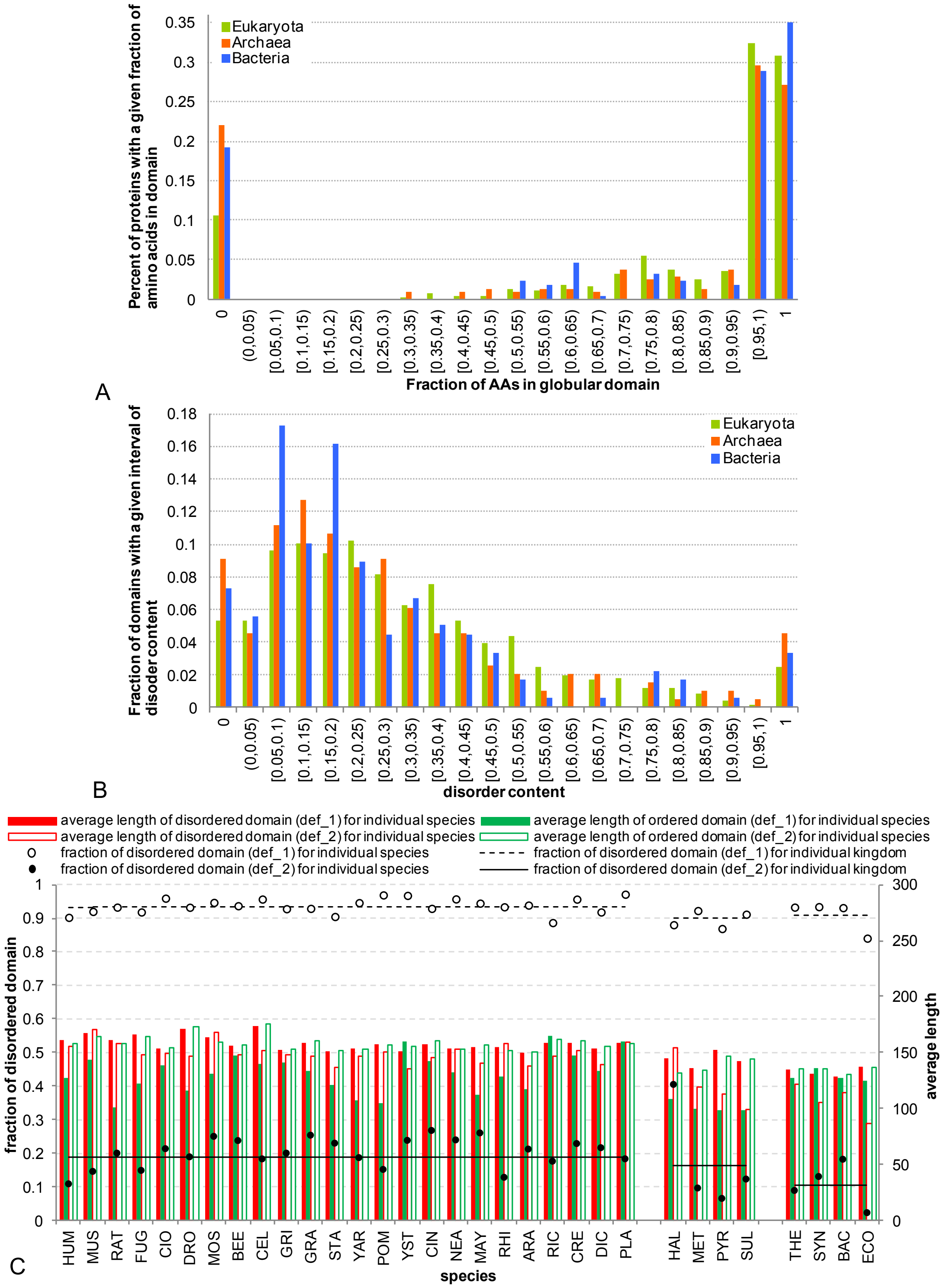

Figure 4.4. Characterization of the domains in ribosomal proteins.

A. The distribution of fraction of amino acids in domain per protein. **B**. The distribution of disorder content per domain. **C**. The fraction of disordered domains (hollow and solid circles, respectively; *y*-axis on the left) and the average length of disordered (solid and hollow bars in red) and ordered domains (solid and hollow bars in green; y-axis on the right). Domains were assumed to be disordered when they contain at least one disordered region with at least four consecutive disordered residues (def_1) or when at least half of their residues are disordered (def_2).

4.3.3 Functional Analysis of Disordered Regions in Ribosomal Proteins

Although ribosomal proteins are relatively short, with the average length of about 100-150 residues, we observed that the sizes of IDRs follow bimodal distribution with a relatively large number of short regions (between 4 and 15 amino acids) and with a second peak for longer fragments (between 25 and 100 amino acids) across the three kingdoms of life. The bimodal length distribution of IDRs motivated us to analyze the function for two classes of the IDRs: short regions with less than 30 amino acids, and long with at least 30 amino acids. We considered 26 functions of IDRs in the ribosomal proteins, which are annotated based on sequence alignment into the functionally characterized disordered regions from the DisProt database (as explained in section 4.2.5), that are summarized in **Table S1** in Appendix B. We exclude functions with less than 20 annotations for both short and long disordered regions.

Figure 4.5 compares the annotations of the 13 remaining predicted (using alignment) functions between the short and long disordered regions of ribosomal proteins. The results reveal that disorder in ribosomal proteins plays several important roles, from facilitating the protein-protein, protein-DNA, protein-RNA, and protein-other-ligand interactions, to involvement in metal binding, post-translational modifications, and implementation of linkers and intra-protein interactions. Overall, both long and short disordered regions are similarly implicated in several functions including interactions with proteins, DNA, and ligands. The short regions are predominant in a larger number of functions, including RNA and metal binding, auto-regulatory functions, transactivation, polymerization, apoptosis, and are more prevalent in the post-translational modification sites. At the same time, the long disordered regions more often serve as linkers and play a strong role in the intra-protein interactions. Our analysis provides useful clues that can be used to narrow down potential functions of IDPs and IDRs, especially knowing the size of the corresponding regions, in ribosomal chains that currently lack functional annotations.

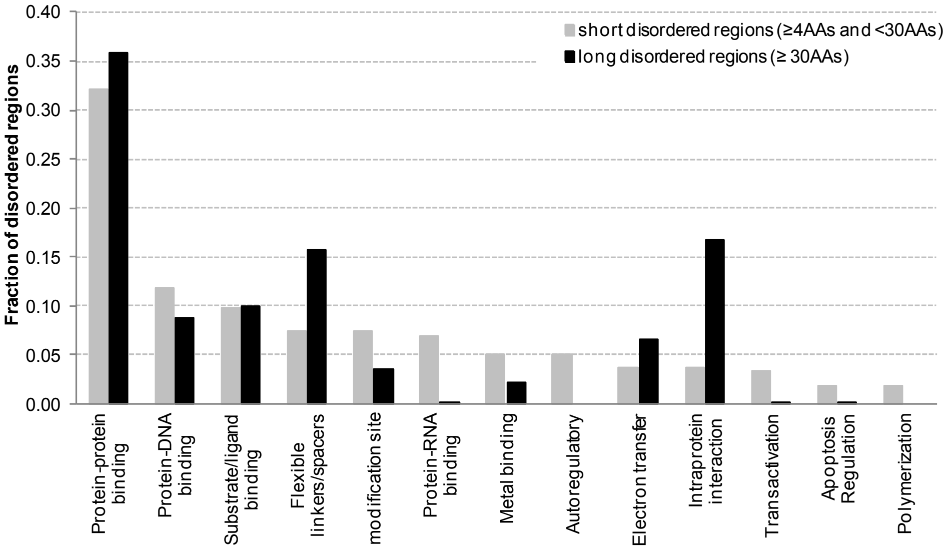

Figure 4.5. Functions of short and long IDRs extracted from ribosomal proteins, respectively. Fraction of short (4 to 30 amino acids) and long (over 30 amino acids) disordered regions for a given function; *x*-axis represents the 13 considered functions sorted by the decreasing number of short regions.

4.3.4 MoRF Regions in Ribosomal, RNA-, and DNA-Binding Proteins

The most prevalent function of disorder in the ribosomal proteins is facilitation of the protein-protein interactions. **Figure 4.5** shows that over 30% of short IDRs in the ribosomal proteins are implicated in these binding events. This motivates our analysis of MoRFs regions (Mohan, et al., 2006; Oldfield, et al., 2005; Uversky, et al., 2010; Vacic, et al., 2007). **Figure 4.6A** demonstrates that there are on average about 0.85 MoRFs per 100 residues (we normalize by unit of length to allow direct comparison to longer DNA- and RNA-binding chains) in the eukaryotic ribosomal proteins, including a large fraction of α-MoRF and γ-MoRF and relatively lower numbers of complex- and β-MoRFs. The complex-MoRFs, γ-MoRFs, and α-MoRFs are similarly abundant in the ribosomal chains from the three kingdoms of life, while bacterial and archaean ribosomal proteins are enriched in β-MoRFs. Both, RNA- (**Figure 4.6B**) and DNA-binding proteins (**Figure 4.6C**) have fewer MoRF regions per 100 residues, and are characterized by rather different distributions of the overall abundance of MoRFs (which vary more widely between species) and their split into α-, β-, γ-, and complex-MoRFs between eukaryotic, archaean and bacterial proteins, particularly for DNA-binding chains that are depleted in β-MoRFs. This suggests that MoRF regions in the ribosomal chains may be involved in different types of protein-proteins interactions across different kingdoms of life.

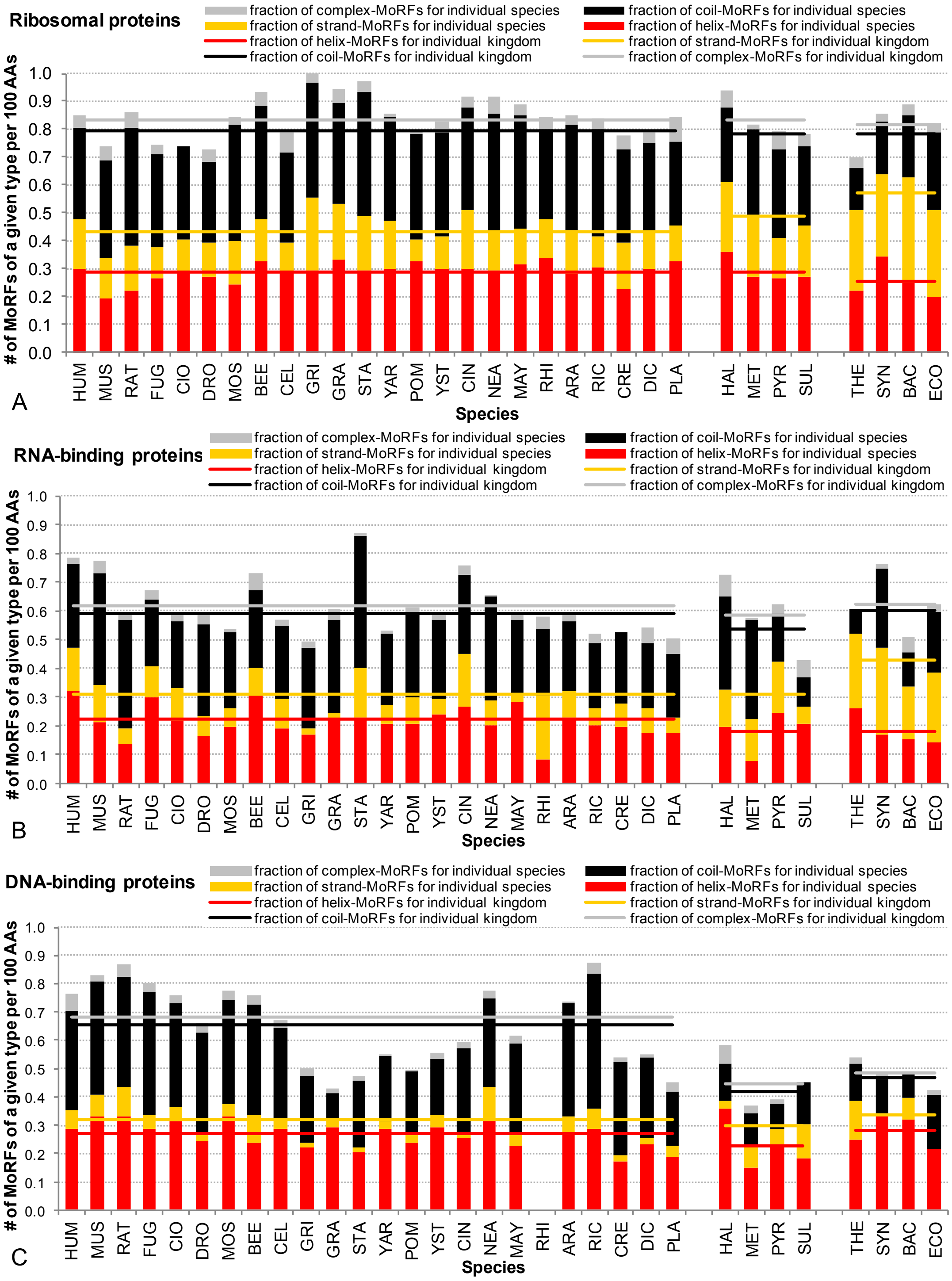

Figure 4.6. MoRFs on ribosomal, RNA- and DNA-binding proteins, respectively.
Number of MoRFs per protein, shown using stacked bars, across different species and kingdoms. The bars are subdivided using colors that correspond to different MoRF types. The solid lines show a cumulative (over MoRF types located below the line) average number of a given MoRF type for each of the three domains. The species, which are shown on the *x*-axis, are grouped into eukaryota, archaea and bacteria kingdoms. Panels **A**, **B** and **C** correspond to ribosomal, RNA- and DNA-binding proteins, respectively.

4.3.5 Evolutionary Conservation of Disorder in Ribosomal Proteins

Next, we investigate evolutionary conservation of intrinsic disorder in the ribosomal proteins. The conservation was quantified using the relative entropy computed from the Weighted Observed Percentages (WOP) profiles generated by PSI-BLAST. Higher values of the relative entropy indicate a higher degree of conservation. **Figure 4.7** shows that ribosomal, RNA-, and DNA-binding proteins in bacteria are characterized by higher levels of conservation when compared with the archaea and eukaryota. This can be also observed in **Figure 4.8** where we compare conservation between disordered and ordered residues. Besides the overall trend that shows higher conservation in bacteria, our results show that disordered residues are more conserved when compared with the structured parts of the ribosomal proteins (see **Figure 4.8A**). This is true for all species in eukaryota and archaea, while in bacteria the disordered and ordered residues have similarly high conservation. Moreover, we show that residues located in long IDRs of ribosomal proteins are more conserved than the overall population of both disordered and ordered amino acids across all three kingdoms of life. In eukaryotic RNA-binding proteins, the situation is reversed and ordered regions are more conserved (**Figure 4.8B**), whereas eukaryotic DNA-binding proteins are characterized by the higher conservation of long disordered and ordered regions (see **Figure 4.8C**). This suggests that disorder plays important role in the three kingdoms of life from the evolutionary perspective, particularly in the ribosomal proteins where it is characterized by higher conservation levels.

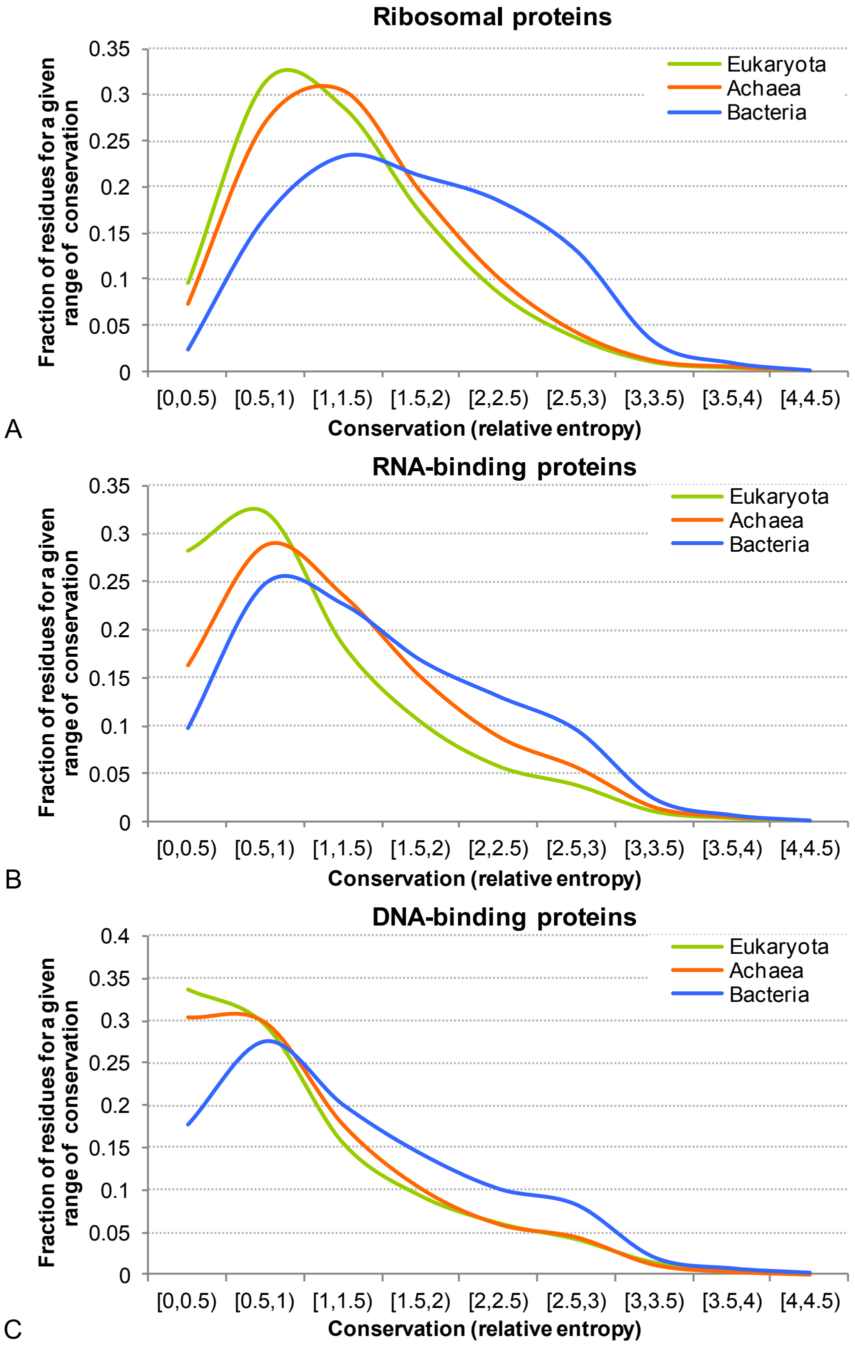

Figure 4.7. Distribution of the evolutionary conservation for ribosomal, RNA- and DNA-binding proteins.

Distribution of the average relative entropy. which quantifies evolutionary conservation, for the proteins from eukaryota, archaea and bacteria. Panel **A**, **B**, and **C** correspond to ribosomal, RNA- and DNA-binding proteins, respectively.

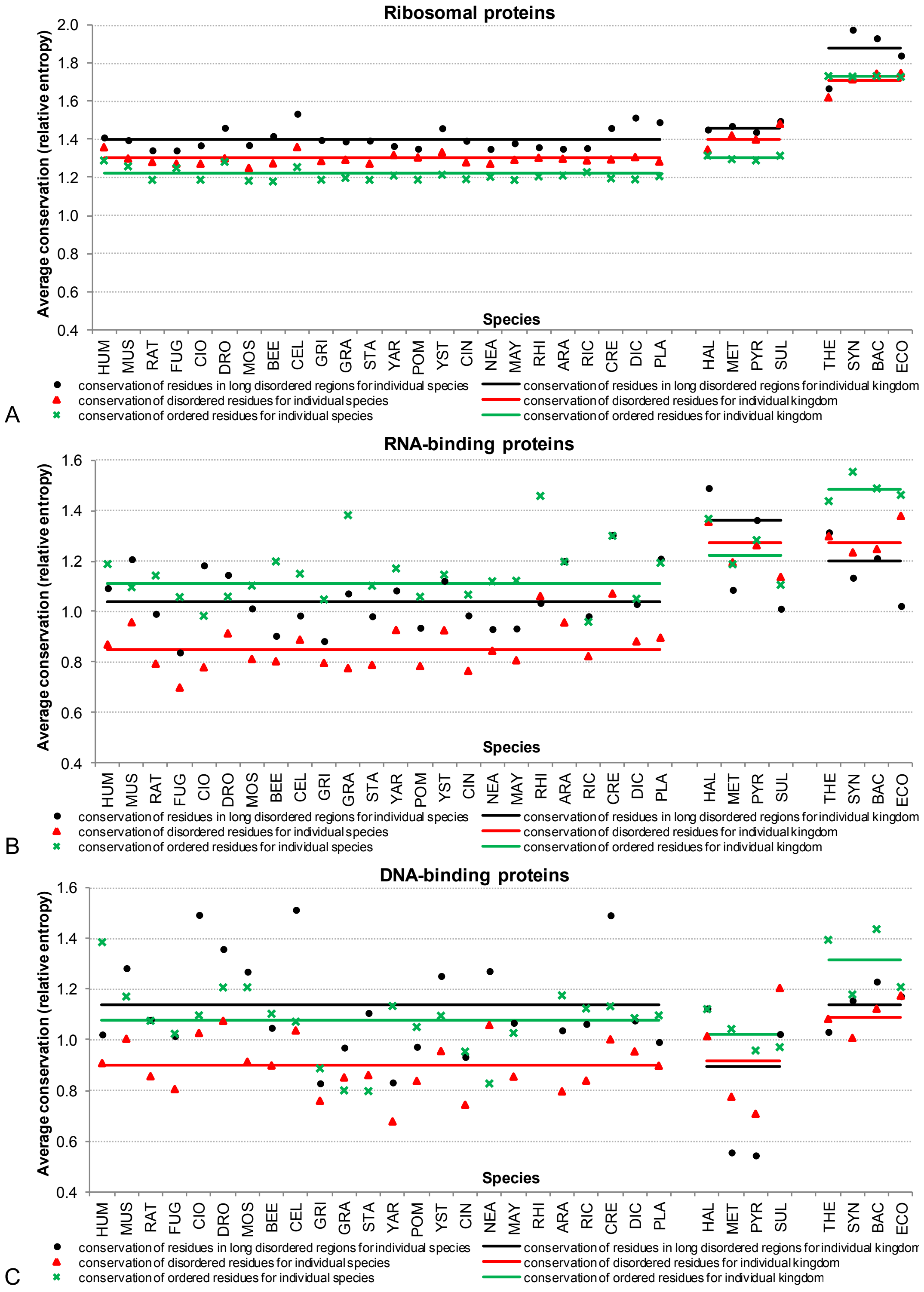

Figure 4.8. The average evolutionary conservation for ribosomal, RNA- and DNA-binding proteins.

The average relative entropy, which quantifies evolutionary conservation, across different species and kingdomss. Blue points/lines, green triangles/lines, and orange crosses/lines denote the average relative entropy of disordered residues in long disordered segments, all disordered residues, and ordered residues, respectively. The species, which are shown on the x-axis, are grouped into eukaryota, archaea and bacteria kingdoms. Panels **A**, **B** and **C** correspond to ribosomal, RNA- and DNA-binding proteins, respectively.

4.4 Discussion and Conclusions for Study on Ribosomal Proteins

4.4.1 Commonness and Peculiarities of Intrinsic Disorder in Ribosomal Proteins

We show that intrinsic disorder is abundant within the ribosomal proteins from the three kingdoms of life. This conclusion is in line with the results of the analysis of crystal structure of the eukaryotic ribosome from the yeast *S. cerevisiae* that revealed that many ribosomal proteins contain regions of intrinsic disorder, which are seen as regions with missing electron density (Ben-Shem, et al., 2011). **Figure 4.9** represents the results of the computational disassembly of protein components of this eukaryotic ribosome (PDBID: 3U5C and 3U5E), and shows that the complex structure of this important nucleoprotein relies on the intrinsic disorder of ribosomal proteins. In fact, many ribosomal proteins contain IDRs that are at least 8 residues long, with some IDRs can be as long as 94 residues. The illustrative examples of such proteins are listed in Appendix A.

