



**UNIVERSITÀ
DEGLI STUDI
DI PADOVA**

Università degli Studi di Padova

Dipartimento di Scienze Biomediche

CORSO DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

31° CICLO

**COMPUTATIONAL CHARACTERIZATION OF TANDEM REPEAT
AND NON-GLOBULAR PROTEINS**

Coordinatore: Ch.mo Prof. Paolo Bernardi

Supervisore: Ch.mo Prof. Silvio C. E. Tosatto

Co-Supervisore: Dott. Damiano Piovesan

Dottoranda: Lisanna Paladin

Summary

The first protein structure to be determined in 1962 was hemoglobin, a globe-like, water-soluble and relatively rigid protein with enzymatic activity. Since then, the science of protein structures and function has been biased towards this type of proteins, termed "globular". Experimental and computational methods for the determination of protein features have been mainly designed on globular proteins, however, over the last decades accumulating experimental evidences demonstrate that there is much more than globularity in the protein conformational space. The definition **non-globular proteins (NGPs)** encompass a full spectrum of phenomena, including tandem repetition, intrinsically disordered regions, aggregating domains and transmembrane domains. During my PhD I worked at the characterization of these phenomena through the development of new resources for the identification, collection and description of NGPs.

Tandem repeat proteins (TRPs) are characterized by a repeated sequence which codes for a modular architecture, where structural modules are called "units". They are widespread in all type of organisms, where they carry out fundamental functions. In addition, their explosion in number in eukaryotes suggests an important role in the evolution of complex organisms. TRPs sequences diverge quickly while maintaining their fold, hampering detection by traditional methods for sequence analysis. The same holds true for functional annotation, which usually relies on the transfer of knowledge between conserved sequences. On the other side, the challenges of repeats detection by structure lies in the multidimensional nature of the data. Addressing these challenges, sequence- and structure-based methods were built recently for the identification of repeat proteins, however a limited number of them address the problem of the annotation of repeat units. However, data about unit position could be a powerful tool not only to classify TRPs, but also to understand TRPs evolution and assess conservation at the sequence level, since the repeat unit is the evolutionary module of repeated structures. Moreover, the collection of an alphabet of tandem repeat units could be useful for protein engineering applications, as repeat proteins are extensively used in protein design. RepeatsDB is a database of tandem repeat protein structures annotated with the position of repeat units and insertions, i.e. non-repeated segments of structure that occur either inside a repeat unit or between two of them. I addressed the problem of the annotation of repeat proteins by contributing to RepeatsDB and related resources,

starting from the curation of RepeatsDB data. I provided insights on TRPs role in the human proteome by characterizing them in terms of function, protein-protein interaction networks and impact on diseases. As a case-study of this phenomenon, I dissected the interactome of Collagen V, a repeat protein associated to Ehlers-Danlos syndrome, in order to identify genotype-phenotype correlations in relation to the interaction network model. Moreover, I compared the sequence-based classification of repeats to the structural one provided by RepeatsDB. This work was based on the observation that the improvement of TRPs recognition and classification is essential to shed a light on the so called dark proteome, i.e. the large fraction that we know almost nothing about. On this line, I took care of the curation and improvement of RepeatsDB database. The new version of the database was populated taking advantage of ReUPred, predictor of tandem repeat units to which I contributed to develop. The quality of RepeatsDB data is guaranteed by an extensive manual validation, a time-consuming task which requires community annotation efforts. To facilitate this process I developed RepeatsDB-lite, web server for the prediction and refinement of tandem repeats in protein structure. I also contributed to the new release of MobiDB database, addressing the problem of **intrinsically disordered proteins (IDPs)** annotation. IDPs are devoid of order in their native state. The discovery of intrinsic disorder and its prevalence and functional importance is transforming the field of molecular biology. It was shown to be prevalent in the human proteome, to play important signaling and regulatory roles and to be frequently involved in disease. As intrinsic disorder is emerging as a general phenomenon, databases are collecting and presenting disorder related data in a systematic manner. During my PhD I had the opportunity to contribute to MobiDB, database of protein disorder and mobility annotations that describes several aspects of NGPs structure and mechanism of function. MobiDB has been a major contributor by providing consensus predictions and functional annotations for all known protein sequences, driving the field ahead.

A common feature of TRPs, IDPs and other NGPs is that they are characterized by **low-complexity regions (LCRs)**, where the distribution of amino acids deviates from the common amino acid usage. LCRs have been estimated at 20% and 8% of all known sequences of eukaryotes and non-eukaryotes, respectively. The functional importance of LCRs is strictly related to their non-globular arrangement and their involvement in disease has also been extensively discussed. Overcoming early reluctance

to consider these regions for biological studies, mainly due to their unknown properties and annoying statistical features, there is an intensification of research on low complexity. I contributed to the field with a critical review focusing on the definition of sequence features of low complexity regions and their relationship to structural features.

Finally, I exploited the knowledge acquired on NGPs in the previous studies to design one of the first sequence-based methods for the prediction of protein **solubility**, SODA. SODA uses the aggregation propensity, intrinsic disorder, hydrophobicity and secondary structure preferences from the sequence to evaluate solubility changes introduced by a mutation. Solubility is an important, albeit not well understood, feature determining protein behavior. The main envisaged applications of SODA are in protein engineering, where it can help the design of proteins with more favourable surface properties and the study of the impact of protein mutations in disease insurgence.

List of publications

1. Paladin L. and S. C. E. Tosatto. "Comparison of Protein Repeat Classifications Based on Structure and Sequence Families." *Biochemical Society Transactions*, vol. 43, no. 5, Jan. 2015, pp. 832837., doi:10.1042/bst20150079.
2. Paladin L., Tosatto S.C.E., Minervini G. "Structural in Silico Dissection of the Collagen V Interactome to Identify Genotype-Phenotype Correlations in Classic Ehlers-Danlos Syndrome (EDS)." *FEBS Letters*, vol. 589, no. 24PartB, 2015, pp. 38713878., doi:10.1016/j.febslet.2015.11.022.
3. Hirsh L., Piovesan D., Paladin L., Tosatto S.C.E. "Identification of Repetitive Units in Protein Structures with ReUPred." *Amino Acids*, vol. 48, no. 6, 2016, pp. 13911400., doi:10.1007/s00726-016-2187-2.
4. Paladin L.*, Hirsh L.*, Piovesan D., Tosatto S.C.E. "RepeatsDB 2.0: Improved Annotation, Classification, Search and Visualization of Repeat Protein Structures." *Nucleic Acids Research*, vol. 45, no. D1, 2016, doi:10.1093/nar/gkw1136. (*joint first authors)
5. Paladin L., Piovesan D., Tosatto S.C.E. "SODA: Prediction of Protein Solubility from Disorder and Aggregation Propensity." *Nucleic Acids Research*, vol. 45, no. W1, May 2017, doi:10.1093/nar/gkx412.
6. Piovesan D., Tabaro F., Paladin L., Necci M., Micetic I., Camilloni C., Davey N., Dosztanyi Z., Mszros B., Monzon A.M., Parisi G., Schad E., Sormanni P., Tompa P., Vendruscolo M., Vranken W.F., Tosatto S.C.E. "MobiDB 3.0: More Annotations for Intrinsic Disorder, Conformational Diversity and Interactions in Proteins." *Nucleic Acids Research*, vol. 46, no. D1, Nov 2017, doi:10.1093/nar/gkx1071.
7. Hirsh L.*, Paladin L.*, Piovesan D., Tosatto S.C.E. "RepeatsDB-Lite: a Web Server for Unit Annotation of Tandem Repeat Proteins." *Nucleic Acids Research*, vol. 46, no. W1, Sept. 2018, doi:10.1093/nar/gky360. (*joint first authors)
8. Mier P., Paladin L., Tamana S., Petrosian S., Hajdu-Soltz B., Urbanek A., Gruca A., Plewczynski D., Grynberg M., Bernad P., Gspri Z., Ouzounis C., Promponas V.J., Kajava A.V., Hancock J.M., Tosatto S., Dosztanyi Z., Andrade-Navarro M.A. "Disentangling the complexity of low complexity proteins" *Submitted*.

632

The Brain- is wider than the Sky-
For- put them side by side-
The one the other will contain
With ease- and You- beside-

The Brain is deeper than the sea-
For- hold them- Blue to Blue-
The one the other will absorb-
As Sponges- Buckets- do-

The Brain is just the weight of God-
For- Heft them- Pound for Pound-
And they will differ- if they do-
As Syllable from Sound-

EMILY DICKINSON

Acknowledgments

I would like to thank those who made it possible not this goal alone, but also this entire scientific and human path. First of all, Professor Silvio Tosatto. For years now he has been continually demonstrating that he believes in me, doing it in the way that suits me the best: proposing me one challenge after another. This PhD experience itself is the most important one. Thanks to him I met my other mentors. Dr. Damiano Piovesan, who has always found the time to give me help and advice, and Dr. Giovanni Minervini, example of enthusiasm and tenacity. For me, the BioComputingUP Lab has always been the right place where to find insightful comments and guidance, but also encouragement and support, thanks to all the colleagues and friends: Alessandro, Emilio, Manuel, Alessandra, Francesco, Ivan, Marco and Marco. "Standing on the Shoulders of Giants" - a heartfelt thanks also to those who demonstrated participation to my scientific activity welcoming me elsewhere in the world: Professor Rita Casadio, to whom I owe a large part of my education, Professor Miguel Andrade, Professor Salvador Ventura and Robert Finn, together with their research groups. Each of them represents for me a growth opportunity that I had, as well as a contribution to the work contained in this thesis. There are still many scientists I think I owe to, many more than I can enumerate here. I had the opportunity to have insightful discussions with some of them on the grounds of the COST Action NGPnet, which I cite here to symbolically thank all of them. "The Soul selects her own Society" - there are other people who have always been there. Diana, since the very beginning of my academic path, and Vanessa. Without them I would never been able to overcome the difficulties along it. Veronica and Viola, who shared with me the beginning of this journey and lately remained in my life. Nadia, who has always been a friend, and Irene, the best rediscovery I have ever done. Thanks to all those who are behind the scenes of this research work: those who call me, those who listen to my voicemails, those who interpret my messages, those that will wait for me at home. Inevitably and immensely thanks to my family. A child's debt to his parents can not be described, but I am fully aware that they shaped the person I am. The same applies to my brothers, Milena and Alessandro, to whom I dedicate this thesis. Everything that I do, I do it to be worthy of them.

Contents

List of Figures	vii
List of Tables	ix
Glossary	xi
1 Introduction	1
1.1 Principles of protein structure	1
1.1.1 Primary structure	1
1.1.2 Secondary structure	4
1.1.3 Amino acids secondary structure preference	5
1.1.4 Tertiary structure	5
1.1.5 Quaternary structure and protein interaction	6
1.1.6 Determination and classification of protein structures	7
1.2 Principles of protein folding	8
1.2.1 Basic principles of thermodynamics	8
1.2.2 Folding and denaturation	9
1.2.3 The hydrophobic effect	10
1.2.4 Energy landscape of protein folding	10
1.2.5 Protein folding and solubility	12
1.3 Relationship between protein sequence, structure and function	13
1.3.1 Sequence-structure-function paradigm	14

CONTENTS

1.3.2	Sequence evolution	14
1.3.3	Protein families	15
1.3.4	Domain architecture	16
1.3.5	Protein function	16
1.4	Non-globular proteins	19
1.4.1	Tandem Repeat Proteins	19
1.4.1.1	Solenoids	20
1.4.1.2	Folding	21
1.4.1.3	Organism distribution	22
1.4.1.4	Identification	23
1.4.1.5	Prediction of repeat units	23
1.4.1.6	Evolution	24
1.4.2	Disordered proteins	25
1.4.2.1	Sequences	25
1.4.2.2	Function	26
1.4.2.3	Detection	28
1.4.3	The relationships between repeats and disorder	28
1.4.4	Low complexity	29
1.4.4.1	Structural features	30
2	Personal Contribution and Thesis outline	33
2.1	Tandem repeats	33
2.2	Intrinsical disorder	35
2.3	Low complexity sequences	36
2.4	Solubility	36
3	Materials & methods	37
3.1	Prediction of tandem repeat units in structures	37
3.1.1	ReUPred	37
3.1.2	RepeatsDB-lite	43
3.2	Prediction of changes in protein solubility	46
3.2.1	SODA	46
3.3	Web resources implementation protocols	48
3.3.1	Databases	49

3.3.2	Interfaces	49
4	Results & Discussion	53
4.1	Protein tandem repeats characterization	53
4.1.1	Comparison of protein repeat classifications based on structure and sequence families	54
4.1.1.1	Class I: Crystalline aggregates of unlimited size	54
4.1.1.2	Class II: Fibrous structures	54
4.1.1.3	Class III: Elongated structures	56
4.1.1.4	Class IV: Closed structures	58
4.1.1.5	Class V: Beads on a string	58
4.1.2	Tandem Repeat proteins at a glance: functions, diseases and role in protein-protein interaction network	59
4.1.2.1	A structure of success	60
4.1.2.2	The perfect binder	61
4.1.2.3	A protein population that colonized every organism dis- trict	61
4.1.2.4	Probable candidates of disease-association	63
4.1.2.5	Centrality comes with a price	66
4.1.3	Structural in silico dissection of the collagen V interactome to identify genotype-phenotype correlations in classic Ehlers-Danlos Syndrome (EDS)	66
4.1.3.1	Genotype/phenotype correlation	67
4.1.3.2	Model construction and evaluation of mutations	67
4.1.3.3	Detection and analysis of interactors	69
4.1.3.4	Ehlers-Danlos association to Collagen V	72
4.2	Protein tandem repeats identification and annotation	78
4.2.1	Identification of repetitive units in protein structures with ReUPred	79
4.2.1.1	Repeat classification	80
4.2.1.2	Unit prediction accuracy	81
4.2.1.3	Expanding the universe of known solenoids	84
4.2.1.4	Benchmarking	86

CONTENTS

4.2.2	RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures.	87
4.2.2.1	Database description	87
4.2.2.2	Database usage	88
4.2.3	RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins	92
4.2.3.1	Web server description	94
4.2.3.2	Usage example	94
4.2.3.3	RepeatsDB-lite performance	96
4.2.4	Improving Repeat Definitions in Pfam	99
4.3	Intrinsically disordered proteins	101
4.3.1	MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.	101
4.3.1.1	Database description	101
4.3.1.2	Usage and annotated data	105
4.4	Low complexity sequences	107
4.4.1	Disentangling the complexity of low complexity proteins	107
4.4.1.1	The many shades of complexity	108
4.4.1.2	Detection of low complexity sequences	110
4.4.1.3	Correlation between low complexity and disorder	117
4.4.1.4	Structural properties of LCRs	120
4.5	Protein aggregation and protein solubility	128
4.5.1	SODA: Prediction of protein solubility from disorder and aggregation propensity	128
4.5.1.1	Benchmarking	129
4.5.1.2	Server description	129
4.5.1.3	Usage examples	130
5	Conclusions	135
	References	143