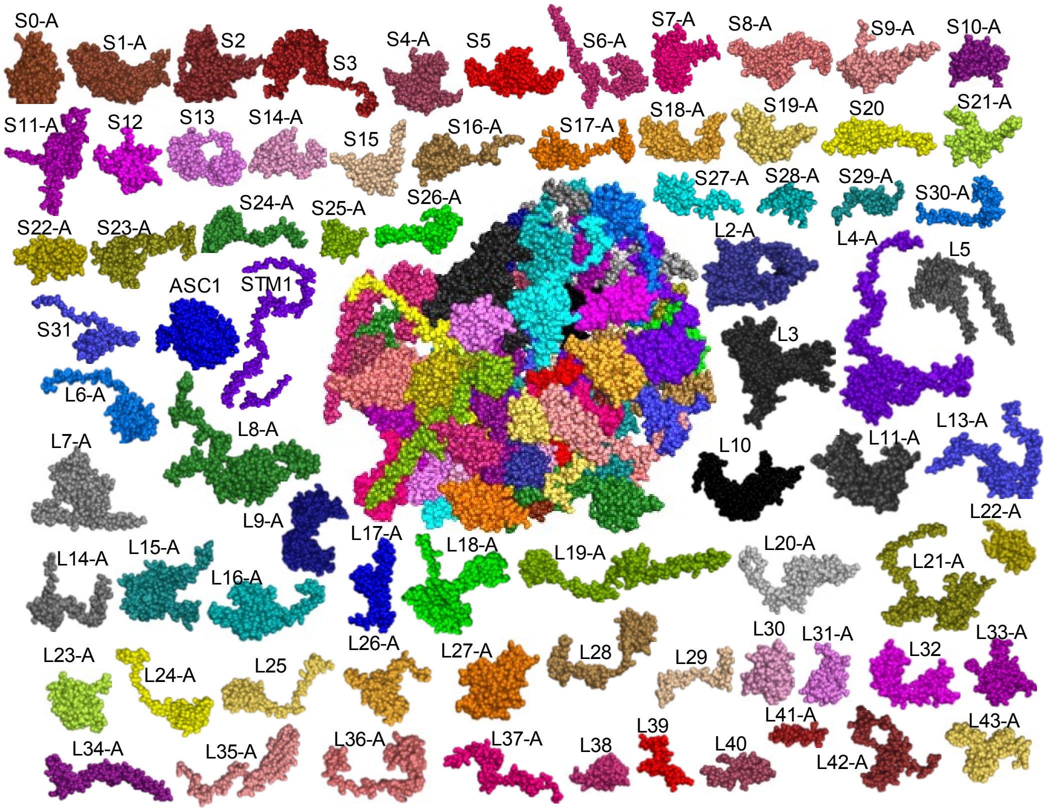

Figure 4.9. Computational disassembly of the eukaryotic ribosome A from the yeast *S. Cerevisiae* (PDBID: 3U5C and 3U5E).

Visual analysis of the crystal structures of individual ribosomal proteins revealed that many of them possess very unusual morphologies inconsistent with globular structures suggesting that these structures are likely to be formed as a result of binding-induced folding. This hypothesis is supported by the computational conclusion that the vast majority of eukaryotic ribosomal proteins is found above the order-disorder boundary (Gunasekaran, et al., 2004). These observations indicate that many eukaryotic ribosomal proteins are intrinsically disordered in their unbound states. To understand how general this statement is, we analyzed a large dataset of ribosomal proteins from the three kingdoms of life. Application of various computational tools unequivocally showed that disorder is very common in all the ribosomal proteins and that many potential globular domains still possess noticeable levels of disorder (from **Figure 4.1** to **Figure 4.4**). Since disorder is reliably predicted using computational tools developed based on the disorder-related data from large databases (e.g., PDB), one can conclude that IDRs of ribosomal proteins are generally similar in their properties to IDRs of many other proteins observed in several large databanks.

The ribosome is a ribonucleoprotein machine whose proteins are involved in interactions with both proteins and RNA. To understand how ribosomal proteins differ from other nucleic acid binding proteins, we compared some of their disorder-related features with disorder characteristics of large randomly selected sets of generic RNA- and DNA-binding proteins. Data shown in **Figure 4.1** to **Figure 4.4** and **Figure 4.6** to **Figure 4.8** suggest that disorder in the ribosomal proteins and its functional roles are different from those aspects of disorder in the DNA- and RNA-binding proteins. It is likely that some of these differences are related to the functional uniqueness of ribosomal proteins, many of which are involved in multiple simultaneous binding events, being involved in the interaction with RNA and other ribosomal proteins. Next, we discuss some of the reasons for the abundance of disorder in the ribosomal proteins.

4.4.2 Functional Viewpoint of Intrinsic Disorder in the Ribosomal Proteins

Protein-rRNA and protein-protein interactions on the ribosome. Being components of a large ribonucleoprotein complex, ribosomal proteins are obviously involved in interaction with both RNA and other proteins. Their ability to bind to RNA is determined by high positive charge. In general, ribosomal proteins are very basic (average pI ~10.1), suggesting that a general function of these proteins may be to counteract the negative charges of the phosphate residues in the rRNA backbone. In agreement with this hypothesis, many ribosomal proteins were shown to serve as RNA chaperones and therefore play crucial roles during the ribosome assembly (Semrad, et al., 2004; Wilson, et al., 2005). The only exceptions from this rule are S1 and S6 in the small subunit and the L7/L12 proteins in the large subunit which do not have intensive contacts with RNA, being predominantly engaged in the protein-protein interactions (**Figure 4.9** shows these proteins). Here, L7/L12 interact directly with L10 to form the pentameric L10 × $(L7/L12)_4$ or heptameric L10 × $(L7/L12)_6$ complex, S6 makes extensive contact with S18, and S1 interacts with S21, S11 and S18 (Wilson, et al., 2005).

Many ribosomal proteins possess complex structure and are often characterized by containing a globular domain, which is generally located on the surface of the ribosome, and a long extended region that penetrates into the ribosome's interior. In fact, all S-proteins (except S4 and S15) and about 50% of the L-proteins possess such extensions which have distinctive amino acid compositions, containing multiple residue G to allow flexibility and tight packing, and are rich in basic amino acids to interact with rRNA (Wilson, et al., 2005). The content of the basic amino acids R/K in the extensions of the large subunit ribosomal proteins (27%) noticeably exceeds that of the globular parts (19%). As a result these extensions that constitute only ~20% of the protein mass of the large subunit are responsible for burying of ~50% of total RNA surface area (Wilson, et al., 2005). It was pointed out that some ribosomal proteins, being studied in isolation, contain globular regions, whereas their extended tails are typically not observed in the isolated structures (Wilson, et al., 2005), suggesting that these regions undergo disorder-to-order transitions induced by interaction with the rRNA. Among the extreme examples are the long extensions of L2 and L3 that reach towards the peptidyl-transferase center. S12 has an extremely long extension that starts from the globular domain located adjacent to the decoding center on the intersubunit side of the small subunit and reaches all the way to the back or solvent side of the 30*S*, where it interacts with S8 and S17. Thus S12 provides an illustrative example of the "penetrator" binding mode, where significant part of an IDP penetrates deep inside the structure of its binding partner (Uversky, 2011). Also, the short 61 amino acid ribosomal protein S14 is completely devoid of any globular domain (Wilson, et al., 2005). Therefore, IDRs of many ribosomal proteins are important foldable regions that serve to ensure the formation of a correctly folded rRNA state during the ribosome assembly process and also support the correct conformation of the rRNA in the final assembled complex (Wilson, et al., 2005).

Besides the intensive contacts with rRNA, several ribosomal proteins are involved in well-developed net of protein-protein interactions. For example, a tight heterodimeric complex is formed by S6 and S18 proteins on the outer edge of the platform of the small subunit, whereas at the back of the 30*S* head, S3, S10, and S14 form a tight complex, and in the large subunit there are previously mentioned pentameric L10 × $(L7/L12)_4$ or heptameric L10 × $(L7/L12)_6$ protein complexes (Wilson, et al., 2005). Formation of these tight protein-protein complexes may also involve disorder-to-order transition, at least in some parts of the interacting proteins.

Specific on-ribosome functions. It was recognized long ago that some ribosomal proteins are mostly essential for the assembly of the ribonucleoprotein particle and are dispensable for function after the ribosomal subunits are fully assembled (Nierhaus, 1991). This fact suggests that the major function of these “dispensable” proteins (e.g., S16, L15, L16, L20, and L24) in the assembled ribosome could be to improve the ribosome stability. Furthermore, there are several ribosomal proteins that are not essential for the translational function of the ribosome, the hypothesis based on the observations *E. coli* strains lacking S6, S9, S13, S17, S20, L1, L9, L11, L15, L19, L24, L27 to L30, and L33 are viable (Dabbs, 1978; Dabbs, 1986; Wilson, et al., 2005). Since the subject of the on-ribosome functions of the ribosomal proteins was covered in a recent in-depth review (Wilson, et al., 2005), we are listing some of these functions in the appendix A. The interested readers are encouraged to look for the original review, where the functional roles of many ribosomal proteins were considered in great detail (Wilson, et al., 2005).

All these functions are relying on multiple interactions with various partners, suggesting that ribosomal proteins can be considered as ribosomal hubs. Earlier, it was shown that binding promiscuity of hubs can be determined by the use of intrinsic disorder in one of the two ways, where one disordered region can bind to many different partners and many disordered region can bind to one partner (Dosztanyi, et al., 2006; Dunker, et al., 2005; Ekman, et al., 2006; Haynes, et al., 2006; Patil, et al., 2006; Singh, et al., 2007; Singh, et al., 2007).

Moonlighting or off-ribosome functions. The core ribosome functions; i.e., the precise interaction of mRNA codon with tRNA anticodon and the catalysis of peptide bond formation are carried out by rRNA molecules of the small and the large ribosomal subunits, respectively. Therefore, the major or core on-ribosome functions of ribosomal proteins are to assist in rRNA folding (i.e., to serve as RNA chaperones) and function, to assist in the ribosome assembly, and to be involved in related protein-protein, protein-rRNA, protein-mRNA, and protein-tRNA interactions. On the other hand, many ribosomal proteins were shown to be involved in some extra-ribosomal or auxiliary functions, thereby serving as an illustrative example of moonlighting proteins. In agreement with this hypothesis, numerous extra-ribosomal functions were assigned to ribosomal proteins (Lindstrom, 2009; Tompa, et al., 2005; Warner, et al., 2009; Weisberg, 2008; Wool, 1996). It was even stated recently that “moonlighting is particularly widespread among ribosomal proteins, many of which have extra-ribosomal employment” (Weisberg, 2008). Even the first systematic analysis of this subject (which was performed in 1996) revealed that ribosomal proteins might have up to 30 extra-ribosomal functions (Wool, 1996). Recently, it was emphasized that the numerous extra-ribosomal functions of ribosomal proteins reported in the literature so far can be grouped into two major categories, where ribosomal proteins (a) control balance among ribosomal components; or (b) control nucleolar stress, or aberrant ribosome synthesis, leading to cell cycle arrest or apoptosis (Warner, et al., 2009). Some of the extra-ribosomal functions of ribosomal proteins within the ribosome system were already described above (e.g., see notes for S1, L1, and L4) and are covered in great detail in a recent review (Warner, et al., 2009). In *E. coli*, these extra-ribosomal include the L4 mediated inhibition of translation of the S10 operon that encodes eleven different ribosomal proteins including L4 itself (Zengel, et al., 1994) and binding of L4 to RNAse E that modulates the RNAse E activity, leading to the stress-related changes in the mRNA composition (Singh, et al., 2009). Among other regulatory ribosomal proteins L4 occupies a unique position due to its ability to regulates both transcription and translation of its transcription unit (Freedman, et al., 1987; Zengel, et al., 1990; Zengel, et al., 1990). Furthermore, via a comprehensive analysis of deletion and point mutants, these two functions of L4 were assigned to different regions of this protein (Li, et al., 1996). In fact, although the C-terminal region of L4 (residues 171-201) was shown to be crucial for the L4-mediated autogenous control, it was not involved in the incorporation of this protein to the ribosome. On the other hand, the central region of L4 (residues 67-103) was involved in the ribosome assembly but did not play significant role in the regulatory L4 functions (Li, et al., 1996). Curiously, the last third of the regulatory C-terminal fragment of L4 is predicted to be highly disordered, whereas central region required for the ribosome assembly is expected to be mostly disordered throughout its entire length.

In eukaryotes, L30 inhibits splicing by binding to its own transcript (Eng, et al., 1991), S14 controls the splicing of the transcript of one of its genes (Fewell, et al., 1999), L2 controls the level of its mRNA through accelerated turnover (Presutti, et al., 1991), S13 binds to the first intron of its transcript to inhibit splicing (Malygin, et al., 2007; Parakhnevich, et al., 2007), and L12 controls its own synthesis by inhibiting the splicing of its own mRNA (Mitrovich, et al., 2000). In addition to these roles in the control of the balance among ribosomal components during the ribosome synthesis, the established off-ribosome functions of ribosomal proteins are related to the surveillance of the ribosome assembly, as well as numerous roles in development, apoptosis and cancer (Warner, et al., 2009). It is very likely that the ability of ribosomal proteins to act off the ribosome can be attributed to their intrinsically disordered nature. This hypothesis is in agreement with the recent analysis which showed that the structural malleability characteristic for the IDPs/IDRs can define the capability of some proteins to be involved in the moonlighting activities (Tompa, et al., 2005).

In conclusion, we presented the results of the comprehensive computational analyses of ribosomal proteins that shows that the vast majority of these important RNA-binding proteins are typical IDPs. We also show that intrinsic disorder is very important for various biological functions of ribosomal proteins, being commonly used in numerous interactions of any given ribosomal protein with its various binding partners of different nature, such as other ribosomal proteins, RNA, and proteins from the translational machinery. The intrinsically disordered nature of ribosomal proteins is highly conserved in different kingdom of life, indicating that the lack of rigid structure, the resulting ability of ribosomal proteins to interact with various binding partners and be involved in the wide spectrum of the moonlighting activities represent strong evolutionary advantage. Therefore, careful consideration and appreciation of intrinsic disorder are crucial for better understanding of structure and conformational behavior of ribosomal proteins, their promiscuity, molecular mechanisms of their numerous extra-ribosomal functions, and mechanisms underlying regulation and control of these very important proteins.

4.5 Results for Study on Programmed Cell Death Proteins

4.5.1 Characterization of the Intrinsic Disorder in Human Proteins Associated with the Programmed Cell Death

Analysis of the compositional biases in human PCD-related proteins. Protein structure is determined by its primary protein sequence, which means that the amino acid composition biases can provide interesting information about the nature of the underlying protein. In fact, IDPs/IDRs show biases in amino acid compositions (Dunker, et al., 2001; Radivojac, et al., 2007; Romero, et al., 2001; Vacic, et al., 2007; Williams, et al., 2001). Therefore, we investigated the amino acid composition of the PCD-related proteins to evaluate their intrinsically disordered nature, by applying Composition Profiler (Vacic, et al., 2007). Results of this analysis are shown in **Figure 4.10A**, which illustrates that, in comparison with typical ordered proteins, human proteins from all three PCD types are depleted in some major order-promoting residues (e.g., W, I, Y, F, V, and N, see **Figure 4.10A**) and are enriched in some major disorder-promoting residues (e.g., Q, S, E, and P). This suggests that these proteins might contain multiple signatures characteristic for the disordered proteins.

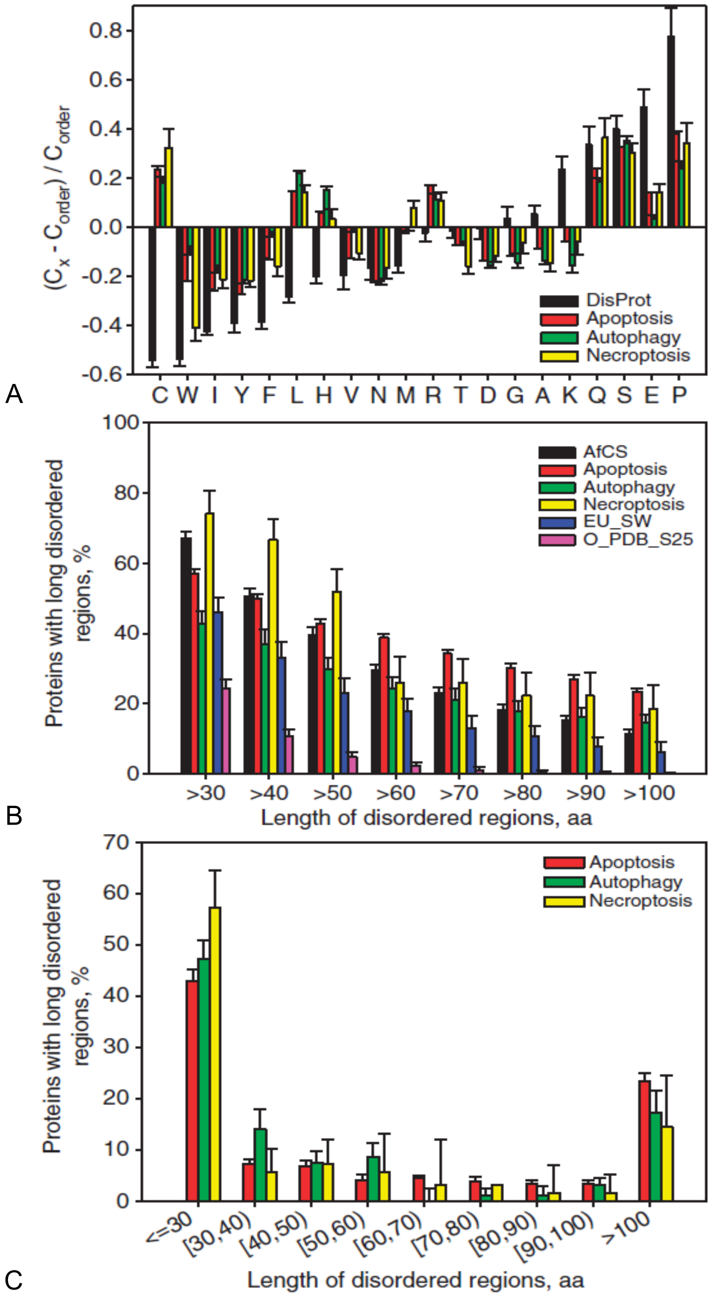

Figure 4.10. Peculiarities of intrinsic disorder distribution in human PCD proteins.
A. Fractional difference in the amino acid composition between the different PCD-related proteins (apoptosis, red bars; autophagy, green bars; and necroptosis, yellow bars) and a set of ordered proteins calculated for each residue (compositional profiles). The fractional difference was evaluated as (C_x-C_{order})/C_{order}, where C_x is the content of a given amino acid in a query set, and C_{order} is the corresponding content in the dataset of fully ordered proteins. Composition profile of IDPs from the DisProt database is shown for comparison (black bars). Positive bars correspond to residues found more abundantly in PCD-related proteins, whereas negative bars show residues, in which PCD-related proteins are depleted. Amino acid types are ranked according to their increasing disorder-promoting potential (Radivojac, et al., 2007). **B**. Abundance of predicted long IDRs in human PCD proteins in comparison with long IDRs in 2,329 proteins involved in cellular signaling (AfCS, black bars), 53,630 eukaryotic proteins from SWISS-PROT (EU_SW, blue bars), and 1,138 sequences corresponding to ordered parts of proteins from PDB Select 25 (O_PDB_S25, pink bars). AfCS collected by Alliance for Cellular Signaling. **C**. Distribution of the length of IDRs, which are extracted from human PCD-related proteins.

Abundance of long disordered regions in human PCD-related proteins. Previous study revealed that intrinsic disorder is abundant in signaling proteins. Specifically, 66% of signaling proteins were shown to contain predicted IDRs of 30 residues or longer (Iakoucheva, et al., 2002). Therefore, we estimated the amount of intrinsic disorder for human PCD-related proteins, based on the putative disorder annotated by the computational method MFDp (Mizianty, et al., 2010); see section 4.2.3. **Figure 4.10B** illustrates that intrinsic disorder is prevalent in the PCD-related proteins too, being comparable with the prevalence observed for signaling and eukaryotic proteins. In fact, the fraction of human PCD-related proteins with long regions of predicted disorder is 3- to 6-fold higher than that of the non-homologous ordered proteins from PDB (Iakoucheva, et al., 2002), being also a bit higher than the corresponding fraction in the eukaryotic proteins. **Figure 4.10C** further illustrates the peculiarities of the disorder content distribution in the three PCD-related datasets and shows that although about 50% of human PCD proteins contain IDRs shorter than 30 consecutive residues, sizable fractions of these datasets (in a range of 15-20%) correspond to proteins with very long IDRs (longer than 100 consecutive residues).

Overall, this analysis revealed that human PCD proteins possess relatively large amount of disorder, with the apoptosis- and necroptosis-related proteins being noticeably more disordered (on average) that the proteins involved in autophagy.

4.5.2 Overall Characteristics of the Intrinsic Disorder in Deathbase

We performed a comprehensive analysis of the manually curated and well-annotated PCD-related proteins from the Daethbase. The average disorder content (i.e., fraction of disordered residues in a protein chain) in these proteins ranged between 0.17 and 0.38 across the considered 28 species; see **Figure 4.11A**. The average content in these proteins across the 28 species was 0.27, which was substantially larger than the overall disorder content in eukaryotic species that was estimated to be 0.19 (Ward, et al., 2004). Side-by-side comparison of disorder content for individual species between the entire proteome and the cell death proteins showed consistent enrichment of the disorder in the latter protein sets. Specifically, in human, the overall disorder content was reported at 0.22 (Ward, et al., 2004) while in human cell death proteins the content was 0.32, in worm they were 0.16 vs. 0.27, and in fly 0.22 vs. 0.35.

The distribution of the disorder content is shown in **Figure 4.11B**. We observed that close to half of the cell death proteins contain up to 20% disordered residues. On the other hand, 17% of these proteins had at least half of their amino acids disordered and about 1% (39 proteins) is fully disordered.

Next, we investigated the enrichment of disorder across various cell death processes. **Figure 4.11C** reveals that proteins involved in necroptosis and apoptosis have larger amounts of disorder at about 0.41 and 0.35, respectively, compared to proteins involved in immune responses, for which the average disorder content equals to 0.2.

Figure 4.11D compares the average number of disordered regions per proteins. It demonstrates that proteins that are involved in apoptosis have fewer numbers of (longer) IDRs, about 2.8 per chain, compared to proteins implicated in necroptosis and immune responses that have on average about 5 and 3.8 IDRs per chain, respectively. Overall, there were about 3.3 IDRs per chain when considering all cell death proteins. Our analysis indicated that intrinsic disorder is important across all cell death processes. However, the disorder profiles, quantified with the average content and number of regions per chain, vary between these processes.

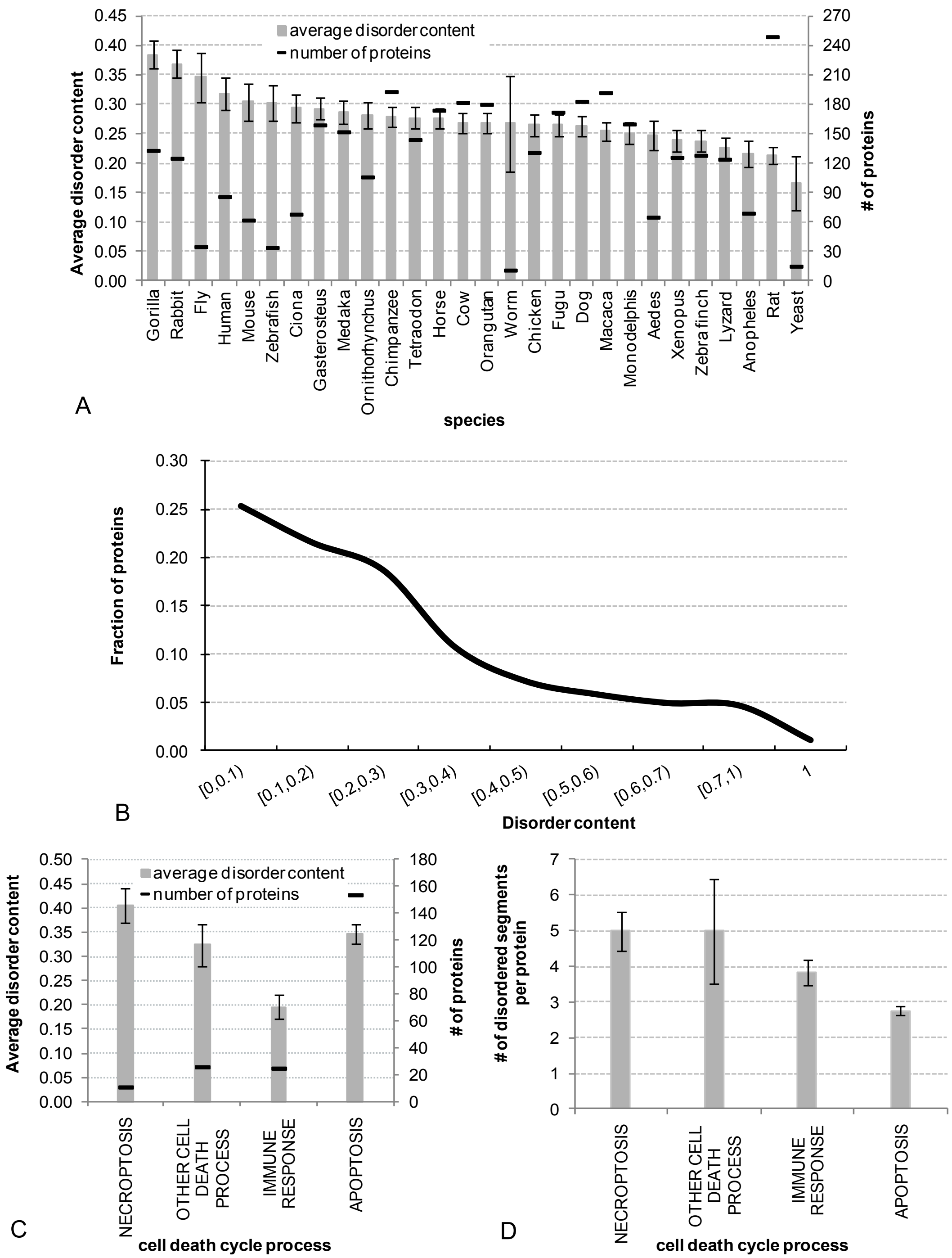

Figure 4.11. Evaluation of abundance of intrinsic disorder in PCD-related proteins from the Deathbase.