List of Figures

1.1	Summary of protein structure vocabulary.	2
1.2	Ramachandran plot.	3
1.3	The folding funnel.	11
1.4	Structure of a globular protein.	17
1.5	Schematic of a linear (or open) solenoid protein domain.	20
1.6	Relationship between sequence, structure and repeats.	21
1.7	Structure of an intrinsically disordered protein.	25
3.1	Schematic description of the ReUPred algorithm	40
3.2	Evaluation of repeat unit predictions.	42
3.3	Schema of implementation technologies	50
4.1	RepeatsDB classification compared to Pfam families	55
4.2	Repeat proteins association to diseases and ubiquity.	61
4.3	Repeat proteins GO cellular compartment enrichment.	62
4.4	Repeat proteins number of interactors.	63
4.5	Repeat proteins mutations and diseases.	64
4.6	Collagen V structure	68
4.7	Collagen V interactome	70
4.8	ReUPred unit prediction	80
4.9	Repeat unit number	82

LIST OF FIGURES

4.10 Repeat unit periodicity.	82
4.11 Large-scale periodicity predictions.	83
4.12 Raphael and ReUPred unit periodicity.	84
4.13 RepeatsDB 1.0 available annotation.	85
4.14 Retrieving RepeatsDB data, browse.	89
4.15 Retrieving RepeatsDB data, search and results.	90
4.16 RepeatsDB entry page.	91
4.17 RepeatsDB growth.	92
4.18 RepeatsDB-lite result page.	93
4.19 RepeatsDB-lite editing.	95
4.20 RepeatsDB-lite editing results.	96
4.21 MobiDB data.	102
4.22 The low complexity diagram: sequence complexity composition versus periodicity.	110
4.23 Shannon entropy of CAST CBRs	113
4.24 Motif graph based on SIMPLE analysis of CO1A1_HUMAN.	116
4.25 Low complexity methods comparison on the dataset.	117
4.26 Low complexity diagram for various sequence datasets.	119
4.27 Structural features of low complexity proteins.	120
4.28 Structural features of low complexity proteins, examples.	124
4.29 C-terminal of CTCF_HUMAN.	124
4.30 Sample SODA mutation input page.	131
4.31 Sample SODA mutation input output.	131
4.32 Sample SODA full protein mode page.	132
4.33 Sample SODA full protein output.	133

List of Tables

1.1	Overview of low complexity terms and their definitions.	30
3.1	Structural alignment constraints for the "master" unit.	39
3.2	Structural alignment constraints for the "secondary" units.	40
3.3	RepeatsDB solenoid dataset used in ReuPred benchmarking.	41
3.4	Validation rules for RepeatsDB-lite predictions.	43
3.5	Evaluation of SODA on the PON-Sol training set.	48
4.1	Dataset sizes in human proteome analysis.	59
4.2	Top 20 Pfam domains associated to diseases.	65
4.3	Collagen V interactors	74
4.4	Solenoid detection performance on the RAPHAEL dataset	79
4.5	ReUPred classification performance.	81
4.6	Comparative repeat unit prediction evaluation	81
4.7	RepeatsDB-lite benchmarking	97
4.8	RepeatsDB-lite versus reviewed entries	98
4.9	MobiDB 3.0 databases.	103
4.10	Overview of tools used into MobiDB 3.0.	105
4.11	Low complexity dataset.	108
4.12	Compositionally biased regions (CBR) detected by CAST.	112
4.13	Repeats detected by SIMPLE.	115

LIST OF TABLES

4.14 Comparison between low complexity methods	116
4.15 Number of residues predicted to be in different structural states.	125
4.16 Comparison of SODA with other predictors on the CamSol dataset.	129

CATH Class, Architecture, Topology and Homology

CBR Compositionally Biased Regions

CSAH Charged Single Alpha-Helix

DPP Dentin phosphophoryn

DSP Dentin sialoprotein

DSPP Dentin sialophosphoprotein

DSSP Define Secondary Structure of Proteins

EDS Ehlers-Danlos Syndrome

FN False Negative

FP False Positive

FV Feature Viewer

HMM Hidden Markov Model

IDP Intrinsically Disordered Protein

IDR Intrinsically Disordered Region

LIST OF TABLES

LC Low Complexity

LCR Low Complexity Region

NMR Nuclear Magnetic Resonance

PDB Protein Data Bank

PPI Protein Protein Interaction

PTM Post Translational Modification

ReUPred Repeat Unit Predictor

SAH Single Alpha-Helix

SCOP Structural Classification of Proteins

SODA SOLubility prediction from Disorder and Aggregation propensities

SRUL Structural Repeat Unit Library

TN True Negative

TP True Positive

TR Tandem Repeat

TRP Tandem Repeat Protein

UniProt Universal Protein resource

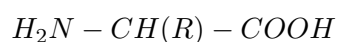
WT Wild Type

1.1 Principles of protein structure

In 1962 the Nobel Prize for Chemistry was given to John Kendrew and Max Perutz, who independently determined the structure of myoglobin (1) and hemoglobin (2), respectively. Since then, tens of thousands of different protein structures have been determined and enable scientists to understand the architectural and energetic principles of proteins, as well as their mechanism of function. The natural proteins are mapped to the primary database of protein annotation, the Universal Protein Resource - UniProt (www.uniprot.org). UniProt contains sequence information, functional annotations, cross-references to other resources on structure. Protein structure is organized in a hierarchical manner (3), for which scientists have devised a hierarchical vocabulary to describe protein architecture, described in next section.

1.1.1 Primary structure

The primary structure (backbone) of a protein refers to the ordered sequence of amino acids composing the polypeptide chain. The general structure of the 20 amino acids that can make up protein is



1. INTRODUCTION

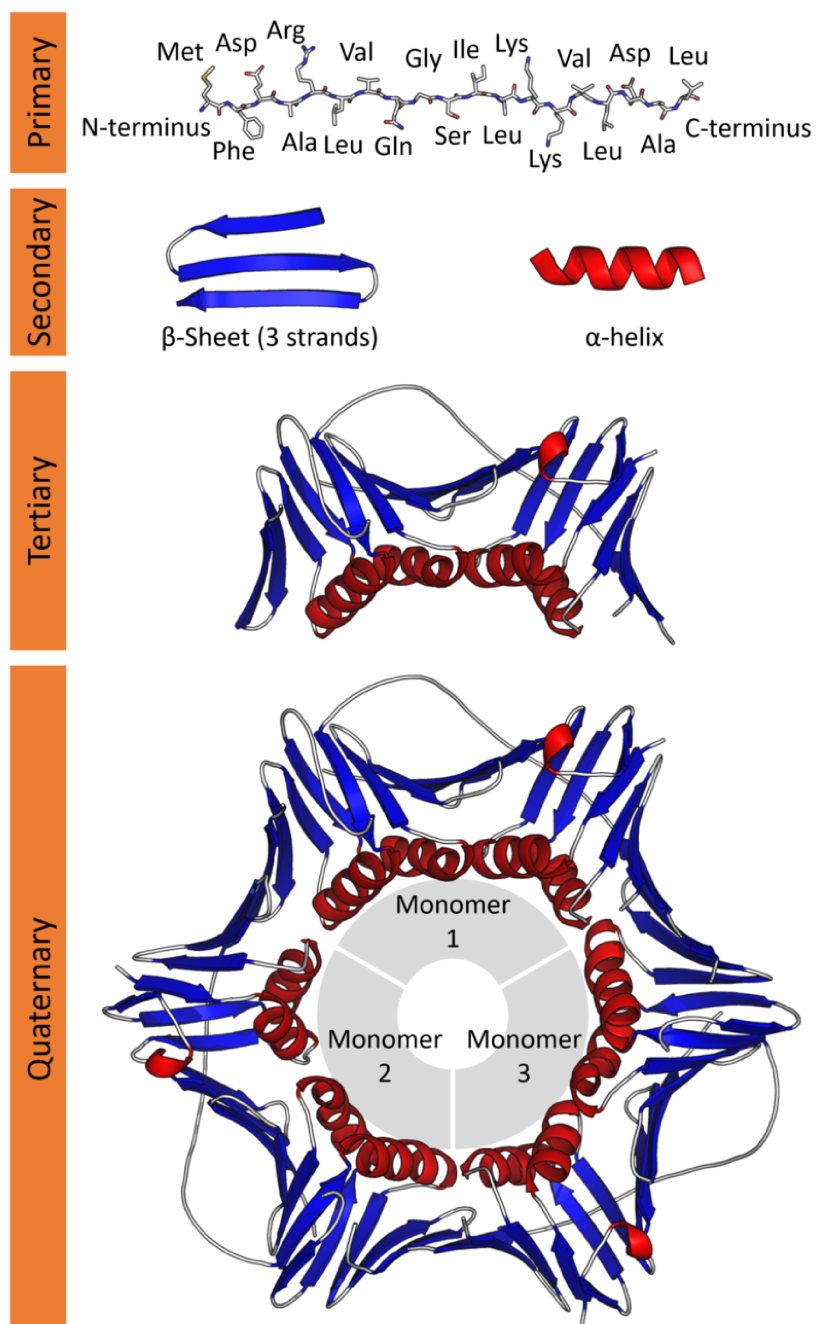


Figure 1.1: Summary of protein structure hierarchical vocabulary (primary, secondary, tertiary, and quaternary) using the example of PCNA (PDB: 1AXC). License: Creative Commons Attribution 4.0 International. Author: Thomas Shafee.

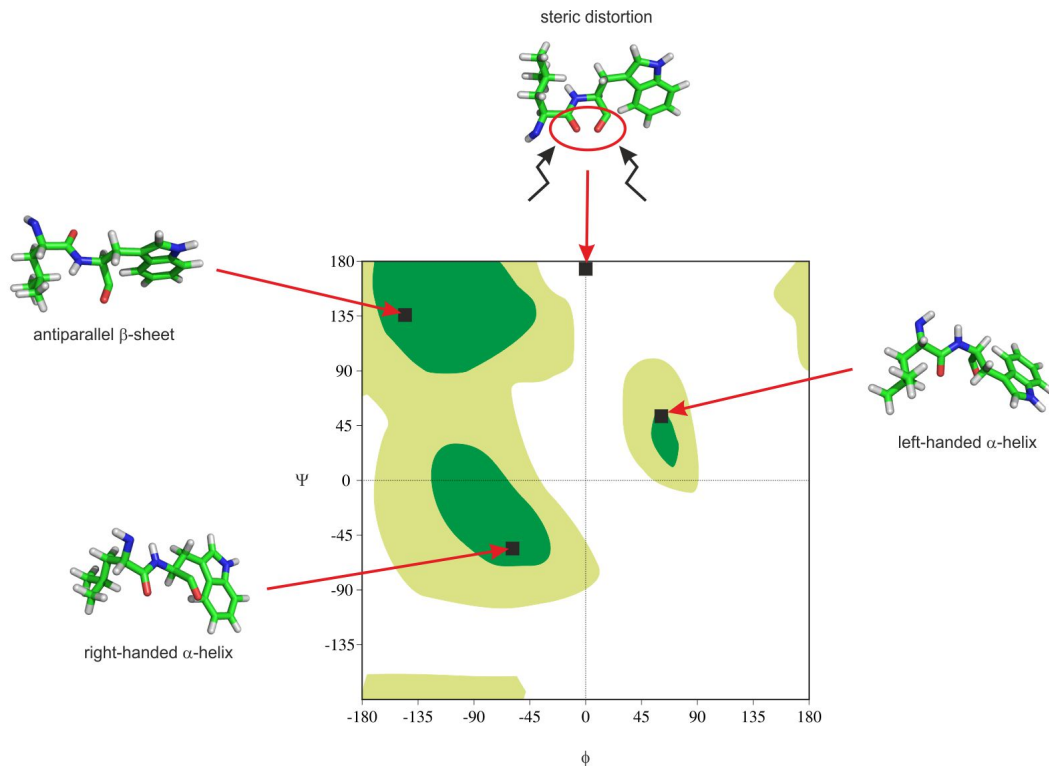


Figure 1.2: Ramachandran plot (4) with the most favoured (dark green) and additional allowed (light green) regions. Examples of relative orientation of amino acids in three secondary structure elements (antiparallel β -sheet, right-handed α -helix and left-handed α -helix) are provided, together with exemplary steric distortion where two oxygen atoms are too close to each other. Author: Krzysztof Brzozowski.

”R” refers to the variable side-chain. Proteins exclusively use the L form of the two possible enantiomers of amino acids, L and D. The variable side-chain confers unique properties, i.e. acidity, basicity, hydrophobicity and hydrophilicity. This way, proteins are provided the chemical toolkit to assemble the amino acid modules in unique complex structures with unique properties and interaction with environment, which ultimately defines their function. The assembly of amino acids is achieved through amide bond between their α amino and carboxylic groups. When linked in the polypeptide chain (see Figure 1.1, Primary), amino acids are also called residues. Their sequence, or indeed primary structure, is rendered in the direction from the first amino acid with a free α -amino group (N-terminus) to the last with free carboxylic group (C-terminus). This convention is due to the direction of protein synthesis.

1. INTRODUCTION

1.1.2 Secondary structure

The polypeptide chain folds up into local spatial elements termed secondary structure. Ramachandran in the late '60s gave an elegant description of the regularity of these local conformational elements through the Ramachandran plot (4), in Figure 1.2. Most of the possible arrangements of adjacent residues fall into four main classes, mapped into the plot of the two dihedral angles around the α -carbon atom of the peptide bond ($C\alpha$). The torsion angles are ϕ (phi) and ψ (psi), corresponding to the rotation of the two adjoining amide plains around the bond connecting them to the $C\alpha$, and respectively x and y axes of the plot. Adjacent residues conformations are described by the ϕ, ψ value pairs which determine the local element of secondary structure (see Figure 1.1, Secondary).