A. The average disorder content (grey bars) with standard errors (error bars) and the number of proteins for the 28 considered species shown on the *x*-axis. **B.** Distribution of disorder content for the cell death proteins for the 28 considered species. **C**. The average disorder content (grey bars) with standard errors (error bars) and the number of proteins for different cell death processes. **D.** The average number of disordered regions per protein (grey bars) with standard errors (error bars) for different cell death processes.

4.5.3 Functional Analysis of Disordered Regions

Distribution of the sizes of the IDRs in PCD-related proteins across the considered 28 species is given in **Figure 4.12A**. Interestingly, we observed that the region sizes followed a bimodal distribution with a relatively large number of short disordered regions, between 4 and 25 consecutive amino acids, and with a second peak for longer sizes, at over 30 amino acids. Thus, we analyzed the functional repertoire of the disordered regions separately for short (less than 30 consecutive amino acids) and long (at least 30 consecutive amino acids) regions.

To this end, we considered 26 functions associated with disorder, which were predicted based on the data in the DisProt database (Sickmeier, et al., 2007), that are summarized in **Table S1** in Appendix B. We excluded functions with less than 20 annotations for both short and long IDRs, as they do not offer enough data to draw statistically sound conclusions. **Figure 4.12B** compares the annotations of the 19 remaining putative functions for the short and long IDRs and reveals that intrinsic disorder plays a diverse set of roles in the PCD-related proteins, from facilitating interactions with other proteins, DNA, RNA, and other ligands, to involvement in the post-translational modification sites, intra-protein interactions, and implementation of linker regions. Both long and short IDRs were implicated with similar frequency in several functions including interactions with DNA and ligands and regulation of proteolysis. The short IDRs were more prevalent in a larger number of functions compared with the long IDRs, including protein-RNA and cofactor/heme binding, nuclear localization, protein inhibition, polymerization, trans-activation, regulation of apoptosis, and in auto-regulatory functions. They were also implicated in the entropic bristle activities and were twice more often associated with the post-translational modification sites. On the other hand, long IDRs more often served as linkers, were involved in the electron transfer, and played a strong role in the protein-protein, protein-lipid, and intra-protein interaction. Overall, disorder seems to be implicated in a diverse repertoire of cellular functions, with the primary function to implement binding events.

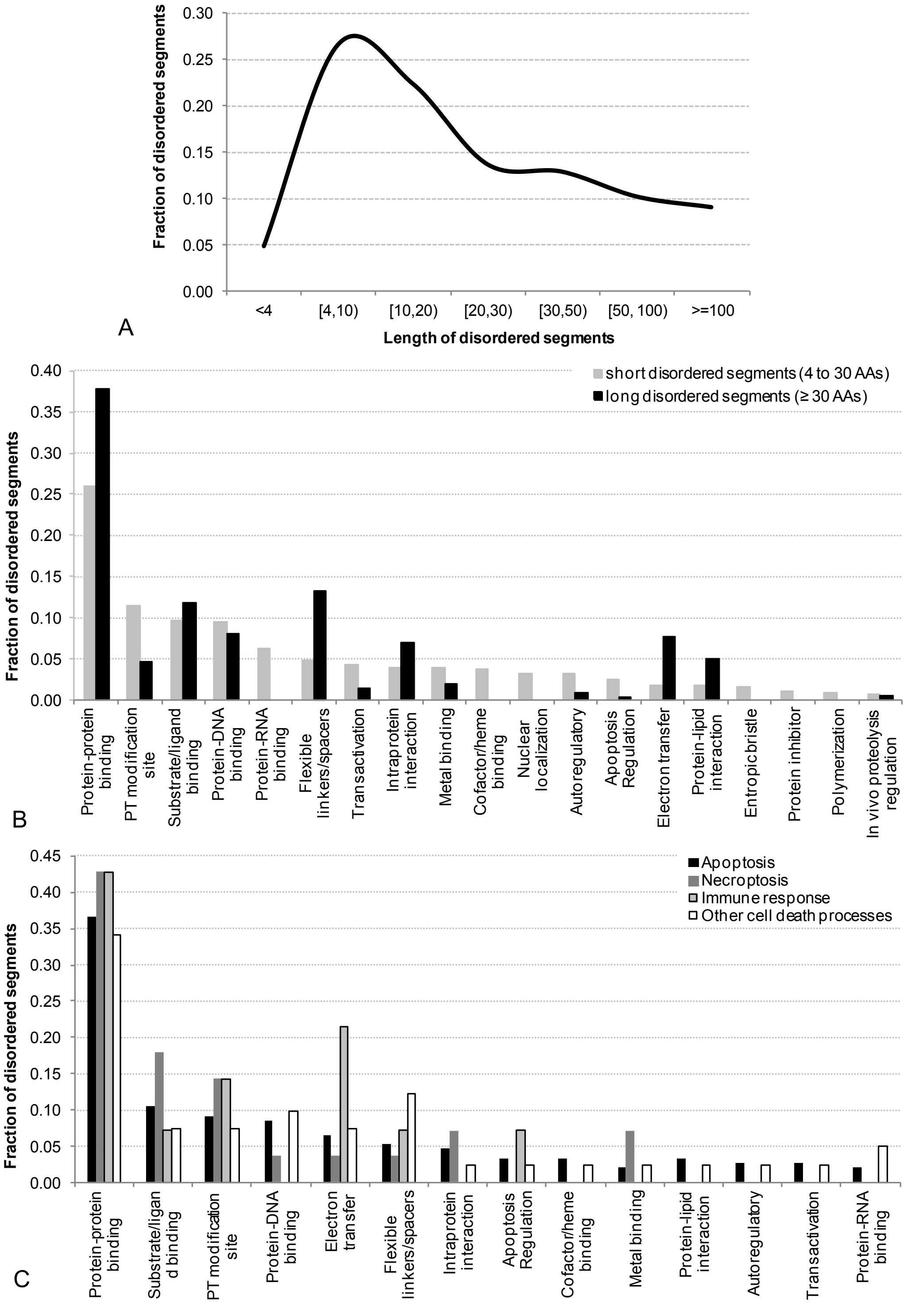

Figure 4.12. Correlation between intrinsic disorder in cell death proteins and their functions. **A**. Distribution of the length of the disordered regions in cell death proteins. **B**. Fraction of short, 4 to 30 consecutive amino acids (AAs), and long, ≥ 30 consecutive AAs, disordered regions for a given function; *x*-axis shows 19 considered functions sorted in descending order by the number of the short regions. **C**. Fraction of disordered regions for a given function for the cell death proteins in the curated species that are annotated with cell death processes; *x*-axis shows 14 considered functions sorted in descending order by the overall number of the disordered regions.

Furthermore, we investigated the functional roles of IDRs for the proteins associated with specific cell death processes. Since there were substantially fewer (only 128) disordered regions with the predicted functions for the five curated species that include cell death process annotations (compared to results in **Figure 4.12B**), we excluded the functions with <5 annotations. The remaining 14 functions are summarized in **Figure 4.12C**. Disordered regions found in all cell death processes were shown to be responsible for five functions: protein-protein and protein-ligand binding, electron transfer, linker regions, and were also prevalent in post-translational modification sites. The remaining functions were specific to selected subsets of the cell death processes. Intrinsic disorder implemented all considered 14 functions in proteins involved in apoptosis. In contrast, the IDRs in the necroptosis-related proteins were responsible for fewer functions, such as protein-protein, protein-ligand, and protein-DNA binding, intra-protein interactions and metal binding. Finally, immune response triggered cell death utilized disorder to facilitate electron transfer and protein-protein and protein-ligand interactions. To sum up, disorder is utilized to facilitate a different set of cellular functions for different cell death processes.

4.5.4 MoRF Regions in Proteins Involved in Programmed Cell Death

The most prevalent function of disorder in the cell death proteins is facilitation of the protein-protein interactions; **Figure 4.12B** shows that about 30% of the functionally annotated disordered regions were implicated in these binding events. This motivated our analysis of MoRFs regions (Mohan, et al., 2006; Oldfield, et al., 2005; Uversky, et al., 2010; Vacic, et al., 2007) that undergo coupled binding and folding upon interaction with protein partners. **Figure 4.13A** demonstrates that there are on average between 1.1 and 3 MoRF regions per protein across the considered 28 species. The average number of MoRFs per chain over all cell death proteins is at 2.3, which means that about two-thirds of all disordered regions, which there are on average 3.3 per chain, include MoRFs. This means that protein-protein interactions dominate the function of the disorder in these proteins and that the disorder is transitional as disordered regions might fold into a structured conformations upon binding to their protein partners. Furthermore, **Figure 4.13A** suggests that MoRFs primarily fold into helical or irregular conformations upon binding; this is consistent across all considered species.

Figure 4.13B shows that the proteins that control apoptotic pathways have the smallest number of MoRFs, at about 2.2 per chain. On the other hand, proteins involved in the necroptosis-driven cell death have the largest number MoRFs per sequence, which equals 3.5. These counts correlate with the overall number of the disordered regions, see **Figure 4.11D**, resulting in a similar ratio between the numbers of disordered and MoRF regions.

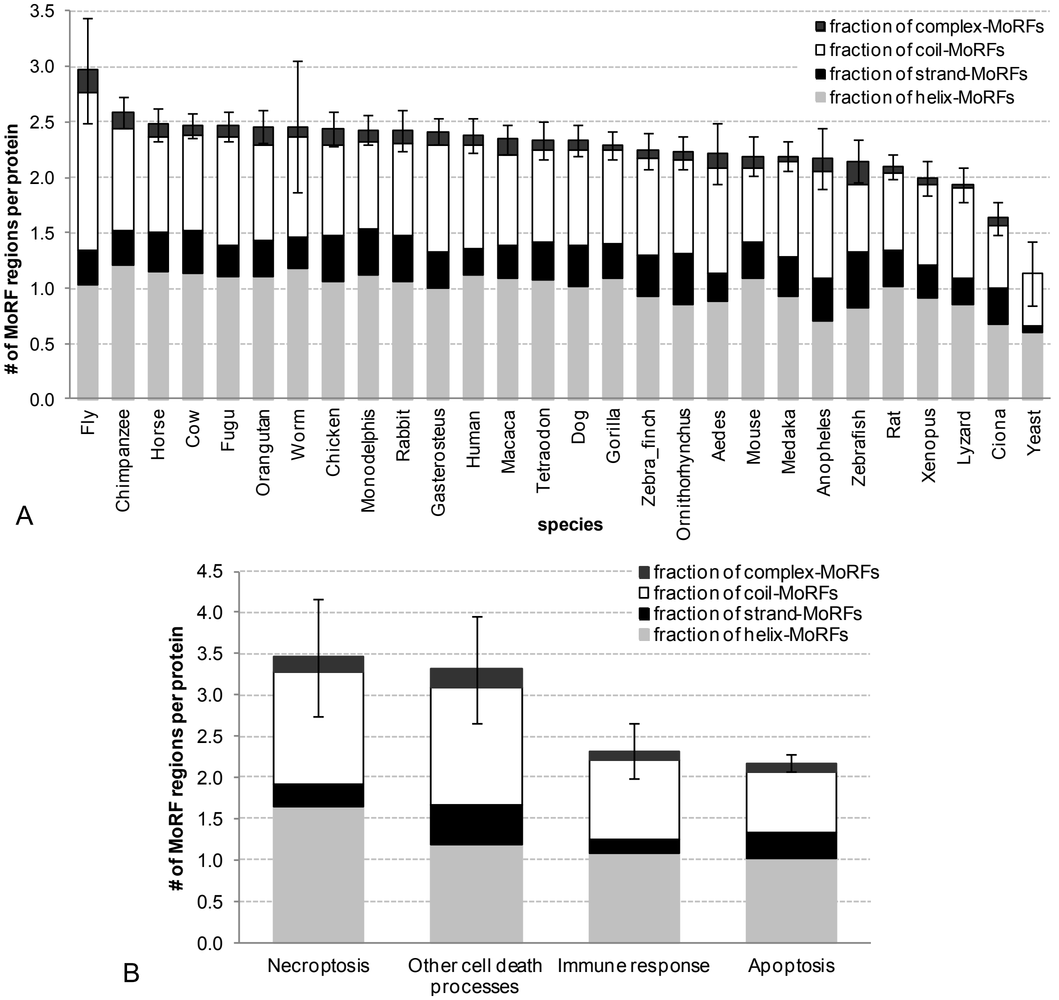

Figure 4.13. MoRF regions in PCD-related proteins.
A. Average number of MoRF regions per protein, shown using bars, with standard errors (error bars) across the considered 28 species shown on the *x*-axis. The bars are subdivided to represent different MoRF types. **B**. Average number of MoRF regions per protein, shown using bars, with standard errors (error bars) across the four types of cell death processes. The bars are subdivided to represent different MoRF types.

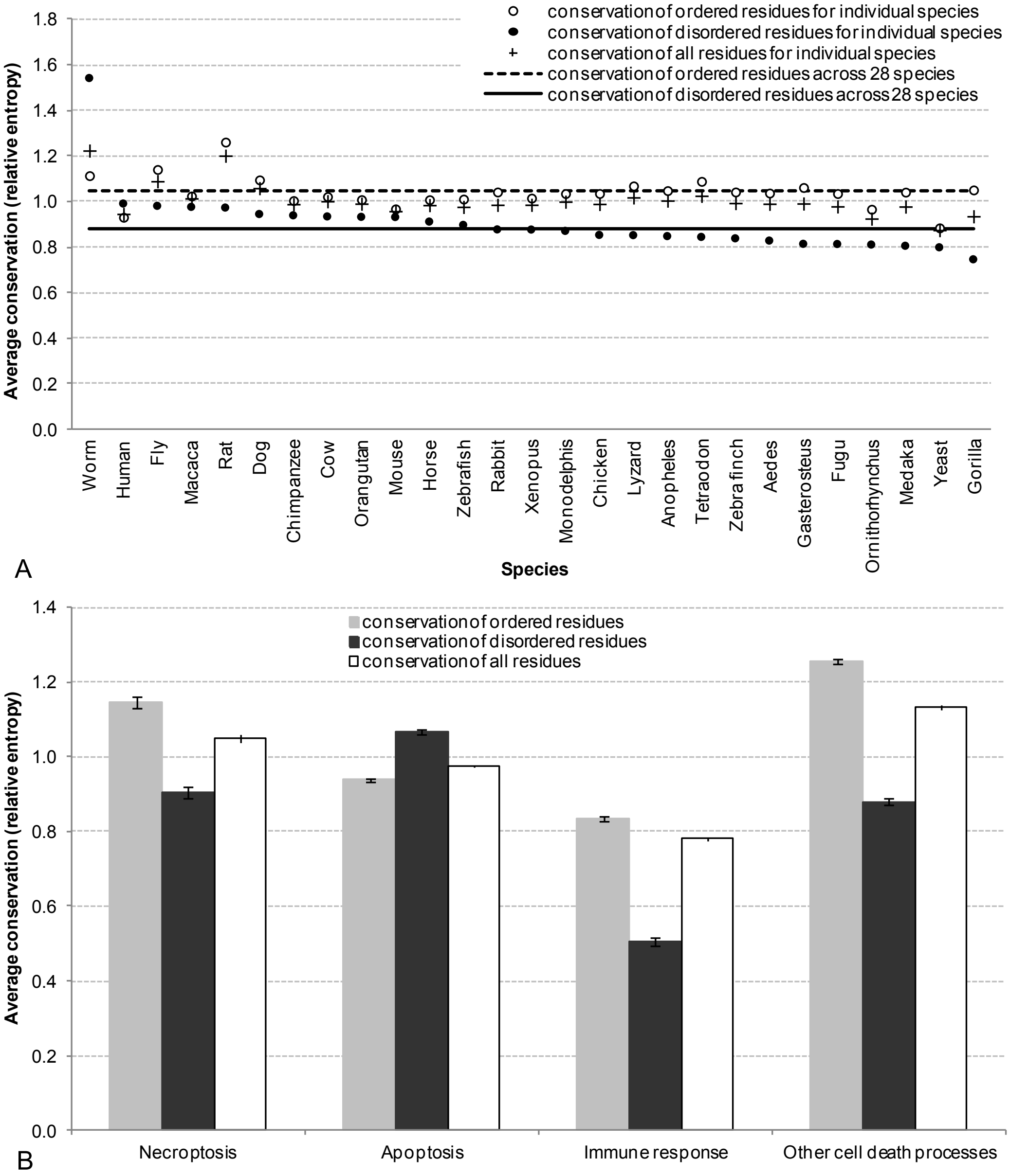

Figure 4.14. Evolutionary conservation for PCD-related proteins.
A. The average evolutionary conservation, quantified with relative entropy, for ordered (hollow circles), disordered (solid circles), and all (crosses) residues across the considered 28 species that are shown on the *x*-axis. Solid horizontal lines show average conservation for ordered (gray line) and disordered (black line) residues. **B**. Average relative entropy, which quantifies evolutionary conservation for ordered, disorder and all residues for each cell death process, shown using bars, with standard errors (error bars).

4.5.5 Evolutionary Conservation of Disorder in PCD-related Proteins

Finally, we investigated evolutionary conservation of the disorder in the cell death proteins across various species, **Figure 4.14A**, and across various processes, **Figure 4.14B**. The conservation was quantified with the relative entropy computed from the

WOP profiles generated by PSI-BLAST. Higher values of the relative entropy indicate a higher degree of evolutionary conservation. **Figure 4.14A** shows that for significant majority of species, except for worm and human, disordered residues were characterized by a consistently lower conservation compared to the structured amino acids. This suggested that disordered regions are under higher evolutionary pressure compared to the structured regions. **Figure 4.14B** demonstrates that the above trend holds for all cell death processes, except for the apoptosis where disordered amino acids have higher conservation. Immune response-related proteins were characterized by the highest difference in conservation where the disordered residues had conservation lower by about 40% compared to the ordered residues. As immune related processes require dynamic adjustments in response to relatively rapid evolution of pathogens, we speculate that these dynamics are facilitated through the use of disorder.

4.6 Discussion and Conclusions for Study on Cell Death Proteins

Our studies suggest that intrinsic disorder is very common in the PCD-related proteins in spite of the fact that many of these proteins are enzymes, such as kinases, ribonucleases, deoxyribonuclease, proteases, protein and ubiquitin ligases, polymerases, oxidureductase, GTPases, and so on.

There are many examples in the literature where the PCD-related proteins were experimentally found to be disordered or possess long IDRs. Since the literature on this topic is vast, only several characteristic cases of the experimentally validated disorder in PCD-regulating proteins are briefly discussed. A well-documented example is the transcription factor p53, which possesses functionally important IDRs (Dawson, et al., 2003; Lee, et al., 2000). About 70% of the interactions of p53 with other proteins and DNA are mediated by IDRs and these regions contain 86, 90, and 100% of the observed acetylation, phosphorylation, and protein conjugation sites (i.e., post-translational modification sites) in this protein, respectively (Oldfield, et al., 2008). Early on, X-ray and NMR structural analysis revealed that an important inhibitor of PCD, BCL-x_L, contains a 60-residue IDR connecting helices αl and α2 (Muchmore, et al., 1996). Other members of the BCL-2 family, BH3-only proteins (such as BIM, BAD, and BMF) that serve as key initiators of PCD are largely disordered in solution (Hinds, et al., 2007). Interaction of the disordered BIM, BAD, BMF, BAK, and tBID with several pro-survival BCL-2 family members leads to the formation of an α-helical segment that anchors these BH3-only proteins to their binding partners (Rautureau, et al., 2010; Sattler, et al., 1997). In addition to the BH3-only proteins, pro-survival BCL-2 family members include MCL-1, BFl-1, and BCB-B, all of which are expected to contain long functional IDRs (Rautureau, et al., 2010). Overall, structural and sequence analyses revealed that many pro-apoptotic and pro-survival proteins of the BCL-2 family are either IDPs or contain functionally important IDRs (Rautureau, et al., 2010). Furthermore, conformational plasticity and ability to fold upon binding were shown to have a crucial role in multifarious interactions of the BCL-2 family members with their numerous partners (Hinds, et al., 2005).

NMR spectroscopy in conjunction with circular dichroism spectroscopy and limited proteolysis revealed that a large central region (~1500 residue) of the BRCA1 tumor suppressor protein is a long IDR that lacks any pre-existing independently folded globular domains and serves as "an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions" (Mark, et al., 2005). NMR analysis revealed that an important regulator of proliferation and apoptosis, cAMPresponsive (CRE)-binding (CREB) protein (CBP), contains an intrinsically disordered ACTR-binding domain (residues 2059–2117) that completely folds upon binding (Demarest, et al., 2002). Combined experimental and computational analysis revealed that N- and C-terminal regions of the human Nogo proteins (Li, et al., 2007) and C-terminal domain of a mitochondrial pro-apoptotic protein ARTS (Reingewertz, et al., 2011) are typical IDRs, and that the prostate apoptosis response factor-4 (PAR-4) (Libich, et al., 2009) prostate-associated gene 4 (PAGE4) protein (Zeng, et al., 2011), and important oncoprotein c-Myc (Andresen, et al., 2012) are mostly disordered apoptosis-related proteins. The abovementioned and various other examples of experimentally validated disorder in the PCD-related proteins support our conclusions.

In the nutshell, we demonstrate that intrinsic disorder is very common across various cell death proteins, especially in the proteins involved in the necroptosis. Disorder also plays a diverse set of functional roles in these proteins, from facilitating interactions with other proteins, DNA, RNA, and other ligands, to involvement in the post-translational modification sites, intra-protein interactions, and implementation of linker regions. Thus, the intrinsic disorder plays a crucial role in the regulation and control of all cell death processes.

Chapter 5

High-Throughput Prediction of RNA, DNA and Protein Binding Mediated by Intrinsic Disorder

5.1 Introduction and Motivation

Previous characterization of a generic (without a given functional label) intrinsic disorder in nature (see Chapter 3 and Chapter 4) shows that IDPs/IDRs are not only common in nature, but also are fascinating entities that are involved in numerous cellular functions (Dunker, et al., 2005; Dunker, et al., 2001; Dunker, et al., 2008; Dunker, et al., 2008; Dyson, et al., 2005; Tompa, 2002; Uversky, et al., 2005; Wright, et al., 1999). For instance, IDPs/IDRs were shown to act as important regulators of protein-protein interaction (PPI) networks and were shown to be enriched in the RNA and DNA binding proteins. More specifically, enrichment in the intrinsic disorder is a common feature in hub proteins (Haynes, et al., 2006) that interact with multiple different protein partners in the PPI networks, and was suggested to participate in the programmed cell deaths processes that are regulated by a series of PPIs (Peng, et al., 2013). In addition to be involved in the protein-RNA interactions (Peng, et al., 2014), intrinsic disorder was also observed in about 50% and >41% of sequences of human transcription factors (Minezaki, et al., 2006) and histone proteins (Peng, et al., 2012), respectively.