- α -helices are righthand-spiral conformations (i.e. helix) in which the polypeptide chain takes turn allowing every backbone NH group to donate a hydrogen bond to the backbone C=O group of the amino acid located three or four residues earlier along the protein sequence (the average is 3.6 residues per turn). The distribution of constituting amino acids usually originates helix sides with different physicochemical properties.
- β -sheets are another basic building block of proteins. They are constituted by β -strands, fully extended element having 2 residues per turn and unable to allow intra-chain interactions. These elements are stabilized by inter-chain H bonds, and the assembly is defined parallel or antiparallel according to the relative orientation of the two interacting strands.
- proteins reverse the direction of their polypeptide chain through secondary structure elements called turns. They come in several variants, depending on the involved amino acids, and thus occupy different regions in the Ramachandran plot.

Other secondary structure arrangements are present in nature, however the exhaustive description of these elements goes beyond the scope of the present dissertation. Secondary structure elements are easily identified in an automated way through the Define Secondary Structure of Proteins (DSSP) software (5) when the 3D structure is provided. It is based not on angles but on the pattern of intra-backbone hydrogen

bonds. As already introduced, indeed, the conformations of secondary structure elements such as α -helices and β -sheets are stabilized by a specific pattern of hydrogen bonds. Among all possible shapes, these two were selected exactly because capable of efficiently packing atoms in a compact way and of pairing backbone amide and carboxyl groups in energetically favourable bonds (6). Protein energetics and the mechanisms driving their folding will be deepened in section 1.2.

1.1.3 Amino acids secondary structure preference

The different 20 amino acid types have different physical and chemical characteristics, including polarity, shape, dimensions. As already apparent from the Ramachandran's prediction of the allowed ϕ and ψ angles (4), the different secondary structure elements place some constraints to the amino acids that they can include. This has been confirmed by statistical analysis of amino acid frequencies in secondary structure elements of protein structures (7) and by the analysis of mutations which disrupt secondary structure elements (8, 9). These studies demonstrate the thermodynamic preferences of the different secondary structure elements for certain amino acids. This framework is at the basis of the prediction of protein secondary structure arrangement when only the sequence is provided.

1.1.4 Tertiary structure

Secondary structure elements are local and usually proceed along one axis of the protein chain. The atoms in different secondary structure elements may establish contacts and further compact the protein structure, shaping its three-dimensional arrangement (tertiary structure). The tertiary structure (see Figure 1.1, Tertiary) optimizes the various attraction forces between amino acids in the chain and with the environment. The evolution of biological molecules took place in aqueous solvent. Inside living organisms these environments include the cell cytosol, interstitial fluids, multi-cellular fluid environment such as blood, saliva, lymph, etc. As a result, most proteins evolved to bear the following properties:

- Compactness: Due to the highly crowded environment (10) in which proteins exist, their structure must be dense enough and still retain the ability to diffuse freely.

1. INTRODUCTION

- Solubility: Proteins need to be water-soluble.
- Binding specificity: As a consequence of the previous property, proteins binding needs to be specific. Non-specific binding to each other or other molecules, in this kind of environment, may result in deleterious aggregation.

The proteins described by the listed features are termed **globular**. They achieve this state by exploiting the different properties of their composing amino acids. Indeed, they contain both polar and non-polar amino acids. To remain water-soluble, they fold so as to allow polar residues to be on the surface and non-polar residues buried in the core, thus termed "hydrophobic core" (11). The partitioning of polar/non-polar residues is to a large extent due to the secondary elements mentioned above. β -sheets are very efficient in the burial of non-polar residues, whereas α -helices allow simultaneous externalization of polar and internalization of non-polar ones (12). The physico-chemical properties driving the folding process are better described in section 1.2.

1.1.5 Quaternary structure and protein interaction

The three levels of structural hierarchy exist in all proteins, but the third is not necessarily the final level. Some proteins include more than one chain. In such cases, the native state of the macromolecule may be constituted by the assembly of multiple folded protein subunits (not directly linked in the same polypeptide chain) in a multi-subunit complex, namely the quaternary structure of a protein. Quaternary structure (see Figure 1.1, Quaternary) is the number and arrangement of the different members involved in a multi-subunit complex. These complexes include organisations from simple homodimers (two copies of the same structure) to large oligomers and complexes with defined or variable numbers of subunits. When the complex is made up by several copies of the same protein, the copies are called "monomers". Proteins can form biomolecular complexes also with nucleic acids and other cofactors. Virtually all cellular processes include key players which are complexes (13). In this sense, it should be noted that many cellular proteins tend to physically interact with other members of the biochemical pathway in which they are involved (14). Therefore, the identification of a protein interactions, i.e. the description of the protein-protein interaction (PPIs) network, is essential to understand their function. These interactions make up the

so-called interactomics of the organism, while aberrant PPIs are the basis of multiple diseases. The interaction that a protein is able to establish are the result of biochemical events steered by electrostatic forces, and similar to the ones that drive their folding, described in section 1.2.

1.1.6 Determination and classification of protein structures

The function that a protein assumes and the interaction that it can establish depend on its structure. Therefore, protein structure determination is of utmost importance in the study of living organisms, in drug design and as a necessary background information for protein design. Several techniques have been designed to build atomic models of biomolecules. These techniques include X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and electron microscopy. X-ray crystallography (15) is based on the purification and crystallization of a single protein species. The crystal is illuminated with a finely focused beam of X-rays, producing a diffraction pattern of regularly spaced spots. Several two-dimensional images are taken at different orientations mounting the crystal on a goniometer. Finally, the three-dimensional model of the structure is reconstructed using the mathematical method of Fourier transforms. The method has been a more commonly used technique and obtaining protein structure information is a routine, highly automated procedure. Yet, it requires crystallization of the protein, which can take months. On the other hand, NMR spectroscopy (16) allows the study of a protein nearly under physiological conditions. NMR uses a large magnet to probe the intrinsic spin properties of atomic nuclei. The sample is placed in a magnetic field and the NMR signal is produced by excitation of the nuclei sample with electromagnetic radiation (radio frequency waves) into nuclear magnetic resonance, which is detected with sensitive radio receivers. The magnetic resonance of a molecule is determined by its electronic structure and its individual functional groups, therefore the molecular structural details can be derived by its NMR signal. Even if this technique offers the possibility to determine the structural behaviour of a protein in its native environment, NMR spectroscopy experiments are hard to automate. All the information derived by structural determination techniques is stored in the Protein Data Bank (PDB, www.pdb.org), database of protein and nucleic acid structures (17). At the moment of writing, the database contains atomic coordinates of more than

1. INTRODUCTION

140,000 structures of natural or designed biomolecules, of which about 80% is determined by X-ray crystallography and 9% by NMR spectroscopy. Other methods provide minor contributions. Proteins contained in the PDB can be classified in groups sharing structural similarities. The two main resources for protein structure classification are the Structural Classification of Proteins, SCOP (<http://scop2.mrc-lmb.cam.ac.uk/>, (18)) and CATH (<http://www.cathdb.info/>, (19)), which states for Class, Architecture, Topology and Homology. They are both based on structural similarities between proteins, at all levels from the secondary to the quaternary. Proteins sharing 3D similarities are grouped into "folds". It was shown (20) that the clustering of these groups does not map perfectly to the clustering of their primary structures (sequences), indicating a non-linear relationship between the protein sequence and structure. This relationship will be described in section 1.3.

1.2 Principles of protein folding

Proteins are physical entities subjected to the physical forces that dominate our universe, and in order to understand the principles of their folding and functioning it is necessary to introduce some concepts about thermodynamics. The field that describes proteins in terms of forces and energies is termed structural biophysics, which can be in turn be separated into two major fields:

- *Energetics* studies the principal forces that affect protein folding and stability.
- *Dynamics* studies the conformational changes of the polypeptide chain during folding, including those that occur in the native state.

The basic principles of thermodynamics applied to the description of proteins will be summarized in following subsections.

1.2.1 Basic principles of thermodynamics

Thermodynamics exploits the energy characterization of states in nature to predict the direction and probability of processes, e.g. folding vs. unfolding of a property. In an isolated system (cannot exchange matter or energy with its surroundings), spontaneous processes tend to increase the system *entropy*, that is the number of possible configurations of the system (21). Biological system, however, are not isolated. They do not

exist in a state of constant volume and energy. In such systems, the most convenient way to define the direction of a process is the Gibbs free energy, symbolized by G . It represents the energy which is "available" to systems under constant temperature and pressure and not volume (22). The G is defined as:

$$G = U + pV - TS \quad (1.1)$$

where U is the internal energy of the system, p is pressure, V is volume, T is the temperature, S is the entropy. Alternatively,

$$G = H - TS \quad (1.2)$$

where H is the so called *enthalpy*, a quantity which represents the internal energy of the system, plus the product of its pressure and volume. The total enthalpy of a system cannot be measured directly: only a change in energy carries a physical meaning, such as in classical mechanics. What one can measure is the difference in enthalpy, ΔH , with respect to a reference. Consequently, the same holds for the Gibbs free energy, measured as ΔG . Spontaneous processes tend to decrease the free energy of the system, so they have negative ΔG . Thus, by measuring the free energy change of any process, one would be able to assess the spontaneity or not of the process (23).

1.2.2 Folding and denaturation

Proteins native structures represent the state of minimum Gibbs free energy in the systems where they are placed. In physiological environment, in the case of globular proteins this minimum corresponds to the folded structure but in other conditions this may change. It was shown by Hisen Wu in the early '30s (24) that in specific conditions protein lose their structure (a process called "denaturation"). In these states, therefore, the conformation corresponding to the lowest Gibbs free energy is the unfolded one. It was also found that after removing the agents that caused the change of environment, some proteins can automatically retake their native structure, this is called "renaturation" or "refolding". Experiments on denaturation and renaturation lay the foundation for the theory of protein folding (25). Within a cell, protein folding must be thermodynamically favourable in order it to be spontaneous. However, protein folding decreases the number of possible configurations of the protein, that is, the entropy. Such events are common in bio-systems, because over-compensated by other factors. In

1. INTRODUCTION

protein folding, the ordering of polypeptide chain is coupled with the effects on water around the protein and by the formation of favourable non-covalent interactions. In the Gibbs free energy formula, the free energy changes resulting from chemical bonding are represented by ΔH , while $-T\Delta S$ represents the difference in the system's degree of order.

1.2.3 The hydrophobic effect

The main force which drives protein folding is the so called *hydrophobic effect* (26), already anticipated in section 1.1, together with hydrogen bonding. The hydrophobic effect is the process through which the polypeptide chains minimize the number of hydrophobic side-chains exposed to water, collapsing into the protein core away from the hydrophilic environment. The process starts from the unfolded state of the protein, where the hydrophobic surfaces are exposed to the water molecules which, in response, tend to aggregate around them in ordered states. The hydrophobic collapse breaks the ordered state of water molecules thus introducing entropy in the system. Within the core of the globular folded protein, the large number of hydrophobic chains interact via van der Waals forces. In the meantime, the backbone secondary structure elements are stabilized by hydrogen bonds enveloped in a hydrophobic environment. Since the strength of hydrogen bonds is influenced by their surroundings, the hydrogen bonds buried in the hydrophobic core contribute more than the ones in the surface to the stability of the native state.

1.2.4 Energy landscape of protein folding

Until now protein folding was presented as a two state process (unfolded/folded). In very small and globular protein, this simplistic view is justifiable, because protein folding is a highly cooperative process and in those cases it happens very fast. However, protein folding happens at very different speed rates according to the protein type, in some cases being much more gradual. This should not be surprising, as because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an enormous number of possible conformations. This observation was firstly made by Cyrus Levinthal, and it is known as the Levinthal's paradox (27). The astronomical number of possible states of a protein structure should slow down the process of its sequential sampling of all possible conformation to a time longer than the

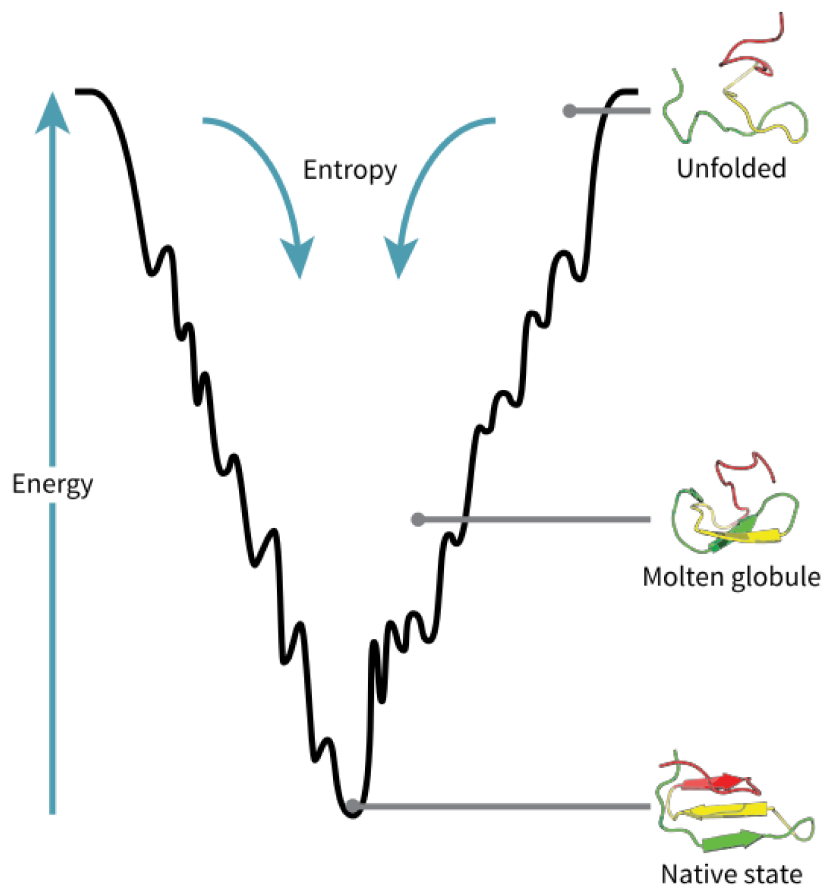


Figure 1.3: The diagram sketches how proteins fold into their native structures by minimizing their free energy. Licensing: Creative Commons Attribution-Share Alike 3.0 Unported. Author: Thomas Splettstoesser.

age of universe. It became clear that protein folding is not simply a sequential sampling of all possibilities, but, as Levinthal himself suggested, it "is sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide; this suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process" (28). To describe this process, the landscape of protein folding was described as a *funnel* (29), characterized by a large number of possible pathways and intermediate states but largely directed toward the native state. This approach introduced the *principle of minimal frustration* (28). The principle states that globular protein sequences were selected so that the folded state of the protein is very stable. The undesired interactions along the folding pathway are

1. INTRODUCTION

reduced, and the whole process is directed towards the native state. However, even though nature has reduced the *frustration* in proteins, some degree of it remains and it can be represented by the local minima in the energy landscape of proteins. The representation in Figure 1.3 shows how:

- The folding process involves both a decrease of energy and of entropy of the protein. The protein native state corresponds to the global minimum of the landscape.
- Many local minima are involved in the folding process, represented by the ragged walls of the funnel. The local minima are kinetic obstacle to the descending to the native state, and even if they do not change the overall free energy between unfolded and folded state they influence the speed of the process, as the protein needs to overcome these energetic barriers to proceed to the folded form.
- The folding process may happen through different paths, represented by the funnel walls.
- During the folding process, different thermodynamic intermediates are formed. Among these, the *molten globule* (30), a compact conformation with similar secondary structure elements to that of the native form. This form lacks though the tertiary structure organization, having the side-chains in a loosely packed conformation. Thus, it tends to change into the native state, which is more stable, ending the folding process.