In recent years, prediction of certain protein functions related to the RNA binding (Cirillo, et al., 2013; Puton, et al., 2012), DNA binding (Kauffman, et al., 2012), and protein-protein interactions (PPIs) (Shen, et al., 2007; Zhang, et al., 2012) generated strong interest. However, these predictions focus on the interactions that are extracted from crystal structures and thus which are primarily implemented by ordered/structured regions. On the other hand, prediction of functions of IDPs/IDRs also gains momentum. Lobley *et al.* show that inclusion of disorder information improves prediction of 26 functional GO categories related to signaling and molecular recognition (Lobley, et al., 2007). The ANCHOR method (Dosztanyi, et al., 2009; Meszaros, et al., 2009) predicts the protein-protein binding residues located in IDRs, while MoRFpred (Disfani, et al., 2012) predicts short protein-binding regions (i.e., 5-25 consecutive residues) located in longer IDRs. These attempts suggest that functions of IDRs are predictable from the protein sequence. The growing amount of annotations of these functions now allows for building and evaluation of the corresponding predictors. Release 6.01 of the DisProt database (Sickmeier, et al., 2007) includes 835 (out of total of 1513; see **Figure 2.5**) IDRs that are annotated with functions, where 35, 94 and 367 IDRs correspond to the disordered RNA-, DNA-, and protein-binding, respectively. For convenience, we use the disordered RNA-, DNA-, and protein-binding term to denote the RNA-, DNA-, and protein-binding mediated by IDRs. The availability of the annotated data, interest in the abovementioned binding events, and predictability of disorder-mediated functions motivated us to propose DisoRDPbind predictor. Our predictor has the following four characteristics:

1. **First attempt to predict multiple functions mediated by IDPs/IDRs**. DisoRDPbind is the first method that predicts disordered RNA- and DNA-binding residues, and it also predicts disordered protein-binding residues.
2. **High-throughput predictions**. DisoRDPbind predicts an average size protein with 450 residues in two seconds on a modern desktop computer; this means that our method can be applied on the genomic scale.
3. **Good predictive quality**. DisoRDPbind is empirically shown to obtain good predictive performance using two independent (from a training dataset) test datasets. Our method also provides accurate predictions when applied to find putative disordered RNA-, DNA-, and protein-binding regions on four complete proteomes/genomes.
4. **Complementarity to other predictors of DNA- and RNA-binding regions**. DisoRDPbind's predictions are empirically shown to complement predictions of representative methods that were built using ordered DNA- and RNA-binding residues, i.e., using annotations based on crystal structures.

5.2 Materials and Methods

5.2.1 Annotation of Disordered RNA-, DNA-, and Protein-binding

The DisProt database (Sickmeier, et al., 2007) includes IDRs that have been experimentally verified to implement over 30 functional subclasses (Dunker, et al., 2002). Using these annotations, the disordered RNA-binding is defined by combining five functional subclasses: *protein-tRNA binding*, *protein-genomic RNA binding*, *protein-rRNA binding*, *protein-mRNA binding*, and *protein-RNA binding*; the disordered DNA-binding combines three subclasses: *protein-DNA binding*, *DNA unwinding*, and *DNA bending*; and disordered protein-binding is composed of five subclasses: *protein-protein binding*, *autoregulatory*, *intraprotein interaction*, *protein inhibitor*, and *regulation of proteolysis in vivo*.

Table S3 in Appendix C provides further details.

5.2.2 Benchmark Datasets

We extracted 430 proteins from the release v5.6 of DisProt by removing proteins containing IDRs that are annotated with "Unknown" and "Disordered region is not essential for protein function". Next, we clustered these 430 proteins by utilizing CD-HIT (Huang, et al., 2010) with 40% sequence similarity. The resulting 385 protein clusters were divided into two subsets at random: one subset was used to create the TRAINING dataset with 315 proteins (283 clusters), and the other set constitutes the TEST115 dataset with 115 proteins (102 protein clusters). Consequently, the sequence similarity between TRAINING and TEST115 datasets is below 40%. The TRAINING dataset includes 16, 60, and 249 IDRs (from 14, 49, and 188 proteins, respectively) that are annotated with the RNA-, DNA- and protein-binding, respectively. TEST115 has 10, 14, and 80 IDRs (from 7, 13, and 61 proteins, respectively) with the RNA-, DNA- and protein-binding annotations, respectively. Next, we considered proteins that were recently deposited in DisProt, between releases v5.6 and v6.01, to build the second test dataset. As a result we collected 36 proteins that constitute the TEST36 dataset with 2, 9 and 11 proteins that have IDRs with the annotations of RNA-, DNA- and protein-binding, respectively. We note that a given IDR/residue can be annotated with multiple functions.

Table S3 in Appendix C summarizes these three datasets. The TRAINING dataset was used to design and compute the proposed DisoRDPbind method based on cross-validation. DisoRDPbind was then assessed and compared with other methods on the TEST115 and TEST36 datasets. The benchmark datasets together with the annotations of DNA-, RNA- and protein-binding are provided at http://biomine.ece.ualberta.ca/DisoRDPbind/.

We used complete proteomes of four popular eukaryotic model organisms that were collected from release 2013_04 of the UniProt database (Consortium, 2012) to evaluate DisoRDPbind on the genomic scale. We removed protein fragments based on the term "Fragment" in the subsection "Sequence status". The resulting proteomes include 42426, 33181, 25159 and 19656 proteins for *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*, respectively. Our predictions were compared against the known DNA and RNA-binding proteins in these proteomes that were annotated based on several recently developed resources including gene ontology (GO) terms (Blake, et al., 2008) in UniProt, RBPDB (Cook, et al., 2011) for the RNA-binding proteins, and animalTFDB (Zhang, et al., 2012) for the DNA-binding proteins. Considering the hierarchical structure of GO, we defined the RNA (DNA) binding by collecting the GO term RNA (DNA) binding itself and all of its children connected by "is_a" relation. Consequently, we collected 3298 RNA-binding proteins (GO_RNA) and 7880 DNA-binding proteins (GO_DNA) across these four proteomes. By mapping accession number of proteins from UniProt into RBPDB and animalTFDB resources, we obtained annotations of 1014 RNA-binding and 4089 DNA-binding proteins, respectively, over the four organisms.

We utilized the latest integrated database of protein-protein interaction (PPI) networks, mentha (Calderone, et al., 2013), for the assessment of the prediction of the disordered protein-binding regions on the genomic scale. We mapped proteins from UniProt into mentha and obtained an average of 21.4, 6.7, 5.2 and 7.3 interactions per protein for the 14547, 8006, 5005 and 8096 proteins from *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*, respectively.

For convenience, we used the source database name (RBPDB, animalTFDB and mentha) to represent the corresponding subset of proteins selected for the four organisms. **Table S2** in appendix C summarizes the five datasets: GO_RNA, GO_DNA, RBPDB, animalTFDB and mentha.

5.2.3 Evaluation Criteria and Statistical Significance

Evaluation criteria. DisoRDPbind outputs real values that quantify propensity of a given AA to participate in the DNA-, RNA-, and protein-binding event mediated by the intrinsic disorder. We assessed the predictive quality of these propensities using area (AUC) under the receiver operating characteristic (ROC) curves; see section 2.8.4. Here positive is a given type of disordered binding residue, and all other residues are set as negatives (non-binding). For example, when predicting the disordered RNA-biding residues, the RNA-biding residues annotated in DisProt (in the training or test datasets) are assumed as positives and all other annotated residues including the remaining disordered residues and all ordered residues are assumed as negatives. BLAST (Altschul, et al., 1997), which we use in our empirical study, provides only binary output and thus we were unable to compute its AUC values.

For the evaluation of binary outputs, we reported TP-rate of DisoRDPbind at FP-rate of 10%. Note that TP-rate of BLAST always equals to its sensitivity along with the changing of the threshold of FP-rate, since BLAST only provides binary prediction.

Statistical significance. We evaluated statistical significance of the differences in the AUC values between each considered predictor and DisoRDPbind. Specifically, we randomly selected half of chains from TEST115 or TEST36 100 times. Next, we compared AUC values of DisoRDPbind to a given considered method over the resulting 10 random subsets of each test dataset. If both of the corresponding vectors of AUC values are normal, as tested using the Anderson-Darling test at the 0.05 significance, then we utilized t-test; otherwise we used the non-parametric Wilcoxon rank sum test. We annotate the significance of the differences at the 0.05 and the 0.01 levels. When the *p*-value > 0.05, we assume a given considered method has predictive quality that is equivalent with DisoRDPbind, i.e., the difference in AUC values is not significant.

5.2.4 Assessment at the Whole Proteome Level

We also assessed the predictive quality of DisoRDPbind at the whole proteome level utilizing the five datasets: GO_RNA, GO_DNA, RBPDB, animalTFDB, and mentha. Since proteins in these datasets are annotated with a given function per sequence, we define the disordered RNA-, DNA-, and protein-binding proteins predicted by DisoRDPbind from our residue-level predictions as follows. First, we binarized the predicted propensities using the default 0.5 cut-off. We assume a given protein as the disordered RNA-, DNA-, and/or protein-binding protein if it has at least one predicted disordered RNA-, DNA-, and/or protein-binding regions composed of at least 4 consecutive residues, respectively. This is based on the prior works that also assume that the disordered regions include at least 4 consecutive disordered residues (Monastyrskyy, et al., 2011; Noivirt-Brik, et al., 2009).

To evaluate prediction of the disordered RNA-binding (DNA-binding) proteins for a given organism we calculated overlap between the set of the predicted disordered RNA-binding (DNA-binding) proteins and the proteins from the RNA-binding datasets GO_RNA and RBPDB (the DNA-binding datasets GO_DNA and animalTFDB). We also assessed statistical significance of this overlap by comparing it to an overlap with a randomly generated set of proteins. First, we selected at random half of the predicted RNA-binding (DNA-binding) proteins 10 times and estimated their overlap with the GO_RNA and RBPDB (GO_DNA and animalTFDB). Next, we selected at random the same number of proteins (compared to the number of predicted RNA-binding (DNA-binding) proteins) from a given complete proteome 10 times and computed their overlap with GO_RNA and RBPDB (GO_DNA and animalTFDB). We compared the 10 corresponding values of overlap to find whether the overlap of our predictions is significantly higher than a baseline defined based on overlap with a random set of proteins. If both vectors of the overlap values are normal, as tested using Anderson-Darling test at the 0.05 significance, then we utilized t-test; otherwise we used the non-parametric Wilcoxon rank sum test. We annotated the significance of the differences at the 0.005 level, i.e., we assume that the overlap of our predictions is equivalent to the overlap of random chains when *p*-value > 0.0005. Moreover, the RNA-binding (DNA-binding) proteins predicted by DisoRDPbind that do not overlap with the known RNA-binding (DNA-binding) proteins from GO_RNA or RBPDB (GO_DNA or animalTFDB) were further analyzed. We computed their sequence identity with the proteins in the GO_RNA and RBPDB (GO_DNA and animalTFDB) datasets in *H. sapiens*. Due to a substantial computational cost, we performed this analysis for the largest *H. sapiens* proteome. Prior work shows that protein function is preserved between proteins in the same genome that share at least 50% sequence identity (Addou, et al., 2009). We use this observation to assess whether our predictions are likely correct, i.e., we assume that a given predicted RNA-binding (DNA-binding) protein is correct if its identity to the known RNA-binding (DNA-binding) proteins is above 50%. We utilized the Needleman–Wunsch algorithm with the BLOSUM62 mutation matrix and gap opening and extension penalty of -11 and -1, respectively. The sequence identity was defined as the number of identical residues in the resulting alignment divided by the length of the shorter of the two aligned sequences.

Since majority of proteins interact with other protein(s), we cannot directly validate these predictions like we defined above for the RNA- and DNA-binding proteins. We note that hub proteins were shown to be enriched in IDRs (Haynes, et al., 2006). Therefore, we investigated relation between the promiscuity of a given protein (number of its proteins partners in the corresponding PPI network) and the number of its predicted disordered protein-binding regions to assess the predictive quality of DisoRDPbind at the whole proteome level. This relation was quantified with the Pearson Correlation Coefficient (PCC) between the average number of partners for proteins with a given number of predicted disordered protein-binding regions and this number of regions. We assert that our predictions of disordered protein-binding regions are likely correct if the PCC value is relatively high and positive. We analyzed the statistical significance of this PCC value by comparing it to a PCC value obtained using the average number of partners for a set of proteins with randomized number of predicted regions. First, we selected at random half of the proteins for each number of predicted disordered protein-binding regions in the mentha dataset 10 times and computed the PCC between the number of their predicted regions and their average promiscuity defined in mentha. We repeated the computation of PCC 10 times using randomly selected sets of proteins of the same size as the number of proteins with a given number of predicted regions and correlating this “randomized” number of regions with their average actual promiscuity extracted from mentha. We computed the statistical significance of the difference between these two vectors of 10 PCC values using the procedure described to assess the overlap for the prediction of the disordered RNA- and DNA-binding proteins.

5.2.5 Architecture of DisoRDPbind

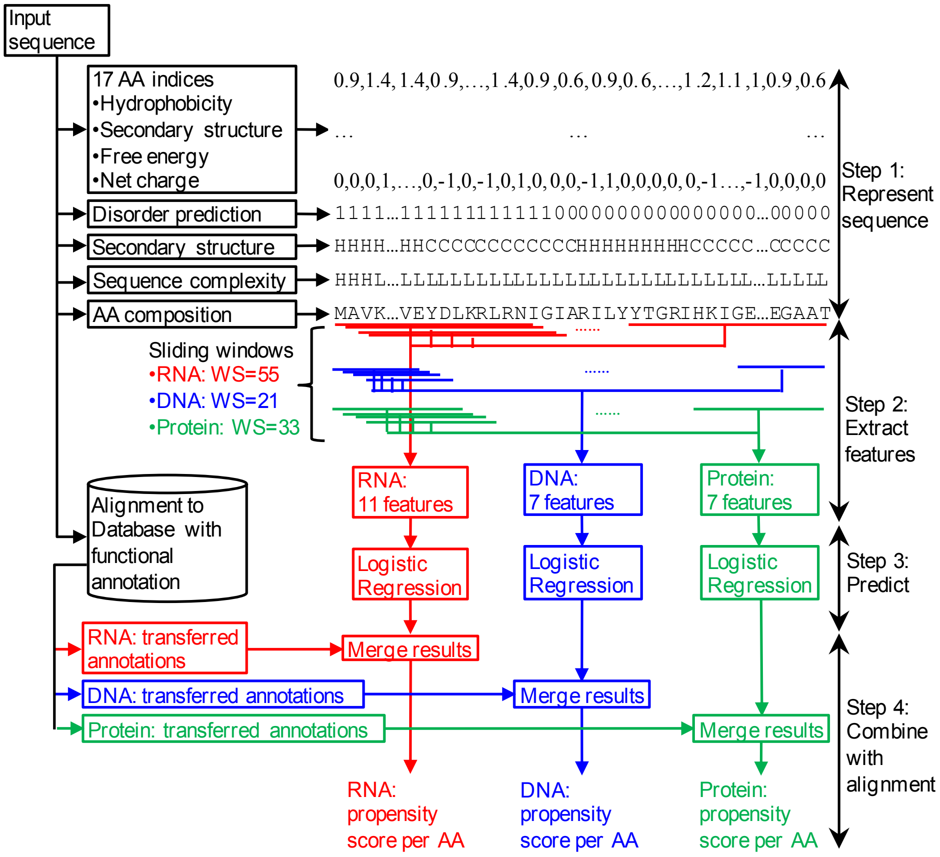

Figure 5.1. Architecture of DisoRDPbind

DisoRDPbind computes predictions based on four steps; see **Figure 5.1**. In step 1, the input protein sequence is represented using several numerical vectors, which quantify a variety of physiochemical properties of amino acids (AAs), predicted intrinsic disorder and secondary structure, estimated sequence complexity, and the AA composition. In step 2, a set of 11, 7, and 7 numerical features (values) are generated from these vectors for the prediction of RNA-, DNA-, and protein-binding residues, respectively. We considered a large number of features generated using sliding windows and performed empirical feature selection to obtain these small feature sets. In step 3, these selected features are inputted into a logistic regression model to predict the propensity score of the central residues in the window to participate in the disordered RNA-, DNA-, and protein-binding events. The choice of logistic regression was based on the fact that this model is popular in related areas, fast to compute, and provides the real-valued propensity score. In step 4, we transferred the annotations of RNA-, DNA-, and protein-binding based on sequence alignment generated by BLAST using annotated chains in the corresponding training dataset. These annotations are merged with the propensity scores generated by the regression to generate the final predictions. The design of DisoRDPbind was performed based on cross-validation using the TRAINING dataset.

5.2.6 Sequence Representation

The input sequence is represented by five types of numerical vectors: AA composition, sequence complexity, predicted secondary structure and disorder, and indices that quantify selected physiochemical properties of AAs.

We considered the AA composition, sequence complexity and secondary structure based on the observations that intrinsically disordered regions are enriched in certain AAs (Dunker, et al., 2001), have low sequence complexity (Romero, et al., 2001), and are biased in their secondary structure (Dosztanyi, et al., 2010; Dyson, et al., 2005; Liu, et al., 2002; Romero, et al., 2001; Uversky, et al., 2000). The sequence complexity was derived with the SEG algorithm (Wootton, 1994; Wootton, 1994; Wootton, et al., 1993; Wootton, et al., 1996). The secondary structure was predicted with the fast version of PSIPRED (McGuffin, et al., 2000), i.e., PSIPRED without using PSI-BLAST (Altschul, et al., 1997).

Several physicochemical properties of AAs, such as hydrophobicity, solvent accessibility, charge, free energy, etc., were successfully used to predict proteins with long disordered regions (Peng, et al., 2014), disordered protein-binding residues (Disfani, et al., 2012), and RNA- and DNA-binding residues annotated using crystal structures (Disfani, et al., 2012; Ma, et al., 2012; Walia, et al., 2012). We utilized a wide range of AA indices that quantify various physicochemical properties of AAs. However, these AA indices may be redundant to each other or irrelevant to our prediction. Thus, we empirically selected a subset of non-redundant and relevant indices using the TRAINING dataset; details are provided in Appendix D.

Inclusion of putative disorder was shown to improve accuracy of prediction of functions related to signaling and molecular recognition (Lobley, et al., 2007) and was successfully utilized to predict disordered protein-peptide binding (Disfani, et al., 2012). To assure that DisoRDPbind is runtime-efficient we utilized disorder prediction generated by fast IUPred (Dosztanyi, et al., 2005). This method predicts long and short disordered regions and globular domains. We used these three versions of IUPred.

Prediction for each residue in a given input chain uses information about the residue itself and its neighbors. We extract information from a sliding window of size *ws* that is centered on the predicted residue to calculate features that are used as inputs into the regression model. The use of the sliding window to calculate the features was inspired by previous related methods (Disfani, et al., 2012; Mizianty, et al., 2010). For the residues at the C or N-terminus of the sequence we reduce the window size on one side so it does not extend outside of the chain. We empirically derive the value of window size *ws* for each predicted function based on the size of the corresponding binding regions in the TRAINING dataset. We set *ws* to the value of 20th centile of the length of a given type of IDR, which translates into 55, 21 and 33 for the prediction of the disordered RNA-, DNA- and protein-binding residues, respectively.

Motivated by the recent work in (Disfani, et al., 2012), we aggregate values of the numerical vectors to generate features by calculating the difference between an average value of the near neighbors, i.e., (*ws*-1)/2 residues in the middle of the sliding window, and remote neighbors, i.e., (*ws*-1)/4 residues at each termini of the sliding window. We utilize this aggregation to contrast the values calculated using positions in the chain that are close to the predicted residue against the values associated with residues in a wider neighborhood in a sequence.

Detailed description of the calculation of features in the sliding windows using the AA composition, sequence complexity, predicted secondary structure and disorder, and AA indices is provided in Appendix E. In total, we consider 398 features.

5.2.7 Feature Selection

Some of the considered features could be redundant with each other or irrelevant to the prediction of the disordered RNA-, DNA- protein-binding residues, and thus we performed empirical two-step feature selection for each of the predicted functions. In step 1, we removed the irrelevant features. We analyzed the strength of relations between values of a given feature and the annotation of disordered RNA-, DNA-, and protein- binding residues in the TRAINING dataset; the relation was quantified with the point-biserial correlation (PBC) (Tate, 1954). If the strength of the relation for a given feature is low, i.e., |PBC value| < 0.02, then we removed this feature. In step 2, we further filtered the redundant and irrelevant features using wrapper feature selection (Kohavi, et al., 1997) utilizing logistic regression as the classifier/prediction model. This step maximizes predictive quality measured with AUC by varying feature sets; the predictions were done using 3+1-fold cross validation on the TRAINING dataset. This type of cross validation was introduced in (Disfani, et al., 2012) to reduce over-fitting. In the 3+1-fold cross validation, we fix one of the four cross validation folds as a test dataset and the remaining three folds are used to perform three-fold cross validation. We test each predictive model twice: based on the three-fold cross validation and based on the fourth test fold. The selection process starts by ranking all features in the descending order of their absolute PBCs computed on the TRAINING dataset. The set of selected features is initialized with the top ranked feature, which has consistent sign and at least 0.02 absolute PBC values across all four folds. We add a subsequently ranked feature to the set of selected features if it satisfy the same condition and if this addition improves AUC on both three-fold cross validation and the independent fourth test fold by at least 0.001. We scanned the ranked feature list once. This procedure resulted in the selection of 11, 7, and 7 features for the prediction of the disordered RNA-, DNA- and protein-binding residues, respectively. Only 17 AA indices are used to calculate the resulting selected features.

5.2.8 Regression Model

Logistic regression is a probabilistic classification algorithm that was extensively used in related prediction efforts including prediction of intrinsic disorder (Peng, et al.,

2013) and the ordered protein-RNA/-DNA/-protein interactions that were annotated using crystal structures (Bader, et al., 2004; Hwang, et al., 2007; Kuznetsov, et al., 2006; Lin, et al., 2004). This and the fact that prediction with the regression model is fast to compute motivated our selection of this model. The regression coefficients were estimated by using the ridge estimator implemented in the Weka platform (Hall, et al., 2009). The regression model provides three real-valued scores that correspond to the predicted propensity of a given AA to participate in the disordered DNA-, RNA-, and protein-binding. These values are merged with the outputs generated using sequence alignment; see Section 5.2.9.

5.2.9 Combining Regression with Sequence Alignment

We use sequence alignment with BLAST(Altschul, et al., 1997) to transfer annotation from the TRAINING dataset. For a given query chain, the annotations are transferred/copied for the similar positions in the alignment with a sequence that has high similarity quantified with the e-value. We chose 0.1 as the e-value cut-off, i.e., if the e-value < 0.1 then the aligned sequence(s) is regarded as similar and the annotations are copied. This cut-off was chosen based on 4-fold cross validation on the TRAINING dataset to maximize the average number of true positives and true negatives for the disordered RNA-, DNA- and protein-binding residues.

Empirical results when we transferred the annotations using alignment with BLAST on the 4-fold cross validation on the TRAINING dataset with the e-value cut-off of 0.1 show that BLAST nearly perfectly predicts negatives (i.e., specificity>=99%) and captures a small number of true positives (i.e., sensitivity<=5%). This conservative prediction (small number of high quality predictions of binding residues) is merged with the prediction from the regression. If a given residue is annotated with a given disordered function by the alignment then its propensity score is set to $(1+p_i)/2$, where p_i is the propensity score produced by the regression model and *i* denotes the particular function: disordered DNA-, RNA- or protein-binding; otherwise we use the prediction generated by the regression model. This raises values of the propensities generated by the regression for residues that were also predicted as binding by the alignment. The final propensities generated by DisoRDPbind combine the results of the regression model and alignment with BLAST against the annotated proteins from the TRAINING dataset.