1.2.5 Protein folding and solubility

The processes focused until now are the ones determining the folding of a protein in an aqueous solute. Life (and proteins) originated in water and water still plays an undeniable role in cells accounting for 70% or more of total cell mass. Protein folding is determined in large part to the interaction with water and therefore protein stability and solubility are inevitably linked. Both depend on extrinsic factors (temperature, pH, ionic strength, etc.) and intrinsic factors (amino acid composition, etc.). However, they can be modulated (almost) independently. Removing charges may lead to lower solubility, sometimes with no impact in folding and stability (31). Redistributing charges on the surface may lead to altered folding and/or stability with low impact in

1.3 Relationship between protein sequence, structure and function

solubility (31). An impairment in both the processes, on the other hand, may impact on aggregation propensity, thus leading to pathogenic pathways. Knowledge about how intrinsic factors influence solubility is limited due to the difficulty of obtaining quantitative solubility measurements (32). The issues in measuring protein solubility may relate to the formation of gel-like or supersaturated solutions in those methods that require the increase in protein concentration (33), or the difficulties in comparing results from different experiments when taking advantage of protein precipitants (31). As a consequence, solubility remains a protein feature not well understood, and still remains a major issue in the detailed structural and functional characterization of many proteins and isolated domains (34, 35). The dissection of the phenomenon is of critical importance not only for its role in protein homeostasis (36, 37), but also because insoluble regions in proteins tend to aggregate (37), leading to a variety of diseases such as Alzheimer's (38) and amyloidoses (39). Aggregation as a flip side of low protein solubility also represents a biotechnological complication. Soluble expression remains a serious bottleneck in protein production (40) and low solubility in drugs may make them ineffective (41) or even toxic (42). Targeted mutagenesis, usually without affecting protein structure or function, has been demonstrated in a number of cases to be a valuable tool to alter protein solubility (43). Especially in the absence of structural knowledge, the identification of residues to mutagenize benefits from dedicated prediction methods. In addition, predictors can contribute to the identification of pathogenic mutations in solubility-related diseases (44, 45).

1.3 Relationship between protein sequence, structure and function

The complex nature of protein sequence/structure relationship has just been introduced in the previous sections. A central challenge in biology is to rationalize the mass of biochemical and biophysical knowledge about proteins collected through very different methods in order to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences. The connection between all this "levels of information" is not so straightforward. This is mainly due to the fact that protein sequences are subject to several disparate evolutionary pressures. They have to fold fast and smoothly, avoiding incorrect and non-functional alternative structures or even the

1. INTRODUCTION

formation of pathogenic aggregates. They need to acquire a certain level of specificity in their function, however the very nature of evolution tends also to be conservative and efficient by re-using structural and functional elements that are already available, eventually modifying their specificity or their context to perform new function. All this forces, sometimes conflicting, complicate the protein sequence-structure-function relationship making it difficult to predict despite its deterministic nature.

1.3.1 Sequence-structure-function paradigm

At the center of the classical biology, lies the idea introduced in the previous section that a protein function depends on a well-defined 3D structure, as the unique spatial pattern of properly placed amino acids residues creates a special physico-chemical microenvironment tailored for the tight and extremely specific interaction with the environment. So, the detailed description of this structure hold the key to understanding the protein role. In turn, the structure is perfectly encoded in the protein sequence as a specific pattern of amino acids is driven through the folding funnel to acquire a specific folded state. However, to fully understand the nature of protein sequence-structure-function relationships, two concepts must be introduced to extend the paradigm:

- The evolutionary relationship between protein sequences (presented in this section).
- The insufficiency of the mantra "function requires (globular) structure" which came into light in the last decades (introduced in next section, [1.4](#)).

1.3.2 Sequence evolution

The relationship between two different proteins may be assessed by their sequence similarity. Very similar proteins in terms of sequence should have the same structure and therefore function. They probably share a recent ancestry, thus they are defined *homologous*. Instead, proteins showing very different structure have usually different 3D arrangements and role. However, this relationship is not linear. Early studies by Chotia and Lesk ([46](#), [47](#)) showed a strong non-linear relationship between sequence and structural similarity in 346 homologous proteins. Very similar sequences showed modest structural differences, but structural differences increased dramatically as sequence identities dropped below 15-20%. The signal of similar structure get blurred in the

1.3 Relationship between protein sequence, structure and function

twilight zone of 20-35% of sequence identity, while the 40% has been established as a threshold to discriminate similar and non-similar protein structures (48). The relationship to function is even more complex, due to the phenomenon of *convergent evolution*, which is the independent evolution of similar specificity, thus function (49). Proteins with distinct three-dimensional fold and sequence pattern may evolve to perform the same function, and even acquire structural similarities which are not directly related to a common evolution. The complexity of the protein sequence-structure-function relationships is the reason why the problem has not yet been solved, as we are not able to predict the structure and function of all proteins based only on their sequence data (50, 51). The general picture is complicated even more from the fact that, as the French biologist Francis Jacob said (52), "nature tends to re-use the old working solutions on new entities". This happens at all levels of protein evolution: at the level of the functions, the basic pathways of all living organisms are the same and they are determined and regulated by the same features (53), at the level of the structures, the wide diversity of folds in existence today have probably evolved from the combination of peptidic ancestors (54, 55, 56), and at the level of the sequence, where the functional units, the "domains", have been recycled in many different proteins which similarity is still detectable (57).

1.3.3 Protein families

As introduced in section 1.1, a domain is a part of the protein structure which often can be independently stable and folded (58). Domains are characterized by a compact structure (59), autonomous folding (60) and independent function and evolution (61). Nature often brings several domain together to form multi-domain proteins, combining their function (61). Such as secondary and tertiary structure modules which serve as building blocks for protein structures, on a higher level domains serve as building blocks for protein function. Many domains in eukaryotic multidomain proteins can be found as independent proteins in prokaryotes (62), suggesting that domains in multidomain proteins have once existed as independent proteins. The sequence similarity between different versions of the same domain in evolution are still trackable, and base the classification of proteins provided by Pfam database (63). Groups of related protein sequences, named *families*, are collected in the protein families database, Pfam (64). Pfam (www.pfam.xfam.org) classifies protein families detected through protein

1. INTRODUCTION

sequence analysis via the HMMER package (65) from a seed alignment to produce a Hidden Markov model (HMM) representing the protein family. HMM are statistical models based on multiple alignments and particularly well-suited to capture the differential variability of protein sequences. Pfam is a collection of HMMs, where each represents a particular non-overlapping protein family. Families are also grouped into clans in order to establish evolutionary relations between large, divergent families (64). Pfam is based on the concept that the relationship between protein sequences can be identified by their similarity, or homology, which is a hint of common evolution. These evolutionary-related groups of protein sequences often share three-dimensional structures and functions. The same concept lay the basis for all existing methods for protein structure and function annotation by homology based inference (66).