5.3 Results and Discussion

5.3.1 Comparative Evaluation of DisoRDPbind

Since there are no methods that predict disordered RNA- and DNA-binding residues (i.e., RNA- and DNA-binding residues located in IDRs), we consider representative sequence-based methods that output propensity score for the prediction of ordered RNA- and DNA-binding (using annotations based on crystal structures) as the closest alternatives. We compare RNA-binding predictions of DisoRDPbind against predictions of BindN+ (Wang, et al., 2010) and RNABindR v2.0 (Walia, et al., 2012), and DNA-binding predictions against BindN+ and DNABR (Ma, et al., 2012). BindN+ is a popular method that predicts both RNA- and DNA-binging which was recently shown to provide accurate results (Chen, et al., 2012). RNABindR v2.0 and DNABR are the latest sequence-based methods for the prediction of ordered RNA- and DNA-binding residues, respectively. MoRFpred (Disfani, et al., 2012) and ANCHOR (Dosztanyi, et al., 2009; Meszaros, et al., 2009) are two recent predictors of the disordered protein-protein interacting residues, which focus on prediction of short and general (i.e., including short and long) regions, respectively. We compare our predictions of the disordered protein-protein binding residues with these two methods. **Table 5.1** compares predictive performance of these six methods with DisoRDPbind on the TEST115 and TEST36 dataset. We also include results for “Regression” which denotes DisoRDPbind without the BLAST-based alignment (without step 4 in **Figure 5.1**). Statistical significance was assessed over 10 repetitions with half of the test dataset; details are given in Section 5.2.3.

DisoRDPbind performs well based on the fact that it obtains AUC values ranging between 0.6 to 0.72, depending on the datasets and the predicted function; see **Table 5.1**. The TP-rate that was computed at the FP-rate of 0.1 shows that our predictions are characterized by a reasonable good predictive performance. For instance, considering prediction of RNA-binding residues DisoRDPbind provides TP-rate of 0.28 and 0.20 at the FP-rate of 0.10 on the TEST115 and TEST36 datasets, respectively; this means that the TP-rate is 2.8 and 2 times higher than the FP-rate of 0.1, respectively. The sensitivity (fraction of true positives from the set of all positives) of BLAST predictions alone is low, between 0.00 and 0.03, which means that BLAST adds only a few predictions to our DisoRDPbind. However, the inclusion of the alignment provides statistically significant improvements for the prediction of the disordered DNA- and protein-binding residues, although the magnitude of these improvements is relatively small; see results for DisoRDPbind and regression. This is consistent with the ROC curves shown in **Figure 5.2**. Comparison with the six considered predictors (including two versions of BindN+) shows that DisoRDPbind provides higher AUCs across predictions of DNA-, RNA- and protein-binding residues on both test datasets; also see **Figure 5.2**. This can be explained by the fact that BindN+, DNABR and RNABindR v2.0 predict binding in the ordered binding regions (regions annotated from crystal structures) and thus they secure lower predictive performance on the disordered binding regions that are considered here. The comparison of the predictions of the disordered protein-binding residues reveals that on average DisoRDPbind improves AUC by 0.02 and 0.11 compared to ANCHOR on the TEST115 and TEST36 datasets, respectively, and by 0.1 and 0.19 compared to MoRFpred, respectively. The analysis of the statistical significance indicates that these improvements are significant. The larger improvements over MoRFpred can be explained by the fact that this method predicts binding with short peptides (up to 25 residues) as opposed to DisoRDPbind that also predicts binding with longer chains.

Next, we investigated the correlation between predictions of DisoRDPbind and each of the six considered predictors, which was measured with PCC between their predicted propensity scores; see **Figure 5.3**. DisoRDPbind predictions have low correlation (PCC < 0.29) with each of the considered methods, except for ANCHOR (PCC >= 0.5). For the prediction of the RNA-binding residues, the PCCs between DisoRDPbind and BindN+ and RNABindR v2.0 are 0.03 (0.05) and 0.25 (0.29), respectively, both of which are about twice lower than the PCC of 0.52 (0.50) between BindN+ and RNABindR v2.0. This means that DisoRDPbind's predictions are complementary to/different from the predictions from BindN+ and RNABindR v2.0. Similarly, for the prediction of the DNA-binding residues the correlations between the predictions of DisoRDPbind, BindN+ and DNABR are also small and ≤ 0.25. This suggests that outputs generated by these three methods are different with each other. These results are consistent with the fact that BindN+, RNABindR v2.0 and DNABR focus on the prediction of RNA/DNA binding mediated by ordered residues, rather than disordered regions as in the case of DisoRDPbind. In contrast, for the prediction of the disordered protein-binding residues, DisoRDPbind's predictions are characterized by relatively high PCC > 0.5 with the outputs of ANCHOR, and lower correlation (PCC ≤ 0.38) with MoRFpred. This is also expected since MoRFpred predicts binding with the shorter peptides while both DisoRDPbind and ANCHOR predict binding of generic (short and long) disordered protein-binding regions.

Table 5.1. Comparison between DisoRDPbind and other six considered methods.
Predictive performance measured with AUC values and significance of difference in AUC when comparing DisoRDPbind with other methods for the prediction of the disordered RNA-, DNA-, and protein-binding residues on the TEST115 (above the double line) and the TEST36 (below the double line) datasets; "--" and "-" in the "Significance" column mean that DisoRDPbind has statistically significantly higher AUCs than the method in a given row at the *p*-value<0.01 and <0.05, respectively. Regression denotes DisoRDPbind without the BLAST-based alignment. AUC is the average AUC value on 10 random subsets with half of proteins from Test115 or from Test36. 'std' is the standard deviation of these 10 AUCs. TP-rate was measured at the FP-rate of 0.10.

Function	Method	AUC±std	*p*-value	Significance	TP-rate
RNA-binding on TEST115	BindN+	0.55±0.06	0.000713	--	0.13
	RNABindR 2.0	0.60±0.09	0.000416	--	0.14
	Regression	0.70±0.13	1	=	**0.28**
	DisoRDPbind	0.70±0.13	NA	NA	**0.28**
DNA-binding on TEST115	BindN+	0.64±0.04	0.001555	--	0.23
	DNABR	0.49±0.02	0.001953	--	0.10
	Regression	0.67±0.04	0.003062	--	**0.26**
	DisoRDPbind	0.68±0.03	NA	NA	**0.27**
Protein-binding on TEST115	ANCHOR	0.58±0.04	0.040446	-	**0.16**
	MoRFpred	0.50±0.02	0.000179	--	0.12
	Regression	0.59±0.05	2.22E-05	--	0.15
	DisoRDPbind	0.60±0.05	NA	NA	**0.16**
RNA-binding on TEST36	BindN+	0.55±0.09	0.001953	--	0.16
	RNABindR 2.0	0.64±0.05	0.027344	-	**0.20**
	Regression	0.66±0.06	1	=	**0.20**
	DisoRDPbind	0.66±0.06	NA	NA	**0.20**
DNA-binding on TEST36	BindN+	0.55±0.03	0.001953	--	0.10
	DNABR	0.52±0.05	0.001953	--	0.11
	Regression	0.62±0.08	0.003906	--	0.20
	DisoRDPbind	0.64±0.09	NA	NA	**0.26**
Protein-binding on TEST36	ANCHOR	0.61±0.09	1.07E-06	--	0.19
	MoRFpred	0.53±0.04	1.13E-05	--	0.12
	Regression	0.70±0.1	0.006006	--	0.31
	DisoRDPbind	0.72±0.1	NA	NA	**0.33**

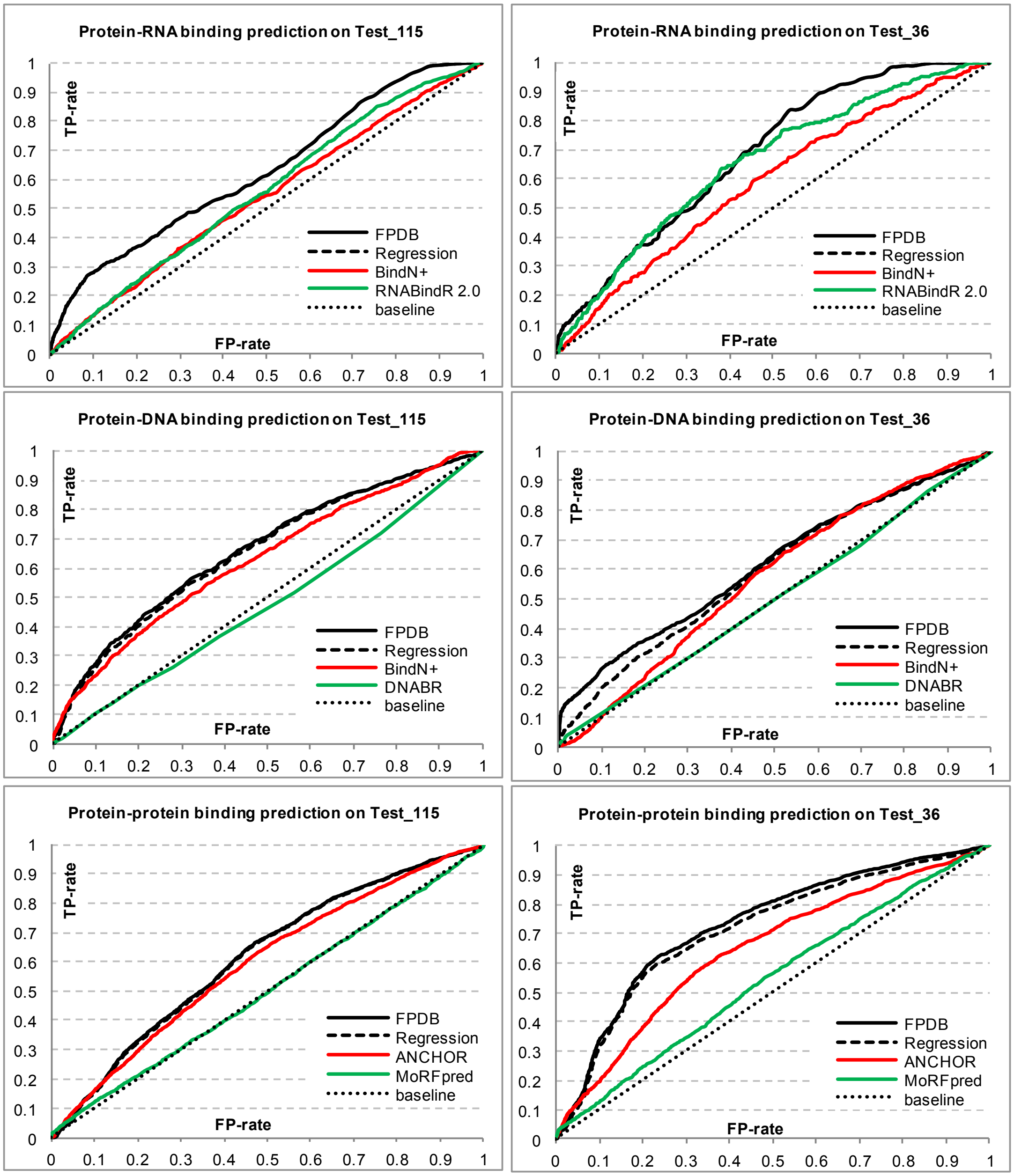

Figure 5.2. ROC curves for DisoRDPbind and other six considered methods.
ROC curves for the prediction of the disordered RNA- (left most panels), DNA- (panels in the middle column), and protein-binding (right most panels) residues on the TEST115 datasets (top three panels) and the TEST36 dataset (lower three panels), respectively. Dotted black line denotes baseline, which corresponds to the results obtained with a random predictor. Regression denotes the DisoRDPbind model without the BLAST-based alignment; we note that the ROC curves for the regression in the left most panels overlap with the curves for DisoRDPbind.

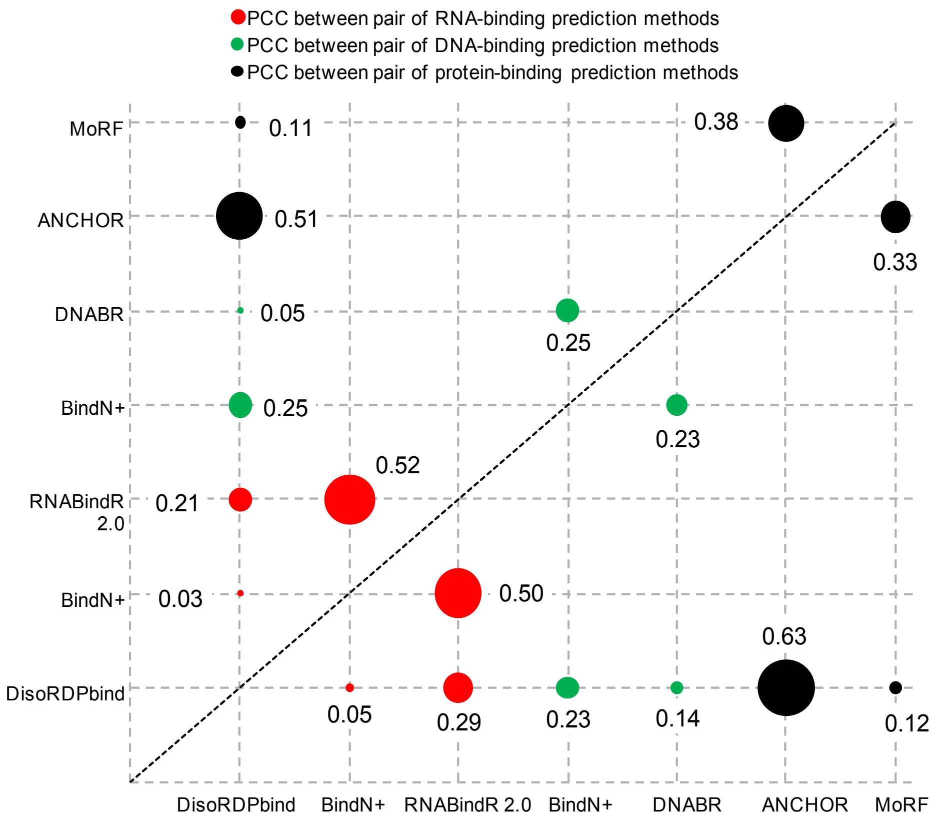

Figure 5.3. Relations between pair of methods.
PCC values between the propensity scores generated by the pairs of RNA- (red dots), DNA- (green dots) and protein- (black dots) binding predictors listed on the *x* and *y*-axes. The PCC values on the TEST115 and TEST36 are shown above and below the diagonal line, respectively. Size of the dots is proportional to the corresponding PCC value.

To sum up, DisoRDPbind offers good predictive quality for the three types of the disordered binding residues. Our empirical analysis shows that DisoRDPbind's outputs are different from (complementary to) the DNA- and RNA-binding predictions generated by BindN+, DNABR and RNABindR v2.0 that focus on the functions mediated by ordered regions. For the prediction of the disordered protein-binding, our predictions secure higher AUC and are correlated to the outputs generated by ANCHOR and MoRFpred.

5.3.2 Evaluation of Runtime

We assessed runtime of DisoRDPbind and compared it with the runtime of ANCHOR and PSI-BLAST (Altschul, et al., 1997) with one iteration (j=1) against the nr database; see **Figure 5.4**. The latter runtime is used to estimate the lower bound of the runtime of the other predictors including BindN+, RNABindR v2.0, DNABR, and MoRFpred. These methods utilize PSI-BLAST (multiple rounds) to generate their inputs. The runtimes were calculated based on the union of TEST115 and TEST36 (151 proteins) using a modern desktop computer. We analyze relative differences in the runtime, rather than the absolute values, since these are hardware independent. Although DisoRDPbind is slower than ANCHOR by up to two folds, we note that DisoRDPbind provides prediction of three considered functions at the same time. DisoRDPbind is at least 150 times faster than the one round of PSI-BLAST when considering different chain sizes. DisoRDPbind's prediction for a single chain takes between 0.3 seconds and 1 minute, depending on the chain length. The runtimes of DisoRDPbind, ANCHOR and PSI-BLAST are characterized by a quadratic increase with the chain size. **Figure 5.4** shows that quadratic function provides a good fit; the estimate with the quadratic fit has strong correlation with the measured runtime for DisoRDPbind (PCC=1) and PSI-BLAST (PCC=0.83).

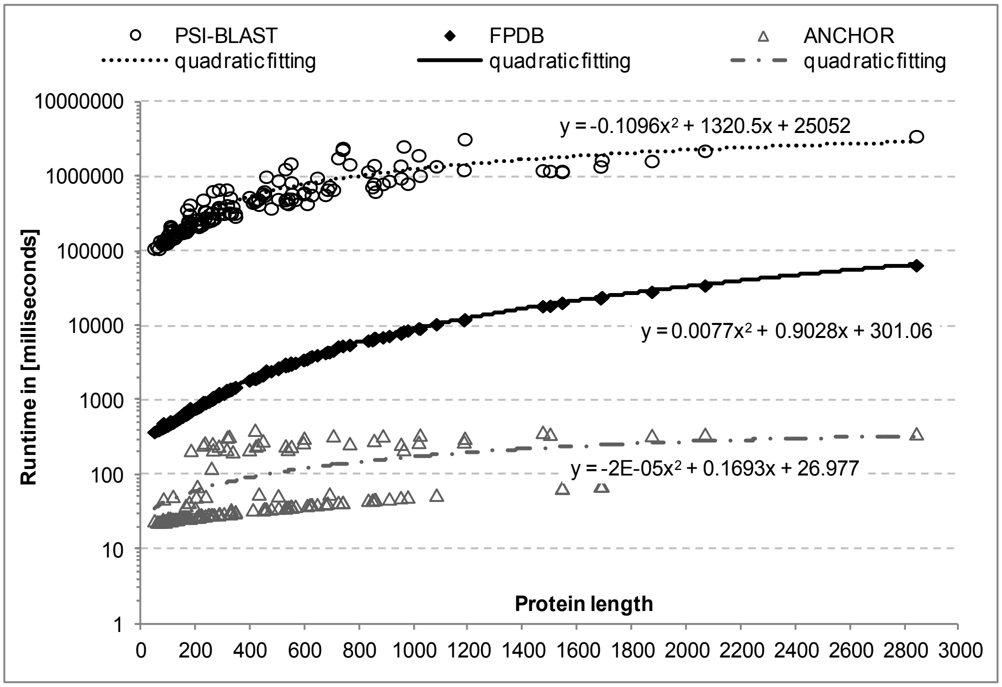

Figure 5.4. Evaluation of runtime of DisoRDPbind.
Relation between the length of protein chains (*x*-axis) and the runtime (*y*-axis in the logarithmic scale) computed for individual chains from the union of TEST115 and TEST36 using a modern desktop. The results include DisoRDPbind (solid diamonds), ANCHOR (hollow triangles), and one PSI-BLAST iteration (hollow circles). The thin black, thick gray and thick black lines represent the quadratic fitting for DisoRDPbind, ANCHOR and PSI-BLAST, respectively.

We also assessed the ability to perform prediction at the proteomic/genomic level with DisoRDPbind. We ran DisoRDPbind on the complete proteome of *H. sapiens*, the largest among the four considered proteomes. The total runtime of DisoRDPbind for these 42,426 chains is 45 hours. Using the quadratic fit from **Figure 5.4** the total runtime for DisoRDPbind and PSI-BLAST is estimated to be 43 hours and 260.8 days, respectively. This shows that the quadratic estimate is relatively accurate for DisoRDPbind (<5% difference to measured data). Compared to PSI_BLAST, DisoRDPbind is reasonably fast as it allows a full genome run in under 2 days.

5.3.3 Validation on the Whole Proteomes

Utilizing the runtime-efficient design of DisoRDPbind, we applied it to perform predictions for the four complete proteomes of *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*. We assessed its predictive quality on the corresponding datasets: GO_RNA, GO_DNA, RBPDB, animalTFDB, and mentha, for each of the four organisms. We assessed prediction of the disordered protein-binding by analyzing the relation between the promiscuity of a given protein in PPI networks and the number of its predicted disordered protein-binding regions. We evaluated the prediction of the disordered DNA- and RNA-binding by quantifying the overlap between the disordered RNA-/DNA-binding proteins predicted by DisoRDPbind and the RNA-/DNA-binding proteins annotated in the GO_RNA/GO_DNA and RBPDB/animalTFDB datasets, respectively; details are discussed in Section 5.2.4.

Validation of the disordered protein-binding predictions. For the four considered species, between 91% and 94% of proteins annotated in the mentha database are predicted by DisoRDPbind to have at least one disordered protein-binding region; see **Table S2** in Appendix C. We analyzed the relation between promiscuity of proteins (number of its proteins partners in the corresponding PPI network) and the corresponding number of the predicted disordered protein-binding regions to investigate whether DisoRDPbind provides accurate predictions. This was quantified with PCC between the average promiscuity of proteins with a given number of the predicted regions and this number of regions. The PCC values are 0.75, 0.57, 0.83 and 0.75 for the proteins from *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*, respectively; see **Figure 5.5**. This suggests that proteins predicted with more disordered protein-binding regions generally interact with more protein partners. We then assessed whether this observation is statistically significant by comparing these correlations with the correlations obtained when using proteins with randomized number of predicted disordered protein-binding regions, referred to as random PCC; details are given in Section 5.2.4. **Figure 5.5** shows that on average the original correlations are at least 2.7 times higher than the random PCC values, and that this increase is statistically significant. This means that the promiscuity of a given protein is significantly correlated with the number of its disordered binding regions predicted with DisoRDPbind. These results are consistent with the prior observation that hub proteins (that interact with at least ten partners) are significantly enriched with disorder compared to the proteins that interact with one partner (Haynes, et al., 2006).

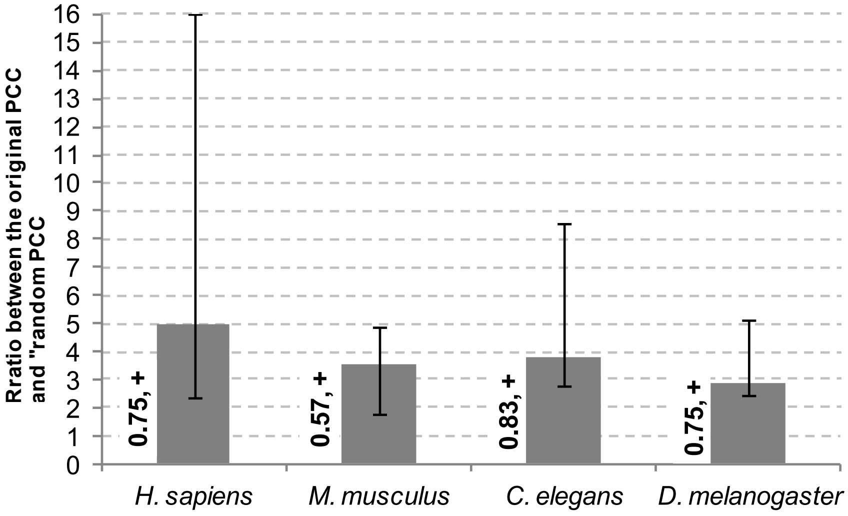

Figure 5.5. Correlation between the promiscuity of a protein and the number of its predicted disordered protein-binding regions.
Relation between the promiscuity of proteins in PPI networks and the number of the disordered protein-binding regions predicted with DisoRDPbind. The number left to a given bar is the PCC of the above relation for the corresponding species collected from mentha. Bars shows median ratio (over 10 repetitions with 50% of data) between these PCC values and the "random PCC", where the promiscuity values are shuffled. '+' inside a given bar means that the differences between these two PCC values is statistically significant at *p*-value < 0.0005. Error bars show the 30% to 70% centiles of this ratio for the 10 repetitions.

Validation of the disordered RNA- and DNA-binding predictions. DisoRDPbind predicted 2764 (2475), 1040 (2231), 722 (1241) and 792 (1140) proteins to be the disordered RNA-binding (DNA-binding) in *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*, respectively. **Table S2** in Appendix C shows that the predicted RNA-binding (DNA-binding) proteins have on average 32 to 37% (30 to 40%) disordered residues, which is 12 to 15% (7 to 17%) higher than the average disorder content (i.e., fraction of disordered residues) in these four species. This is consistent with prior works that showed that disorder is enriched in the RNA-/DNA-binding proteins (Dixon, et al., 2011; Peng, et al., 2012; Peng, et al., 2014); this also support our claim that DisoRDPbind's predictions are accurate.

We assessed accuracy of the DisoRDPbind's predictions of the disordered RNA-binding (DNA-binding) proteins by analyzing the overlap between the set of predicted RNA-binding (DNA-binding) protein and the known binding proteins from the GO_RNA and RBPDB (GO_DNA and animalTFDB) datasets, respectively; details are given in Section 5.2.4. **Figure 5.6** shows the overlap between the predictions of DisoRDPbind and the corresponding datasets. The overlap for the DNA-binding is larger than that for the RNA-binding. **Figure 5.7** further reveals that 11 to 14% (depending on the organism) of GO_RNA and 16 to 21% of RBPDB, and 20 to 39% GO_DNA and 38 to 50% of animalTFDB are predicted to be disordered RNA- and DNA-binding proteins by DisoRDPbind, respectively. Moreover, this overlap is 1.8 to 5 (3.5 to 10.3), depending on the organism, times higher than an overlap for a random set of the same number of proteins for the RNA-binding (DNA-binding), respectively (**Figure 5.7**); these differences are statistically significant. These results suggest that our predictions are accurate.

However, **Figure 5.6** shows that majority of the predicted disordered RNA-binding (DNA-binding) proteins are not included in the GO_RNA and RBPDB (GO_DNA animalTFDB) datasets. We refer to these proteins as novel putative binders. We computed sequence similarity between these novel RNA (DNA) binders and the proteins that are known to bind RNA (DNA) from the GO_RNA and RBPDB (GO_DNA and animalTFDB) datasets. Since computation of the pairwise sequence alignment is time consuming, we performed this analysis on the largest of the four proteomes from *H. sapiens*. The results reveal that 89% (98%) of the novel putative RNA (DNA) binders share at least 50% sequence similarity to the chains in the GO_RNA and RBPDB (GO_DNA and animalTFDB) datasets; see **Figure 5.8**. Previous studies show that functional annotations can be transferred between a pair of sequences from the same genome that share ≥ 50% sequence similarity (Addou, et al., 2009). This suggests that that these 89% and 98% of novel binders are likely correctly identified by DisoRDPbind as the RNA- and DNA-binding proteins, respectively.

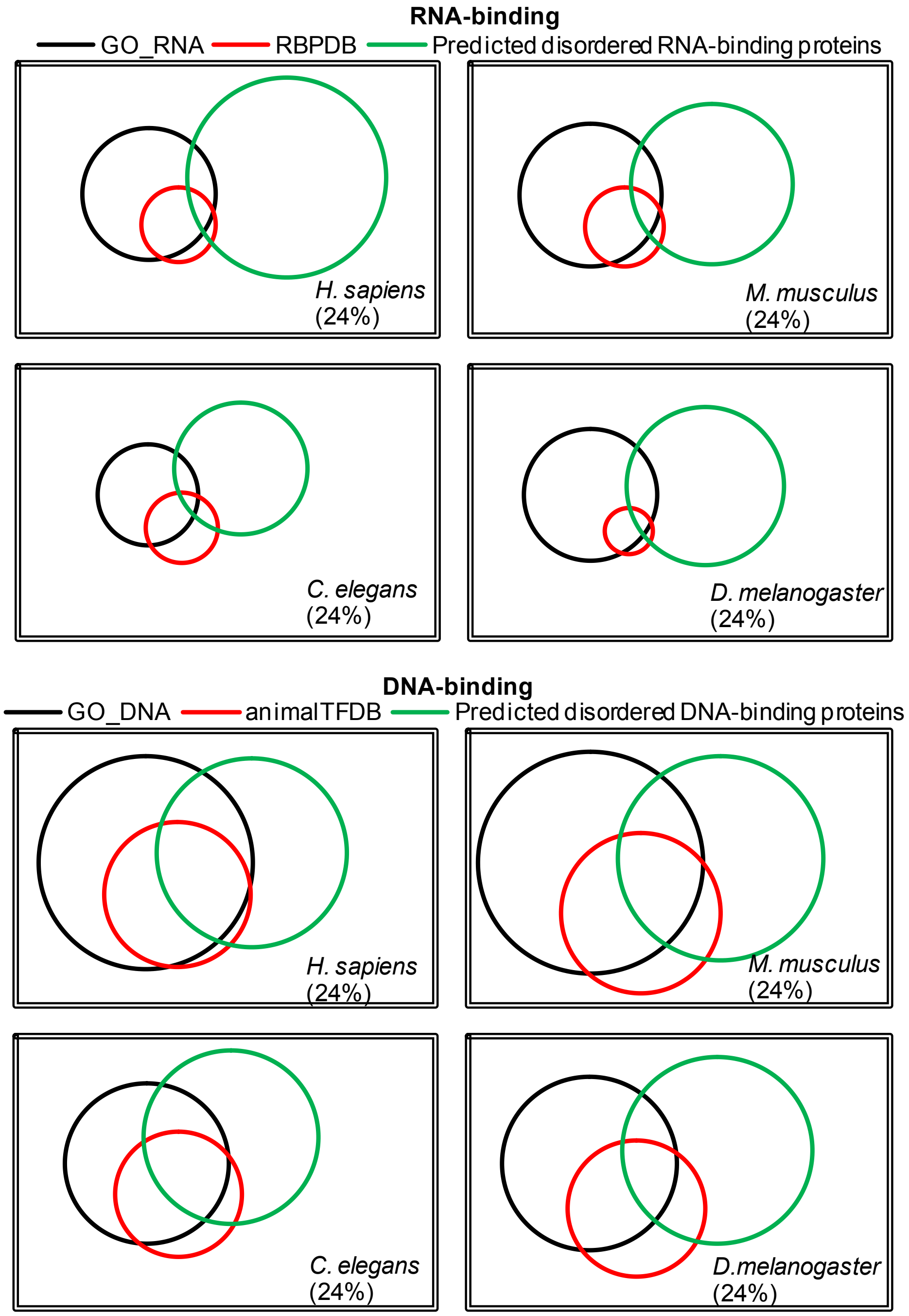

Figure 5.6. Venn diagrams of the overlap between the GO annotations and the prediction for RNA and DNA binding proteins.

Venn diagrams of the overlap between the set of disordered RNA-binding (DNA-binding) proteins predicted by DisoRDPbind and the known binding proteins collected from the GO_RNA (GO_DNA) and RBPDB (animalTFDB) datasets, respectively. Area of the rectangles represents 24% of the total size of a given proteome.

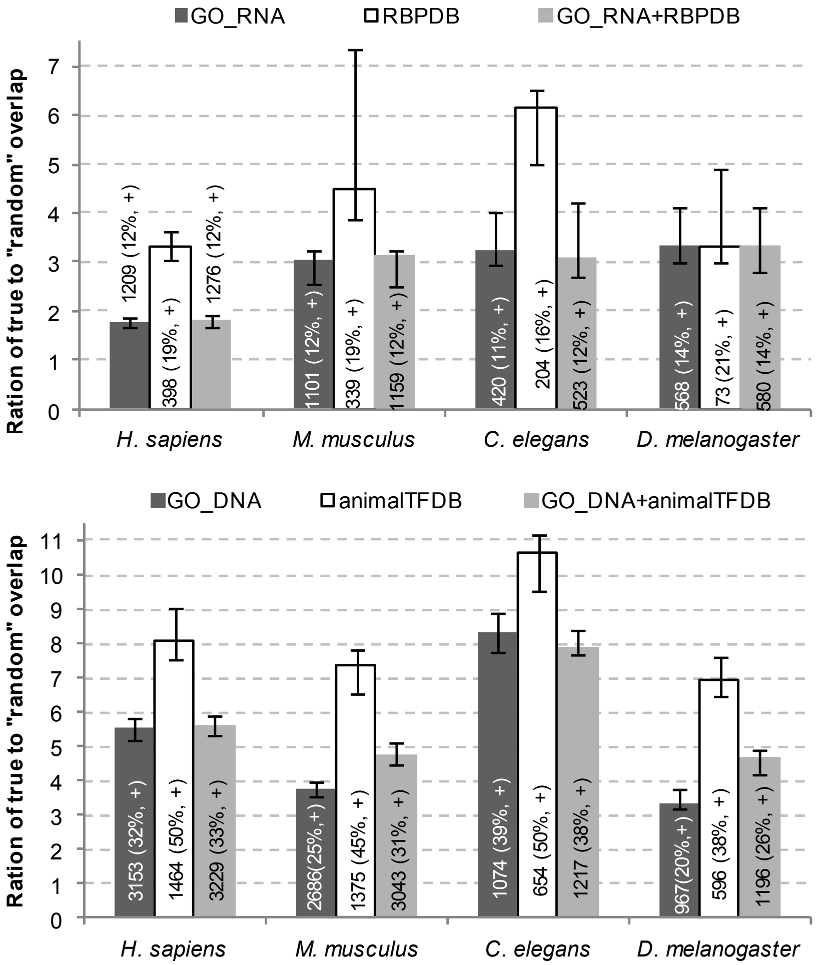

Figure 5.7. Significance of overlap between the GO annotations and the prediction for RNA and DNA binding proteins.

Median ratio (over 10 repetitions with 50% of the data) between the actual overlap between the RNA-binding (DNA-binding) proteins predicted by DisoRDPbind and proteins annotated in the GO_RNA and RBPDB (GO_DNA and animalTFDB), respectively, and the overlap of the proteins from these databases with a randomly chosen set of proteins. The numbers inside bars correspond to the number of chains in the corresponding database and the fraction shows amount of overlap with the predictions of DisoRDPbind; '+' means that the differences between the two values of overlap are statistically significant at *p*-value < 0.0005. Error bars show the 30% to 70% centiles of this ratio for the 10 repetitions.

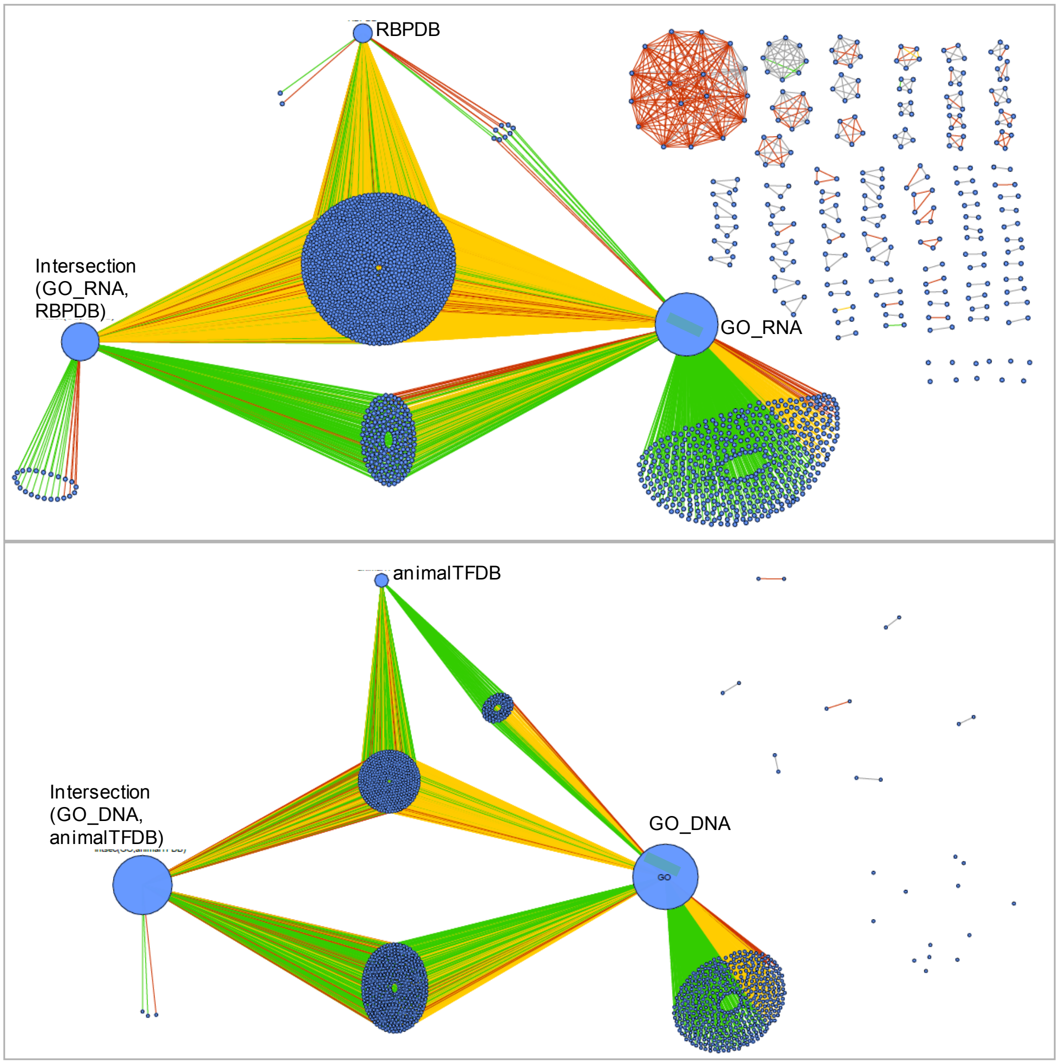

Figure 5.8. Sequence similarity between the novel putative RNA (DNA) binding proteins and proteins from GO_RNA and/or RBPDB (GO_DNA and animalTFDB) databases.
Sequence similarity between the novel putative RNA (DNA) binding proteins and proteins from GO_RNA and/or RBPDB (GO_DNA and animalTFDB) databases; top/bottom panel corresponds to RNA/DNA binding. Large blue dot on the right corresponds to the set of proteins in GO_RNA (GO_DNA) excluding the proteins from RBPDB (animalTFDB); at the top for the set of proteins in RBPDB (animalTFDB) but not in GO_RNA (GO_DNA); on the left for the set of proteins in both GO_RNA and RBPDB (GO_DNA and animalTFDB). The novel putative binders are shown using small blue circles that are connected by color-coded lines/edges to the large dots. The sequence similarity is represented by the color of lines, where red, yellow, green and gray indicate the similarity ≥80%, 60 to 80%, 50 to 60%, and 25 to 50%, respectively.

5.4 Conclusions

Although IDPs and IDRs are substantially different from structured proteins and regions, abundant and functionally important, nearly all existing sequence-based predictors, except ANCHOR and MoRFpred, focus on finding RNA-, DNA- and protein-binding residues that are annotated based on crystal structures, i.e., that are biased to be structured/ordered. To this end, we developed first-of-its-kind sequence-based method DisoRDPbind to predict these three functions facilitated by IDPs/IDRs. The selection of these functions was motivated by the availability of the corresponding experimental annotations and the fact that hub proteins and RNA- and DNA-binding proteins are enriched in disorder.

We utilized a comprehensive set of fast-to-compute features, which quantify selected physiochemical properties, AA composition, sequence complexity, and putative secondary structure and disorder, a regression classifier and sequence alignment to design DisoRDPbind. The resulting method can be used to perform predictions on the genomic scale. Empirical evaluation on two test datasets shows that DisoRDPbind offers good predictive quality for the three considered functions. Analysis of correlations between predictions of DisoRDPbind and the existing predictors of ordered (based on crystal structures) RNA- and DNA-binding residues show that their predictions are different and complimentary (since both types of methods provide good predictive performance on the respective benchmarks). DisoRDPbind provides predictions that are better and correlated with the outputs of ANCHOR when considering prediction of the disordered protein-binding. Analysis of DisoRDPbind's predictions on the four complete proteomes further strengthens our claim that DisoRDPbind's predictions are accurate. Our predictions suggest that promiscuity of proteins in PPI networks is correlated with the number of their disordered protein-binding, and reveal that a relatively large fraction of our predicted DNA- and RNA-binding proteins overlap with the known DNA- and RNA-binding proteins, and that majority of the non-overlapping predictions are similar to the known DNA and RNA binders.

A web server that implements the DisoRDPbind method and the benchmark datasets used to evaluate and compare our method are freely available for the research community at http://biomine.ece.ualberta.ca/DisoRDPbind/.

Chapter 6

Summary and conclusions

This thesis is focused on the systematic characterization of intrinsic disorder and the high-throughput prediction of RNA-, DNA- and protein-binding mediated by IDRs. These works provided interesting insights into the profile and the cellular functions and localization of the intrinsic disorder across all kingdoms of life, and demonstrated that high-throughput prediction of functions of intrinsic disorder is feasible. Our journey with the intrinsic disorder has started with the project that investigated the natural abundance, the functional roles, and the cellular localizations of intrinsic disorder in 965 complete proteomes across the four kingdoms of life including eukaryota, bacterial, archaea and viruses. The results show that IDPs/IDRs are very common in all the kingdoms of life, including viruses, perform specific set of cellular functions, and are preferentially located in certain parts of the cell, such as nucleus and ribosome. In each kingdom, intrinsic disorder has a unique profile, is involved in different functions, and has its own preference for cellular location. These observations led us to perform systematic investigation of the intrinsic disorder in two specific protein families – ribosomal and cell death proteins. This study reveals that intrinsic disorder is present in majority of ribosomal proteins and that it helps to implement functions that are crucial to the assembly of the ribosome and the translation process. We also demonstrated that intrinsic disorder is very common across various cell death proteins, especially in proteins involved in the necroptosis, and it also plays a diverse set of functional roles in these proteins. The abovementioned works suggests that intrinsic disorder is functionally important in RNA-, DNA- and protein-binding proteins. However, when analyzing the functional roles of IDRs in ribosomal and cell death proteins, we found that there were no existing methods focused on the prediction of function of intrinsic disorder. Since we can now obtain sufficient amount (for predictive and evaluative purposes) of functionally annotated IDRs from a recent release of the DisProt database of (i.e., after 2010; see **Figure 2.5**), we used these data to build a novel predictive model

DisoRDPbind. We concluded our journey by building and empirically assessing DisoRDPbind, which predicts the disordered RNA-, DNA- and protein-binding residues. We found that this method provides relatively accurate predictions, which due to the short runtime can be generated at the proteomic level.

6.1 Major Contributions

My major contributions are divided by the below aims.

- Aim 1: To perform first-its-kind comprehensive and detailed analysis of the abundance and the cellular functions of IDPs/IDRs in all complete proteomes.
    - Designed a majority-vote consensus method, by combining several fast disorder predictors, to obtain putative intrinsic disorder across 965 complete proteomes.
    - Estimated the amount of intrinsic disorder for each proteome and each kingdom of life, which was then utilized to discuss the difference in disorder profiles among the considered kingdoms of life.
    - Performed the systematic and large-scale investigation of the biological processes and molecular functions across the four kingdoms of life.
    - Conducted the empirical and detailed analysis of the cellular localization of intrinsic disorder across the kingdoms of life.
    - Mapped the cellular localization of intrinsic disorder into the eukaryotic, bacterial and archaean cell organelles.
    - Investigated the relations between intrinsic disorder and evolutionary pace for eukaryotic and bacterial species.
- Aim 2: To investigate the prevalence and the biological importance of intrinsic disorder in two protein families -- ribosomal proteins and proteins involved in the programmed cell death.

    - Compared the disorder profiles in ribosomal proteins with the corresponding profiles in the RNA and DNA binding proteins.
    - Estimated the abundance of intrinsic disorder in ribosomal proteins and in programmed cell death proteins across multiple species.
    - Investigated the difference in putative functions between short (4 to 30 consecutive disordered residues) and long (≥30 residues) disordered regions extracted from the ribosomal proteins and from the programmed cell death proteins.
    - Analyzed the functional roles of intrinsic disorder in different programmed cell death processes.
    - Investigated occurrence of different types of MoRFs in the ribosomal and cell death proteins.
    - Computed and discussed the evolutionary conservation of intrinsic disorder in the ribosomal and cell death proteins.
- Aim 3: To develop the first computational method that accurately and in high-throughput fashion predicts RNA, DNA and protein binding regions located in IDRs in protein sequences.
    - Investigated the relations between intrinsic disorder and a variety of physiochemical properties of amino acids.
    - Designed the first computational method DisoRDPbind that predict RNA, DNA and protein binding mediated by intrinsic disorder in the high-throughput manner.
    - Empirically assessed the predictive quality of DisoRDPbind on two independent (from training) test datasets.
    - Compared predictive performance and predicted outputs between DisoRDPbind and several related methods that predict RNA, DNA and protein binding..
    - Evaluated the runtime of DisoRDPbind, and compared it with the runtime of the related methods.
    - Applied DisoRDPbind on four eukaryotic proteomes to validate predictions from DisoRDPbind at the proteomic level.

This thesis includes materials and results from the following publications (including submitted works):

- Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Uversky, V.N. and Kurgan, L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in a thousand proteomes from all kingdoms of life. *Cellular and Molecular Life Science*, accepted on Jun 18, 2014.
- Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L. and Uversky, V.N. (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cellular and Molecular Life Science*, 71(8), 1477-1504
- Peng, Z., Xue, B., Kurgan, L. and Uversky, V.N. (2013) Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. *Cell Death & Differentiation*, 20, 1257-1267.
- Peng, Z. and Kurgan, L. High-throughput prediction of RNA, DNA, and protein binding regions mediated by intrinsic disorder. Submitted.

6.2 Major Findings

By analyzing the abundance of the intrinsic disorder across all kingdoms of life, including viruses, we confirmed that intrinsic disorder is common across these kingdoms of life. More importantly, we observed that each kingdom of life has a unique profile of intrinsic disorder, i.e., disorder content and number/size of IDRs in relation to the size of proteins. Specifically, viruses have relatively large number of long (over 300 AAs) fully disordered chains and the widest range of disorder content. Compared to bacteria and archaea, eukaryotic proteomes are characterized by a larger fraction of proteins with larger amounts of disorder, and comprised of a larger number of longer disordered regions. Moreover, short eukaryotic proteins contain the highest amount of predicted disorder and also long disordered proteins in eukaryotes seem to have a preferred range of length (1,500-2,000 residues) where they have higher amounts of disorder. These characteristics suggest that intrinsic disorder is utilized by different cellular functions/processes in difference kingdoms of life.

We validated this hypothesis by investigating the enrichment of intrinsic disorder in a wide range of molecular functions and/or biological processes. In eukaryota, intrinsic disorder seems to be important for the protein-RNA, protein-DNA, and protein–nucleotide interactions. In bacteria, the disorder is enriched in many key processes including transcription, translation, nucleosome assembly/chromosome condensation, protein polymerization and dimerization, catabolic and metabolic processes, and pathogenesis. The archaean proteins use intrinsic disorder for translation, and viruses use it for the RNA binding and for interactions with other organisms. In agreement with these observations of functional roles of the intrinsic disorder, our analysis of cellular localization of disorder shows that IDPs/IDRs are preferentially located in certain cellular component including ribosome, nucleus/nucleiod, peroxisome, cytoskeleton, etc.

Our systematic analysis of ribosomal proteins and proteins involved in programmed cell death further demonstrated that intrinsic disorder is a common feature of the protein-RNA, protein-DNA, and protein-protein interactions.

Lastly, we developed the first computational method, DisoRDPbind, which predicts RNA, DNA and protein binding mediated by the intrinsic disorder. Our empirical analysis revealed that certain physiochemical properties of amino acids are associated with these three binding events mediated by the disorder. The empirical assessment showed that DisoRDPbind provides good predictive quality, and is complementary to/different from the existing methods that predict these three functions for the ordered regions. These observations suggest that these three functions of the intrinsic disorder are predictable from the protein sequence.

Overall, the work presented in this thesis confirmed some prior observations about intrinsic disorder and also generated new knowledge concerning profiles and cellular functions and localization of IDPs/IDRs. Our predictor of functional IDRs provides a viable solution to annotate the RNA, DNA, and protein binding events mediated by IDRs in the high-throughput fashion.

Bibliography

Addou, S., Rentzsch, R., Lee, D. & Orengo, C.A. (2009) Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer, *Journal of Molecular Biology*, **387**, 416-430.

Adler, A.J., Greenfield, N.J. & Fasman, G.D. (1973) Circular dichroism and optical rotatory dispersion of proteins and polypeptides, *Methods in Enzymology*, **27**, 675-735.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389-3402.

Anderson, T.W. & Darling, D.A. (1952) Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes, *Annals of Mathematical Statistics*, **23**, 193-212.

Andresen, C., Helander, S., Lemak, A., Fares, C., Csizmok, V., Carlsson, J., Penn, L.Z., Forman-Kay, J.D., Arrowsmith, C.H., Lundstrom, P. & Sunnerhagen, M. (2012) Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding, *Nucleic Acids Research*, **40**, 6353-6366.

Anson, M.L. & Mirsky, A.E. (1925) On Some General Properties of Proteins, *Journal of General Physiology*, **9**, 169-179.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics*, **25**, 25-29.

Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks, *Nature Biotechnology*, **22**, 78-85.

Ban, N., Nissen, P., Hansen, J., Moore, P.B. & Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution, *Science*, **289**, 905-920.

Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G. & Yusupov, M. (2011) The structure of the eukaryotic ribosome at 3.0 Å resolution, *Science*, **334**, 1524-1529.

Ben-Shem, A., Jenner, L., Yusupova, G. & Yusupov, M. (2011) Crystal structure of the eukaryotic ribosome, *Science*, **330**, 1203-1209.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The Protein Data Bank, *Nucleic Acids Research*, **28**, 235-242.

Bertolazzi, P., Bock, M.E. & Guerra, C. (2013) On the functional and structural characterization of hubs in protein-protein interaction networks, *Biotechnology Advances*, **31**, 274-286.

Bhalla, J., Storchan, G.B., MacCarthy, C.M., Uversky, V.N. & Tcherkasskaya, O. (2006) Local flexibility in molecular function paradigm, *Molecular & Cellular Proteomics*, **5**, 1212-1223.

Bialik, S., Zalckvar, E., Ber, Y., Rubinstein, A.D. & Kimchi, A. (2010) Systems biology analysis of programmed cell death, *Trends in Biochemical Sciences*, **35**, 556-564.

Blake, J.A. & Harris, M.A. (2008) The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis, *Current Protocols in Bioinformatics*, **Chapter 7**, Unit 7.2.

Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R. & Klug, A. (1978) Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits, *Nature*, **276**, 362-368.

Bode, W., Schwager, P. & Huber, R. (1978) The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution, *Journal of Molecular Biology*, **118**, 99-112.

Bracken, C. (2001) NMR spin relaxation methods for characterization of disorder and folding in proteins, *Journal of Molecular Graphics and Modelling*, **19**, 3-12.

Brodersen, D.E., Clemons, W.M., Carter, A.P., Wimberly, B.T. & Ramakrishnan, V. (2002) Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA, *Journal of Molecular Biology*, **316**, 725-768.

Brown, C.J., Johnson, A.K. & Daughdrill, G.W. (2010) Comparing models of evolution for ordered and disordered proteins, *Molecular Biology and Evolution*, **27**, 609-621.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J. & Dunker, A.K. (2002) Evolutionary rate heterogeneity in proteins with long disordered regions, *Journal of Molecular Evolution*, **55**, 104-110.

Burra, P.V., Kalmar, L. & Tompa, P. (2010) Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes, *PLoS One*, **5**, e12069.

Calderone, A., Castagnoli, L. & Cesareni, G. (2013) mentha: a resource for browsing integrated protein-interaction networks, *Nature Methods*, **10**, 690-691.

Cessie, L. & VAN Houwelingen, J.C. (1992) Ridge Estimators in Logistic Regression, *Applied Statistics*, **41(1)**, 191-201.

Chang, C.K., Sue, S.C., Yu, T.H., Hsieh, C.M., Tsai, C.K., Chiang, Y.C., Lee, S.J., Hsiao, H.H., Wu, W.J., Chang, W.L., Lin, C.H. & Huang, T.H. (2006) Modular organization of SARS coronavirus nucleocapsid protein, *Journal of Biomedical Science*, **13**, 59-72.

Chen, J.W., Romero, P., Uversky, V.N. & Dunker, A.K. (2006) Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder, *Journal of Proteome Research*, **5**, 888-898.

Chen, K., Mizianty, M.J. & Kurgan, L. (2012) Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors, *Bioinformatics*, **28**, 331-341.

Chen, Y.C., Wright, J.D. & Lim, C. (2012) DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry, *Nucleic Acids Research*, **40**, W249-256.

Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N. & Dunker, A.K. (2006) Rational drug design via intrinsically disordered protein, *Trends in Biotechnology*, **24**, 435-442.

Cheng, Y., Oldfield, C.J., Meng, J., Romero, P., Uversky, V.N. & Dunker, A.K. (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments, *Biochemistry*, **46**, 13468-13477.

Choy, W.Y., Mulder, F.A., Crowhurst, K.A., Muhandiram, D.R., Millett, I.S., Doniach, S., Forman-Kay, J.D. & Kay, L.E. (2002) Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques, *Journal of Molecular Biology*, **316**, 101-112.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. & Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life, *Science*, **311**, 1283-1287.

Cirillo, D., Agostini, F. & Tartaglia, G.G. (2013) Predictions of protein–RNA interactions, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **3**, 161-175.

Consortium, U. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, **40**, D71-D75.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T.R. (2011) RBPDB: a database of RNA-binding specificities, *Nucleic Acids Research*, **39**, D301-308.

Cortese, M.S., Uversky, V.N. & Dunker, A.K. (2008) Intrinsic disorder in scaffold proteins: getting more from less, *Progress in Biophysics & Molecular Biology*, **98**, 85-106.

Dabbs, E.R. (1978) Mutational alterations in 50 proteins of the Escherichia coli ribosome, *Molecular and General Genetics*, **165**, 73-78.

Dabbs, E.R. (1986) Mutant studies on the prokaryotic ribosome. In Hardesty, B. and Kramer, G. (eds), *Structure, Function and Genetics of Ribosomes*. Springer-Verlag, New York, pp. 733-748.

Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H. & Buchner, J. (2003) The N-terminal domain of p53 is natively unfolded, *Journal of Molecular Biology*, **332**, 1131-1141.

Declercq, W., Van Herreweghe, F., Vanden Berghe, T. & Vandenabeele, P. (2009) Death receptor-induced necroptosis, *Encyclopedia of Life Sciences*.

Demarest, S.J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H.J., Evans, R.M. & Wright, P.E. (2002) Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators, *Nature*, **415**, 549-553.

Diez, J., Walter, D., Munoz-Pinedo, C. & Gabaldon, T. (2010) DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death, *Cell Death & Differentiation*, **17**, 735-736.

Disfani, F.M., Hsu, W.L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N. & Kurgan, L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, *Bioinformatics*, **28**, i75-83.

Dixon, S.E., Bhatti, M.M., Uversky, V.N., Dunker, A.K. & Sullivan, W.J. (2011) Regions of intrinsic disorder help identify a novel nuclear localization signal in Toxoplasma gondii histone acetyltransferase TgGCN5-B, *Molecular and Biochemical Parasitology*, **175**, 192-195.

Doolittle, R.F. (1973) Structural aspects of the fibrinogen to fibrin conversion, *Advances in Protein Chemistry*, **27**, 1-109.

Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I. & Tompa, P. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution, *Journal of Proteome Research*, **5**, 2985-2995.

Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, **21**, 3433-3434.

Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins, *Journal of Molecular Biology*, **347**, 827-839.

Dosztanyi, Z., Meszaros, B. & Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins, *Bioinformatics*, **25**, 2745-2746.

Dosztanyi, Z., Meszaros, B. & Simon, I. (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, *Briefings in Bioinformatics*, **11**, 225-243.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K., Jones, D.T., Longhi, S., Metallo, S.J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R.V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J.L., Tompa, P. & Uversky, V.N. (2013) What’s in a name? Why these proteins are intrinsically disordered, *Intrinsically Disordered Proteins*, **1**, 0-4.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. & Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry*, **41**, 6573-6582.

Dunker, A.K., Brown, C.J. & Obradovic, Z. (2002) Identification and functions of usefully disordered proteins, *Advances in Protein Chemistry*, **62**, 25-49.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M. & Uversky, V.N. (2005) Flexible nets: The roles of intrinsic disorder in protein interaction networks, *FEBS Journal*, **272**, 5129-5148.

Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C. & Villafranca, J.E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Pacific Symposium on Biocomputing*, 473-484.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C. & Obradovic, Z. (2001) Intrinsically disordered protein, *Journal of Molecular Graphics and Modelling*, **19**, 26-59.

Dunker, A.K. & Obradovic, Z. (2001) The protein trinity–linking function and disorder, *Nature Biotechnology*, **19**, 805-806.

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C. & Brown, C.J. (2000) Intrinsic protein disorder in complete genomes, *Genome Inform Ser Workshop Genome Inform*, **11**, 161-171.

Dunker, A.K., Obradovic, Z., Romero, P., Kissinger, C. & Villafranca, E. (1997) On the importance of being disordered., *PDB Newsletter*, **81**, 3-5.

Dunker, A.K., Silman, I., Uversky, V.N. & Sussman, J.L. (2008) Function and structure of inherently disordered proteins, *Current Opinion in Structural Biology*, **18**, 756-764.

Dunker, A.K. & Uversky, V.N. (2008) Signal transduction via unstructured protein conduits, *Nature Chemical Biology*, **4**, 229-230.

Dyson, H.J. & Wright, P.E. (1998) Equilibrium NMR studies of unfolded and partially folded proteins, *Nature Biotechnology*, **5 Suppl**, 499-503.

Dyson, H.J. & Wright, P.E. (2002) Coupling of folding and binding for unstructured proteins, *Current Opinion in Structural Biology*, **12**, 54-60.

Dyson, H.J. & Wright, P.E. (2005) Intrinsically unstructured proteins and their functions, *Nature Reviews Molecular Cell Biology*, **6**, 197-208.

Ekman, D., Light, S., Bjorklund, A.K. & Elofsson, A. (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?, *Genome Biology*, **7**, R45.

Eng, F.J. & Warner, J.R. (1991) Structural basis for the regulation of splicing of a yeast messenger RNA, *Cell*, **65**, 797-804.

Esposito, G., Fogolari, F., Damante, G., Formisano, S., Tell, G., Leonardi, A., Di Lauro, R. & Viglino, P. (1996) Analysis of the solution structure of the homeodomain of rat thyroid transcription factor 1 by 1H-NMR spectroscopy and restrained molecular mechanics, *European Journal of Biochemistry*, **241**, 101-113.

Feng, Z.P., Zhang, X., Han, P., Arora, N., Anders, R.F. & Norton, R.S. (2006) Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes, *Molecular and Biochemical Parasitology*, **150**, 256-267.

Fewell, S.W. & Woolford, J.L. (1999) Ribosomal protein S14 of Saccharomyces cerevisiae regulates its expression by binding to RPS14B pre-mRNA and to 18S rRNA, *Molecular and Cellular Biology*, **19**, 826-834.

Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme, *Berichte der deutschen chemischen Gesellschaft*, **27**, 2985-2993.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. (2002) Evolutionary rate in the protein interaction network, *Science*, **296**, 750-752.

Freedman, L.P., Zengel, J.M., Archer, R.H. & Lindahl, L. (1987) Autogenous control of the S10 ribosomal protein operon of Escherichia coli: genetic dissection of transcriptional and posttranscriptional regulation, *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 6516-6520.

Galea, C.A., High, A.A., Obenauer, J.C., Mishra, A., Park, C.G., Punta, M., Schlessinger, A., Ma, J., Rost, B., Slaughter, C.A. & Kriwacki, R.W. (2009) Large-scale analysis of thermostable, mammalian proteins provides insights into the intrinsically disordered proteome, *Journal of Proteome Research*, **8**, 211-226.

Galluzzi, L., Vanden Berghe, T., Vanlangenakker, N., Buettner, S., Eisenberg, T., Vandenabeele, P., Madeo, F. & Kroemer, G. (2011) Programmed necrosis from molecules to health and disease, *International Review of Cell and Molecular Biology*, **289**, 1-35.

Garrett, R.A. (1983) Structure and role of eubacterial ribosomal proteins, *Horizons in biochemistry and biophysics*, **7**, 101-138.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W. & Bryant, S.H. (2010) The NCBI BioSystems database, *Nucleic Acids Research*, **38**, D492-496.

Gilman, A.G. (1987) G proteins: transducers of receptor-generated signals, *Annual Review of Biochemistry*, **56**, 615-649.

Gunasekaran, K., Tsai, C.J. & Nussinov, R. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers, *Journal of Molecular Biology*, **341**, 1327-1341.

Gutteridge, A. & Thornton, J.M. (2005) Understanding nature's catalytic toolkit, *Trends in Biochemical Sciences*, **30**, 622-629.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I.H. (2009) The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter*, **11**, 10-18.

Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F. & Yonath, A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium, *Cell*, **107**, 679-688.

Haynes, C. & Iakoucheva, L.M. (2006) Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins, *Nucleic Acids Research*, **34**, 305-312.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M. & Iakoucheva, L.M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes, *PLOS Computational Biology*, **2**, e100.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N. & Dunker, A.K. (2009) Predicting intrinsic disorder in proteins: an overview, *Cell Research*, **19**, 929-949.

Hinds, M.G. & Day, C.L. (2005) Regulation of apoptosis: uncovering the binding determinants, *Current Opinion in Structural Biology*, **15**, 690-699.

Hinds, M.G., Smits, C., Fredericks-Short, R., Risk, J.M., Bailey, M., Huang, D.C. & Day, C.L. (2007) Bim, Bad and Bmf: intrinsically unstructured BH3-only proteins that undergo a localized conformational change upon binding to prosurvival Bcl-2 targets, *Cell Death & Differentiation*, **14**, 128-136.

Howard, J. & Hyman, A.A. (2007) Microtubule polymerases and depolymerases, *Current Opinion in Cell Biology*, **19**, 31-35.

Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, **26**, 680-682.

Huber, R. (1987) Flexibility and rigidity, requirements for the function of proteins and protein pigment complexes. Eleventh Keilin memorial lecture, *Biochemical Society Transactions*, **15**, 1009-1020.

Huber, R. & Bennett, W.S. (1983) Functional significance of flexibility in proteins, *Biopolymers*, **22**, 261-279.

Hwang, S., Gou, Z. & Kuznetsov, I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, *Bioinformatics*, **23**, 634-636.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. & Dunker, A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *Journal of Molecular Biology*, **323**, 573-584.

Ishida, T. & Kinoshita, K. (2008) Prediction of disordered regions in proteins based on the meta approach, *Bioinformatics*, **24**, 1344-1348.

James, L.C., Roversi, P. & Tawfik, D.S. (2003) Antibody multispecificity mediated by conformational diversity, *Science*, **299**, 1362-1367.

Jirgenesons, B. (1966) Classification of proteins according to conformation, *Makromolekulare Chemie*, **91**, 74-86.

Johansson, F. & Toh, H. (2010) A comparative study of conservation and variation scores, *BMC Bioinformatics*, **11**, 388.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, **292**, 195-202.

Jones, D.T. & Swindells, M.B. (2002) Getting the most from PSI-BLAST, *Trends in Biochemical Sciences*, **27**, 161-164.

Karush, F. (1950) Heterogeneity of the binding sites of bovine serum albumin., *Journal of the American Chemical Society*, **72**, 2705-2713.

Kauffman, C. & Karypis, G. (2012) Computational tools for protein-DNA interactions, *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, **2**, 14-28.

Kawashima, S. & Kanehisa, M. (2000) AAindex: amino acid index database, *Nucleic Acids Research*, **28**, 374.

Klein, D.J., Moore, P.B. & Steitz, T.A. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit, *Journal of Molecular Biology*, **340**, 141-177.

Kohavi, R. & John, G.H. (1997) Wrappers for feature subset selection, *Artificial Intelligence*, **97**, 273-324.

Kuznetsov, I.B., Gou, Z.K., Li, R. & Hwang, S.W. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins, *Proteins: Structure, Function, and Bioinformatics*, **64**, 19-27.

Landsteiner, K. (1936) *The specificity of serological reactions*. Dover, New York.

Lee, H., Mok, K.H., Muhandiram, R., Park, K.H., Suk, J.E., Kim, D.H., Chang, J., Sung, Y.C., Choi, K.Y. & Han, K.H. (2000) Local structural elements in the mostly unstructured transcriptional activation domain of human p53, *Journal of Biological Chemistry*, **275**, 29426-29432.

Li, B.Q., Hu, L.L., Chen, L., Feng, K.Y., Cai, Y.D. & Chou, K.C. (2012) Prediction of protein domain with mRMR feature selection and analysis, *PLoS One*, **7**, e39308.

Li, M. & Song, J. (2007) The N- and C-termini of the human Nogo molecules are intrinsically unstructured: bioinformatics, CD, NMR characterization, and functional implications, *Proteins: Structure, Function, and Bioinformatics*, **68**, 100-108.

Li, X., Lindahl, L. & Zengel, J.M. (1996) Ribosomal protein L4 from Escherichia coli utilizes nonidentical determinants for its structural and regulatory functions, *RNA*, **2**, 24-37.

Libich, D.S., Schwalbe, M., Kate, S., Venugopal, H., Claridge, J.K., Edwards, P.J., Dutta, K. & Pascal, S.M. (2009) Intrinsic disorder and coiled-coil formation in prostate apoptosis response factor 4, *FEBS Journal*, **276**, 3710-3728.

Lin, N., Wu, B., Jansen, R., Gerstein, M. & Zhao, H. (2004) Information assessment on predicting protein-protein interactions, *BMC Bioinformatics*, **5**, 154-164.

Linding, R., Russell, R.B., Neduva, V. & Gibson, T.J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder, *Nucleic Acids Research*, **31**, 3701-3708.

Lindstrom, M.S. (2009) Emerging functions of ribosomal proteins in gene-specific transcription and translation, *Biochemical and Biophysical Research Communications*, **379**, 167-170.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N. & Dunker, A.K. (2006) Intrinsic disorder in transcription factors, *Biochemistry*, **45**, 6873-6888.

Liu, J., Tan, H. & Rost, B. (2002) Loopy proteins appear conserved in evolution, *Journal of Molecular Biology*, **322**, 53-64.

Lobley, A., Swindells, M.B., Orengo, C.A. & Jones, D.T. (2007) Inferring function using patterns of native disorder in proteins, *PLOS Computational Biology*, **3**, e162.

Ma, X., Guo, J., Liu, H.D., Xie, J.M. & Sun, X. (2012) Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**, 1766-1775.

Malygin, A.A., Parakhnevitch, N.M., Ivanov, A.V., Eperon, I.C. & Karpova, G.G. (2007) Human ribosomal protein S13 regulates expression of its own gene at the splicing step by a feedback mechanism, *Nucleic Acids Research*, **35**, 6414-6423.

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

McGuffin, L.J., Bryson, K. & Jones, D.T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics*, **16**, 404-405.

McMeekin, T.L. (1952) Milk proteins, *Journal of Milk and Food Technology*, **15**, 57-63.

Meszaros, B., Simon, I. & Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins, *PLOS Computational Biology*, **5**, e1000376.

Minezaki, Y., Homma, K., Kinjo, A.R. & Nishikawa, K. (2006) Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation, *Journal of Molecular Biology*, **359**, 1137-1149.

Mirsky, A.E. & Pauling, L. (1936) On the Structure of Native, Denatured, and Coagulated Proteins, *Proceedings of the National Academy of Sciences of the United States of America*, **22**, 439-447.

Mitrovich, Q.M. & Anderson, P. (2000) Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans, *Genes Dev*, **14**, 2173-2184.

Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M. & Kurgan, L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, *Bioinformatics*, **26**, i489-496.

Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K. & Uversky, V.N. (2006) Analysis of molecular recognition features (MoRFs), *Journal of Molecular Biology*, **362**, 1043-1059.

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A. & Kryshtafovych, A. (2011) Evaluation of disorder predictions in CASP9, *Proteins: Structure, Function, and Bioinformatics*, **79 Suppl 10**, 107-118.

Muchmore, S.W., Sattler, M., Liang, H., Meadows, R.P., Harlan, J.E., Yoon, H.S., Nettesheim, D., Chang, B.S., Thompson, C.B., Wong, S.L., Ng, S.L. & Fesik, S.W. (1996) X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death, *Nature*, **381**, 335-341.

Nakai, K., Kidera, A. & Kanehisa, M. (1988) Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Engineering*, **2**, 93-100.

Nakao, A., Yoshihama, M. & Kenmochi, N. (2004) RPG: the Ribosomal Protein Gene database, *Nucleic Acids Research*, **32**, D168-170.

Nierhaus, K.H. (1991) The assembly of prokaryotic ribosomes, *Biochimie*, **73**, 739-755.

Noivirt-Brik, O., Prilusky, J. & Sussman, J.L. (2009) Assessment of disorder predictions in CASP8, *Proteins: Structure, Function, and Bioinformatics*, **77 Suppl 9**, 210-216.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J. & Dunker, A.K. (2003) Predicting intrinsic disorder from amino acid sequence, *Proteins: Structure, Function, and Bioinformatics*, **53 Suppl 6**, 566-572.

Oldfield, C.J., Cheng, Y., Cortese, M.S., Romero, P., Uversky, V.N. & Dunker, A.K. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements, *Biochemistry*, **44**, 12454-12470.

Oldfield, C.J., Meng, J., Yang, J.Y., Yang, M.Q., Uversky, V.N. & Dunker, A.K. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, *BMC Genomics*, **9 Suppl 1**, S1.

Ouyang, L., Shi, Z., Zhao, S., Wang, F.T., Zhou, T.T., Liu, B. & Bao, J.K. (2012) Programmed cell death pathways in cancer: a review of apoptosis, autophagy and programmed necrosis, *Cell Proliferation*, **45**, 487-498.

Parakhnevich, N.M., Ivanov, A.V., Malygin, A.A. & Karpova, G.G. (2007) Human ribosomal protein S13 inhibits splicing of the own pre-mRNA, *Molekuliarnaia Biologiia (Moskva)*, **41**, 51-58.

Patil, A. & Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks, *FEBS Letter*, **580**, 2041-2045.

Pauling, L. (1940) A theory of the structure and process of formation of antibodies, *Journal of the American Chemical Society*, **62**, 2643-2657.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K. & Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder, *BMC Bioinformatics*, **7**, 208.

Peng, Z. & Kurgan, L. (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions, *Current Protein and Peptide Science*, **13**, 6-18.

Peng, Z. & Kurgan, L. (2012) On the complementarity of the consensus-based disorder prediction, *Pacific Symposium on Biocomputing*, 176-187.

Peng, Z., Mizianty, M.J. & Kurgan, L. (2014) Genome-scale prediction of proteins with long intrinsically disordered regions, *Proteins: Structure, Function, and Bioinformatics*, **82**, 145-158.

Peng, Z., Mizianty, M.J., Xue, B., Kurgan, L. & Uversky, V.N. (2012) More than just tails: intrinsic disorder in histone proteins, *Molecular Biosystems*, **8**, 1886-1901.

Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L. & Uversky, V.N. (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome, *Cellular and Molecular Life Sciences*, **71**, 1477-1504.

Peng, Z., Xue, B., Kurgan, L. & Uversky, V.N. (2013) Resilience of death: intrinsic disorder in proteins involved in the programmed cell death, *Cell Death & Differentiation*, **20**, 1257-1267.

Pentony, M.M. & Jones, D.T. (2010) Modularity of intrinsic disorder in the human proteome, *Proteins: Structure, Function, and Bioinformatics*, **78**, 212-221.

Presutti, C., Ciafre, S.A. & Bozzoni, I. (1991) The ribosomal protein L2 in S. cerevisiae controls the level of accumulation of its own mRNA, *EMBO Journal*, **10**, 2215-2221.

Pruitt, K.D., Tatusova, T. & Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, **35**, D61-65.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K. & Bujnicki, J.M. (2012) Computational methods for prediction of protein-RNA interactions, *Journal of Structural Biology*, **179**, 261-268.

Radivojac, P., Iakoucheva, L.M., Oldfield, C.J., Obradovic, Z., Uversky, V.N. & Dunker, A.K. (2007) Intrinsic disorder and functional proteomics, *Biophysical Journal*, **92**, 1439-1456.

Rautureau, G.J., Day, C.L. & Hinds, M.G. (2010) Intrinsically disordered proteins in bcl-2 regulated apoptosis, *International Journal of Molecular Sciences*, **11**, 1808-1824.

Reingewertz, T.H., Shalev, D.E., Sukenik, S., Blatt, O., Rotem-Bamberger, S., Lebendiker, M., Larisch, S. & Friedler, A. (2011) Mechanism of the interaction between the intrinsically disordered C-terminus of the pro-apoptotic ARTS protein and the Bir3 domain of XIAP, *PLoS One*, **6**, e24655.

Rodgers, J.L. & Nicewander, W.A. (1988) 13 Ways to Look at the Correlation-Coefficient, *American Statistician*, **42**, 59-66.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. & Dunker, A.K. (2001) Sequence complexity of disordered protein, *Proteins: Structure, Function, and Bioinformatics*, **42**, 38-48.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z. & Dunker, A.K. (2006) Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8390-8395.

Saeys, Y., Inza, I. & Larranaga, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23**, 2507-2517.

Sattler, M., Liang, H., Nettesheim, D., Meadows, R.P., Harlan, J.E., Eberstadt, M., Yoon, H.S., Shuker, S.B., Chang, B.S., Minn, A.J., Thompson, C.B. & Fesik, S.W. (1997) Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis, *Science*, **275**, 983-986.

Schlessinger, A., Liu, J. & Rost, B. (2007) Natively unstructured loops differ from other loops, *PLOS Computational Biology*, **3**, e140.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L. & Rost, B. (2009) Improved disorder prediction by combination of orthogonal approaches, *PLoS One*, **4**, e4433.

Schlessinger, A., Yachdav, G. & Rost, B. (2006) PROFbval: predict flexible and rigid residues in proteins, *Bioinformatics*, **22**, 891-893.

Schuwirth, B.S., Borovinskaya, M.A., Hau, C.W., Zhang, W., Vila-Sanjurjo, A., Holton, J.M. & Cate, J.H. (2005) Structures of the bacterial ribosome at 3.5 A resolution, *Science*, **310**, 827-834.

Selmer, M., Dunham, C.M., Murphy, F.V.t., Weixlbaumer, A., Petry, S., Kelley, A.C., Weir, J.R. & Ramakrishnan, V. (2006) Structure of the 70S ribosome complexed with mRNA and tRNA, *Science*, **313**, 1935-1942.

Semrad, K., Green, R. & Schroeder, R. (2004) RNA chaperone activity of large ribosomal subunit proteins from Escherichia coli, *RNA*, **10**, 1855-1860.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. & Jiang, H. (2007) Predicting protein-protein interactions based only on sequences information, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4337-4341.

Shojania, S. & O'Neil, J.D. (2011) Intrinsic disorder and function of the HIV-1 Tat protein, *Protein & Peptide Letters*, **17**, 999-1011.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z. & Dunker, A.K. (2007) DisProt: the Database of Disordered Proteins, *Nucleic Acids Research*, **35**, D786-793.

Singh, D., Chang, S.J., Lin, P.H., Averina, O.V., Kaberdin, V.R. & Lin-Chao, S. (2009) Regulation of ribonuclease E activity by the L4 ribosomal protein of Escherichia coli, *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 864-869.

Singh, G.P. & Dash, D. (2007) Intrinsic disorder in yeast transcriptional regulatory network, *Proteins: Structure, Function, and Bioinformatics*, **68**, 602-605.

Singh, G.P., Ganapathi, M. & Dash, D. (2007) Role of intrinsic disorder in transient interactions of hub proteins, *Proteins: Structure, Function, and Bioinformatics*, **66**, 761-765.

Singh, G.P., Ganapathi, M., Sandhu, K.S. & Dash, D. (2006) Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes, *Proteins: Structure, Function, and Bioinformatics*, **62**, 309-315.

Smith, T.F. & Waterman, M.S. (1981) Identification of common molecular subsequences, *Journal of Molecular Biology*, **147**, 195-197.

Stigler, S.M. (1989) Francis Galton's Account of the Invention of Correlation, *Statistical Science*, **4**, 73-79.

Tan, M.L., Ooi, J.P., Ismail, N., Moad, A.I. & Muhammad, T.S. (2009) Programmed cell death pathways and current antitumor targets, *Pharmaceutical Research*, **26**, 1547-1560.

Tate, R.F. (1954) Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation, *The Annals of Mathematical Statistics*, **25**, 603-607.

Timsit, Y., Acosta, Z., Allemand, F., Chiaruttini, C. & Springer, M. (2009) The role of disordered ribosomal protein extensions in the early steps of eubacterial 50 S ribosomal subunit assembly, *International Journal of Molecular Sciences*, **10**, 817-834.

Tomii, K. & Kanehisa, M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Engineering*, **9**, 27-36.

Tompa, P. (2002) Intrinsically unstructured proteins, *Trends in Biochemical Sciences*, **27**, 527-533.

Tompa, P. & Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones, *FASEB Journal*, **18**, 1169-1175.

Tompa, P., Dosztanyi, Z. & Simon, I. (2006) Prevalent structural disorder in E. coli and S. cerevisiae proteomes, *Journal of Proteome Research*, **5**, 1996-2000.

Tompa, P., Fuxreiter, M., Oldfield, C.J., Simon, I., Dunker, A.K. & Uversky, V.N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins, *Bioessays*, **31**, 328-335.

Tompa, P., Szasz, C. & Buday, L. (2005) Structural disorder throws new light on moonlighting, *Trends in Biochemical Sciences*, **30**, 484-489.

Turoverov, K.K., Kuznetsova, I.M. & Uversky, V.N. (2010) The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation, *Progress in Biophysics and Molecular Biology*, **102**, 73-84.

Uversky, V.N. (2003) Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?, *Cellular and Molecular Life Sciences*, **60**, 1852-1871.

Uversky, V.N. (2011) Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes, *Chemical Society Reviews*, **40**, 1623-1634.

Uversky, V.N. & Dunker, A.K. (2010) Understanding protein non-folding, *Biochimica et Biophysica Acta - Proteins and Proteomics*, **1804**, 1231-1264.

Uversky, V.N., Gillespie, J.R. & Fink, A.L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins: Structure, Function, and Bioinformatics*, **41**, 415-427.

Uversky, V.N., Oldfield, C.J. & Dunker, A.K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, *Journal of Molecular Recognition*, **18**, 343-384.

Uversky, V.N., Oldfield, C.J. & Dunker, A.K. (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept, *Annual Review of Biophysics*, **37**, 215-246.

Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z. & Dunker, A.K. (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases, *BMC Genomics*, **10 Suppl 1**, S7.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N. & Dunker, A.K. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners, *Journal of Proteome Research*, **6**, 2351-2366.

Vacic, V., Uversky, V.N., Dunker, A.K. & Lonardi, S. (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences, *BMC Bioinformatics*, **8**, 211-217.

Vandenabeele, P., Galluzzi, L., Vanden Berghe, T. & Kroemer, G. (2010) Molecular mechanisms of necroptosis: an ordered cellular explosion, *Nature Reviews Molecular Cell Biology*, **11**, 700-714.

Vucetic, S., Xie, H., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z. & Uversky, V.N. (2007) Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions, *Journal of Proteome Research*, **6**, 1899-1916.

Walia, R.R., Caragea, C., Lewis, B.A., Towfic, F., Terribilini, M., El-Manzalawy, Y., Dobbs, D. & Honavar, V. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art, *BMC Bioinformatics*, **13**, 89.

Walsh, I., Martin, A.J., Di Domenico, T. & Tosatto, S.C. (2012) ESpritz: accurate and fast prediction of protein disorder, *Bioinformatics*, **28**, 503-509.

Walsh, I., Martin, A.J., Di Domenico, T., Vullo, A., Pollastri, G. & Tosatto, S.C. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs, *Nucleic Acids Research*, **39**, W190-196.

Wang, K. & Samudrala, R. (2006) Incorporating background frequency improves entropy-based residue conservation measures, *BMC Bioinformatics*, **7**, 385.

Wang, L., Huang, C., Yang, M. & Yang, J. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, *BMC Syst Biol*, **4**, S3.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. & Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *Journal of Molecular Biology*, **337**, 635-645.

Warner, J.R. & McIntosh, K.B. (2009) How common are extraribosomal functions of ribosomal proteins?, *Molecular Cell*, **34**, 3-11.

Weisberg, R.A. (2008) Transcription by moonlight: structural basis of an extraribosomal activity of ribosomal protein S10, *Molecular Cell*, **32**, 747-748.

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods, *Biometrics Bulletin*, **1**, 80-83.

Williams, R.J. (1978) The conformational mobility of proteins and its functional significance, *Biochemical Society Transactions*, **6**, 1123-1126.

Williams, R.M., Obradovi, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J. & Dunker, A.K. (2001) The protein non-folding problem: amino acid determinants of intrinsic order and disorder, *Pacific Symposium on Biocomputing*, 89-100.

Wilson, D.N. & Nierhaus, K.H. (2005) Ribosomal proteins in the spotlight, *Critical Reviews in Biochemistry and Molecular Biology*, **40**, 243-267.

Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit, *Nature*, **407**, 327-339.

Wool, I.G. (1996) Extraribosomal functions of ribosomal proteins, *Trends in Biochemical Sciences*, **21**, 164-165.

Wootton, J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures, *Computers & Chemistry*, **18**, 269-285.

Wootton, J.C. (1994) Sequences with ‘unusual’ amino acid compositions, *Current Opinion in Structural Biology*, **4**, 413-421.

Wootton, J.C. & Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases, *Computers & Chemistry*, **17**, 149-163.

Wootton, J.C. & Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases, *Methods in Enzymology*, **266**, 554-571.

Wright, P.E. & Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *Journal of Molecular Biology*, **293**, 321-331.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z. & Uversky, V.N. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins, *Journal of Proteome Research*, **6**, 1917-1932.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N. & Obradovic, Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, *Journal of Proteome Research*, **6**, 1882-1898.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K. & Uversky, V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, *Biochimica et Biophysica Acta*, **1804**, 996-1010.

Xue, B., Dunker, A.K. & Uversky, V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, *Journal of biomolecular structure & dynamics*, **30**, 137-149.

Xue, B., Mizianty, M.J., Kurgan, L. & Uversky, V.N. (2012) Protein intrinsic disorder as a flexible armor and a weapon of HIV-1, *Cellular and Molecular Life Sciences*, **69**, 1211-1259.

Xue, B., Williams, R.W., Oldfield, C.J., Dunker, A.K. & Uversky, V.N. (2010) Archaic chaos: intrinsically disordered proteins in Archaea, *BMC Systems Biology*, **4 Suppl 1**, S1.

Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., Cate, J.H. & Noller, H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution, *Science*, **292**, 883-896.

Zeng, Y., He, Y., Yang, F., Mooney, S.M., Getzenberg, R.H., Orban, J. & Kulkarni, P. (2011) The cancer/testis antigen prostate-associated gene 4 (PAGE4) is a highly intrinsically disordered protein, *The Journal of Biological Chemistry*, **286**, 13985-13994.

Zengel, J.M. & Lindahl, L. (1990) Escherichia coli ribosomal protein L4 stimulates transcription termination at a specific site in the leader of the S10 operon independent of L4-mediated inhibition of translation, *Journal of Molecular Biology*, **213**, 67-78.

Zengel, J.M. & Lindahl, L. (1990) Ribosomal protein L4 stimulates in vitro termination of transcription at a NusA-dependent terminator in the S10 operon leader, *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2675-2679.

Zengel, J.M. & Lindahl, L. (1994) Diverse mechanisms for regulating ribosomal protein synthesis in Escherichia coli, *Progress in Nucleic Acid Research and Molecular Biology*, **47**, 331-370.

Zhang, H.M., Chen, H., Liu, W., Liu, H., Gong, J., Wang, H. & Guo, A.Y. (2012) AnimalTFDB: a comprehensive animal transcription factor database, *Nucleic Acids Research*, **40**, D144-149.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A. & Honig, B. (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale, *Nature*, **490**, 556-560.

Appendix A
List of Ribosomal Proteins with IDRs and On-Ribosome Functions

Ribosomal proteins with long IDRs. The illustrative examples of yeast ribosomal proteins with IDRs are listed in order of the increasing length of their IDRs (for proteins with two IDRs, the longer region is taken into consideration): L16 (residues 103-110), L30e (residues 1-8), L34e (residues 114-121), L11 (residues 1-8 and 156-165), S10e (residues 97-105), S11 (residues 1-10), S17e (residues 90-94 and 127-136), S8e (residues 124-134), S4 (residues 187-197), S19 (residues 1-6 and 132-142), S10 (residues 1-14), S3 (residues 226-240), S6e (residues 219-236), L6e (residues 110-128), S7 (residues 1-19), S12e (residues 1-19), L23 (residues 1-21), S26e (residues 99-119), L22 (residues 157-184), S5 (residues 1-33), S25e (residues 1-36), L24e (residues 99-135), S2 (residues 208-257), P2 (residues 52-110), P1 (residues 47-106), L40e (residues 1-76), S31e (residues 1-81), P0 (residues 108-181 and 222-312), S1e (residues 1-199 and 234-255), and Stm1 (residues 1-24 and 180-273).

Some on-ribosome functions of the ribosomal proteins. S1 is involved in the delivery of the mRNA into the proximity of the ribosome during initiation and also is responsible for the translational feedback regulation of S1 operon. S3, S4, and S5 form the mRNA entry pore and may have a helicase activity during translation to unwind mRNA secondary structure. S4, S5 and S12 are involved in the decoding and fidelity of translation, with S5 facilitating changes of rRNA conformations that alters the selection mode of the ribosome from accurate to error prone and vice versa, and with S12 participating in decoding of the second and third codon positions at the A site of the ribosome. L1 and L16/L27 are involved in the release and binding of tRNAs to the ribosome, respectively. L1 is also responsible for the translational feedback regulation of the L11 operon, whereas proteins L7/L12 (which are involved in the elongation-factor binding and GTPase activation) together with L10 are involved in translational feedback regulation of the L10 operon. L4 and L22 that protrude into the ribosomal tunnel confer resistance to macrolide antibiotics. In addition, L22 may also interact with nascent chains to control translation of particular proteins, whereas L4 may play a role in rRNA transcription antitermination. L9 influences tRNA stability at the P site, regulates the mRNA movement, and controls the efficiency of the translational bypassing. L11 and L10 × $(L7/L12)_4$ proteins, that are located at the stalk region, are involved in binding of the elongation factors EF-G and EF-Tu to the ribosome. L16 is involved in controlling the correct position of the acceptor stem of A- and P-site tRNAs and also in correct positioning of the ribosome recycling factor (RRF) on the ribosome. L17, L22, L23, L24, L29, and L32 are known to form a ring around the tunnel exit site of the ribosome, with two members of this ring, with L22 being able to interact with specific nascent chains to regulate translation, and with L23 and L29 being involved binding of the signal recognition.

Appendix B

Functional Annotations for Intrinsic Disorder

Table S1. Functional annotations of IDRs extracted from DisProt

List of 26 functional annotations of IDRs extracted from DisProt of release v5.9, which were utilized to investigate the functional roles of intrinsic disorder in ribosomal and programmed cell death (PCD) proteins. The “protein-RNA/-DNA binding” and “modification sites” combine several sub-functions, were considered. The third and fourth columns indicate if a given function is considered to be compared between short and long disordered regions extracted from RPG_3411 (ribosomal proteins) and Deathbase (PCD-related proteins), respectively. The last column shows whether a given function is included to investigate the relations between intrinsic disorder and different PCD processes. "Y"/"N" denotes yes/no.

Function	Sub-function	RPG_3411	Deathbase	PCD processes
Protein-DNA binding	Protein-DNA binding	Y	Y	Y
	DNA bending			
	DNA unwinding			
Protein-RNA binding	Protein-rRNA binding	Y	Y	Y
	Protein-tRNA binding			
	Protein-genomic RNA binding			
	Protein-mRNA binding			
	Protein-RNA binding			
Post-translational modification site	Phosphorylation	Y	Y	Y
	Acetylation			
	Fatty acylation			
	Glycosylation			
	Methylation			
Apoptosis Regulation		Y	Y	Y
Autoregulatory		Y	Y	Y
Cofactor/heme binding		N	Y	Y
Electron transfer		Y	Y	Y
Entropic bristle		N	Y	N
Entropic clock		N	N	N
Entropic spring		N	N	N
Flexible linkers/spacers		Y	Y	Y
Intra-protein interaction		Y	Y	Y
Metal binding		Y	Y	Y
Nuclear localization		N	Y	N
Polymerization		Y	Y	N
Protein-protein binding		Y	Y	Y
Protein inhibitor		N	Y	N
Protein-lipid interaction		N	Y	Y
Protein detergent		N	N	N
Protein-Biocrystal binding		N	N	N
Regulation of proteolysis in vivo		N	Y	N
Substrate/ligand binding		Y	Y	Y
Self-transport through channel		N	N	N
Sulfation		N	N	N
Structural mortar		N	N	N
Transactivation		Y	Y	Y

Appendix C

Benchmark Datasets for DisoRDPbind

Table S2. Summary of large scale datasets to evaluate DisoRDPbind at the proteomic level
Summary of the datasets extracted from the four considered complete genomes/proteomes: *H. sapiens*, *M. musculus*, *C. elegans* and *D. melanogaster*. The table includes the number of proteins and the average disorder content (i.e., Fraction of disordered residues) for the proteins sets in the GO_RNA, GO_DNA, RBPDB, animalTFDB and mentha datasets, where GO_RNA and RNPDB include RNA-binding proteins; GO_DNA and animalTFDB include DNA-binding proteins; and mentha is the latest integrated resources of protein-protein interaction (PPI) networks. The predRNA/predDNA/predProtein_UniProt denotes the set of disordered DNA/RNA/protein binding proteins from UniProt that is predicted with at least one disordered DNA/RNA/protein-binding region (>=4 consecutive AAs) by DisoRDPbind. The predProtein_mentha represents the predicted disordered protein-binding proteins from the mentha database. The disorder is predicted by the majority vote of ESPRITZ and IUPred, by following the procedures described in section 3.2.

Species (taxID)	*H. sapiens*	*M. musculus*	*C. elegans*	*D. melanogaster*
	(9606)	(10090)	(6239)	(7227)
# proteins collected from UniProt	42426	33181	25159	19656
Average disorder content in UniProt	0.24	0.21	0.17	0.23
# proteins in GO_RNA	1209	1101	420	568
Average disorder content in GO_RNA	0.28	0.28	0.23	0.29
# proteins in RBPDB	398	339	204	73
Average disorder content in RBPDB	0.37	0.37	0.32	0.39
# proteins in predRNA	2769	1401	722	792
Average disorder content in predRNA	0.37	0.34	0.32	0.35
# proteins in GO_DNA	3153	2686	1074	967
Average disorder content in GO_DNA	0.32	0.37	0.23	0.41
# proteins in animalTFDB	1464	1375	654	596
Average disorder content in animalTFDB	0.32	0.34	0.22	0.41
# proteins in predDNA	2475	2231	1241	1140
Average disorder content in predDNA	0.31	0.31	0.30	0.40
# proteins in mentha	14547	8006	5005	8096
Average number of interactors in mentha	21.4	6.7	5.2	7.3
# proteins in predProtein_UniProt	36150	28243	19683	17439
# proteins in predProtein_mentha	13525	7559	4553	7431

Table S3. Summary of benchmark datasets including Training, TEST115 and TEST36.
The RNA, DNA and protein binding mediated by intrinsic disorder (2nd column) are defined by combining several functional subclasses listed in the 3rd column. The "Others" row (given in *italic*) includes all other functional subclasses that are not included in the TRAINING, TEST115 and TEST36 datasets. The source data was taken from DisProt of release 6.01. The 4th and 5th column list the number of IDRs and disordered amino acids (AAs), respectively, which are annotated with a given function/functional subclass. We only count the IDRs with at least 4 consecutive disordered residues. An IDR/disordered residue could be annotated with multiple functions/ subclasses, so the total number may be different than the corresponding sum.

Dataset	Function	Functional subclass	# IDRs	# disordered AAs
TRAINING	Protein-RNA binding	Protein-tRNA binding	4	308
		Protein-genomic RNA binding	6	435
		Protein-rRNA binding	3	971
		Protein-mRNA binding	3	319
		Protein-RNA binding	0	0
		Total number	**16**	**2033**
	Protein-DNA binding	Protein-DNA binding	59	5091
		DNA unwinding	2	90
		DNA bending	0	0
		Total number	**60**	**5146**
	Protein-protein binding	Protein-protein binding	215	22535
		Autoregulatory	18	1670
		Intraprotein interaction	23	1292
		Protein inhibitor	9	679
		Regulation of proteolysis in vivo	3	237
		Total number	**238**	**24290**
TEST115	Protein-RNA binding	Protein-tRNA binding	5	761
		Protein-genomic RNA binding	2	123
		Protein-rRNA binding	1	600
		Protein-mRNA binding	0	0
		Protein-RNA binding	3	387
		Total number	**10**	**1271**
	Protein-DNA binding	Protein-DNA binding	14	1420
		DNA unwinding	0	0
		DNA bending	1	102
		Total number	**14**	**1420**
	Protein-protein binding	Protein-protein binding	72	6689
		Autoregulatory	1	197
		Intraprotein interaction	3	208
		Protein inhibitor	3	48
		Regulation of proteolysis in vivo	0	0
		Total number	**77**	**6940**
TEST36	Protein-RNA binding	Protein-tRNA binding	2	42
		Protein-genomic RNA binding	0	0
		Protein-rRNA binding	0	0
		Protein-mRNA binding	0	0
		Protein-RNA binding	7	280
		Total number	**9**	**322**
	Protein-DNA binding	Protein-DNA binding	20	948
		DNA unwinding	0	0
		DNA bending	1	5
		Total number	**20**	**948**
	Protein-protein binding	Protein-protein binding	45	2634
		Autoregulatory	1	61
		Intraprotein interaction	19	1217
		Protein inhibitor	1	65
		Regulation of proteolysis in vivo	0	0
		Total number	**52**	**2752**
Others			*339*	*26501*

Appendix D
Selection of Physicochemical Properties of Amino Acids for DisoRDPbind

We collected 531 amino acid (AA) indices from the version 9.1 of the AAindex database (Kawashima, et al., 2000; Nakai, et al., 1988; Tomii, et al., 1996), after removing 13 AA indices with unknown values. Some of these AA indices are redundant with each other or irrelevant to the prediction of disordered DNA-, RNA-, and protein-binding regions. Thus, we empirically selected a subset of non-redundant and relevant indices using the TRAINING dataset. We collected disordered regions with a given functional annotations (set A) and all other regions (including disordered and ordered regions; set B) from the TRAINING dataset. Next, we randomly selected 40% of regions from set A and the same number of regions from set B (the choice of 40% is motivated by the size of the annotation sets to assure that they can be matched), and considered a given AA index by averaging the corresponding numerical values in all regions in each of the two sets. This was repeated 10 times for a given AA index. Consequently, we obtained two vectors of 10 averages. We evaluated significance of the differences between these two vectors. If the measurements are normal, as tested with the Anderson-Darling test (Anderson, et al., 1952) at the 0.05 significance, then we utilized the *t*-test; otherwise we used the non-parametric Wilcoxon rank sum test (Wilcoxon, 1945). Since we considered the disordered DNA-, RNA- and protein-binding, we obtained three *p*-values for each AA index. We averaged these three *p*-values for each AA index and assumed that lower average indicates stronger relations between the corresponding AA index and the disordered DNA-, RNA- and protein-binding regions. The averages were used to rank the AA indices in the ascending order. We selected the top ranked index to initialize the set of selected indices and added a subsequently ranked index if its Pearson Correlation Coefficient (PCC) (Rodgers, et al., 1988; Stigler, 1989) with each AA index that is already in the selected set is < 0.75; otherwise we rejected a given index since it is similar/redundant (i.e., PCC ≥ 0.75) with the already chosen indices. The entire list of ranked indices was scanned once. In total, 159 AA indices were selected. These indices were used to represent the input protein sequence to predict the disordered DNA-, RNA-, and protein-binding regions.

Appendix E
Sequence Representation for DisoRDPbind

We utilize a sliding window to represent information used to perform prediction of the central (in the center of the window) residue; this was done for each residue in the input sequence. The window sizes *ws* were set to 55, 21, and 33 for the disordered RNA, DNA and protein binding residues, respectively. Each position/residue in the input sequence is represented by the following six sets of features:

1. Amino acid (AA) composition, which is defined as the fraction of a given type of AA within the sliding window (20 features).
2. Features based on sequence complexity generated by SEG algorithm (Wootton, 1994; Wootton, 1994; Wootton, et al., 1993; Wootton, et al., 1996) (7 features). Within a given sliding window, we calculated the fraction of AAs in low complexity regions (1 feature), and the average/maximum/minimum length of the low/high complexity regions that is normalized by dividing the number of corresponding complexity regions (2*3 = 6 features). If there is no low (high) complexity region in the sliding window then we set the normalized average/maximum/minimum length of the low (high) complexity regions to 0.
3. Features based on the secondary structure predicted with PSIPRED (McGuffin, et al., 2000) without using PSI-BLAST (12 features). Using the sliding window, we computed the fraction of AAs in helix, strand and coil confirmations, respectively (3 features), and the average/maximum/minimum regions length for a given type of the secondary structure that is normalized by dividing by the number of regions of the corresponding type (3*3 = 9 features).
4. Features based on the putative disorder and globular domain that are predicted with IUPred (Dosztanyi, et al., 2005) (11 features). Based on IUPred prediction for

long and short disordered regions and globular domains, we computed the disorder content (i.e., fraction of disordered residues) and the fraction of AAs in globular domains (3 features), the normalized average/maximum/minimum length of disordered regions with at least 4 residues (3*2 = 6 features), and the average of the two raw propensity values generated by IUPred (2 features).

5. Features based on the selected AA indices (i.e., physicochemical properties of AAs) (159 features; see Section 1.2). We averaged the numerical values of a given AA index in the sliding window.
6. Aggregated features that consider difference between an average value of particular property of the near neighbors, i.e., (*ws*-1)/2 residues in the middle of the sliding window, and remote neighbors, i.e., (*ws*-1)/4 residues at each termini of the sliding window (189 features). We compute these differences for the values of AA composition (20 features), the fractions of residues in low complexity regions (Wootton, 1994; Wootton, 1994; Wootton, et al., 1993; Wootton, et al., 1996) (1 features) and in a given type of secondary structure (3 features), the content of predicted disordered and structured residues (3 features) and the average of predicted propensity scores (3 features) using IUPred's outputs, and the average of the selected AA indices to represent the physicochemical properties (159 features)

In total, we considered 398 features.