1.3.4 Domain architecture

The evolutionary unit in proteins is the domain. Multiple domain arrangements arise from events such as recombination, exon-shuffling, gene fusion, domain loss (57). The representation of protein sequences as sets of ordered functional domains is termed "protein architecture" and provides a useful way of investigating protein evolution. Multidomain proteins are likely to have emerged from selective pressure during evolution to create new combinations of functions (61). Various proteins have diverged from common ancestors by different combinations and associations of domains. The domain organization therefore is an advantage both in protein folding, with each domain being able to individually fold and accelerating the process, and in the diversification of protein roles. Through the combination of different domains, some proteins are involved in structural support and movement, others in enzymatic activity, and still others in interaction with the outside world. Indeed, the functions of individual proteins are as varied as their unique amino acid sequences and complex three-dimensional physical structures.

1.3.5 Protein function

Protein function is how a protein interacts with other molecules in the cell environment and the consequences of this interaction. A critical function of proteins is their activity as enzymes, which are needed to catalyze almost all biological reactions. The most intuitive way to envisage the interaction between a biological catalyst and its substrate

1.3 Relationship between protein sequence, structure and function

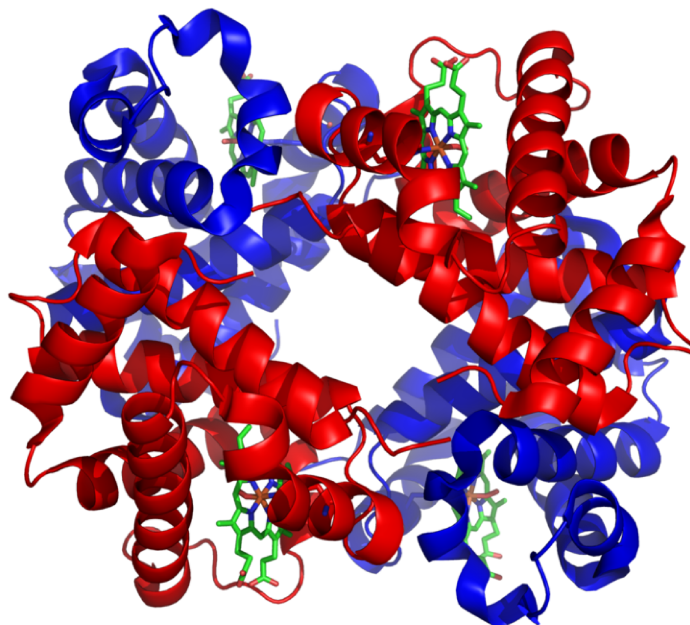


Figure 1.4: 3D structure of hemoglobin, a globular protein. Licensing: GNU Free Documentation License. PDB; PDB ENZYME.

is to imply a strict geometric complementarity between the two partners. This idea was firstly postulated in 1894 by Emil Fischer (67), and it is referred to as the *lock-and-key model*. In this analogy, the lock is the enzyme and the key is the substrate. Only the correctly shaped key (substrate) fits into the key hole (active site) of the lock (enzyme), triggering the reaction. In this model, the nature of amino acid side chains in the vicinity of the active site, in addition to its shape, are determinants for the establishment of the interaction. This framework is relatively simple but it implies an absolute specificity between the two partners. On the contrary, enzymes have varying degree of specificity, from being able to interact with one substrate and no others to multiple substrates with similar functional groups, side chains, or positions on a chain. The least specific enzymes catalyze a reaction at a particular chemical bond regardless of other structural features. To explain the adaptability of these structures, the lock-and-key model should be expanded. This description of protein nature is

1. INTRODUCTION

only a simplification of how proteins work in the cell, as a static picture of how they function in a specific moment, environment and condition. Such as a single photo shoot compared to a video, this picture is reductive compared to the whole ensemble of protein states in space and time, during its life-cycle and activity. Protein structures are, nevertheless, highly dynamic, and their biological functions depend intimately on this. The dynamics or motions in a protein allow its conformation to change and respond to the presence of other molecules and/or to variations in the environment (68). Biological and biochemical processes such as signal transduction, antigen recognition, protein transport and enzyme catalysis rely on this ability to change conformation or to adapt to change. In order to introduce this framework in the lock-and-key model, Daniel Koshland proposed the "induced fit" model (67). In this theory, the active site and the binding portion of the substrate are not exactly complementary, however, the active site is flexible and can remodel its shape until the substrate is completely bound. Once the reaction is completed, the reaction products will move away from the enzyme and the active site returns to its initial shape. Even if still simplistic for the vast ensemble of protein interaction mechanisms, this theory introduces a fundamental concept for the study of protein interactions: the two partners can recognize each other even in cases of imperfect structural compatibility. In addition, protein function is not limited to enzymatic activity. Other roles require fast adaptability, such as antibodies. Antibodies bind to specific foreign particles (e.g. belonging to infectious agents) to recognize them and trigger the protective reaction of the body. Other proteins are messenger proteins, such as some types of hormones. This protein specie is exploited to quickly and efficiently transmit signals to coordinate biological processes. Other proteins are incorporated in the hydrophobic environment of cell membranes, or function as "gates", or "pores", to control the flux of materials between different compartments. Moreover, some macromolecules are used for transport and storage of atoms and small molecule throughout the body. Finally, a large number of proteins provide structural support inside and outside of the cells. These "structural proteins" are often active not only as structural components but also as platforms for PPIs. By looking at the variety of protein roles, it becomes evident that the protein universe includes much more than the case of the spherical-like, water-soluble, relatively rigid and highly specific globular enzyme as it was the first protein ever crystallized, the hemoglobin (Figure 1.4).

1.4 Non-globular proteins

Non-globular proteins (NGPs) encompass different molecular phenomena that defy the traditional view of the sequence-structure-function paradigm. NGPs include intrinsically disordered regions, tandem repeats, aggregating domains and transmembrane domains. These protein species will be presented in this section. Although growing evidence suggests that NGPs are central to many human diseases (69, 70), and more in general to the evolution of complex organisms (71, 72, 73, 74) functional annotation is very limited (75, 76, 77). This is mainly because traditional methods for the description and classification of proteins were designed tailored on globular proteins, basing on their typical sequence conservation and structural features. Despite all the efforts done in order to increase the quality of these methods, indeed, the description of eukaryotic and prokaryotic proteomes remained far from completion. It was estimated that about 50% of all residues in the human proteome lack Pfam annotation (77), which is annotation transferred by sequence homology. Even lower is the coverage in terms of structural annotation, as only 4% of the eukaryotic proteome features a detailed structural description and about half of it is inaccessible to homology modeling, that means, that absolutely no information is available or derivable about their 3D arrangement (78). Due to this limited coverage of existing methods, it became evident that they were simply not suitable for the description of all biological phenomena. To fully comprehend human molecular physiopathology and biology as a whole, a better understanding of NGPs is crucial, as it is the development of specialized method for their description.

1.4.1 Tandem Repeat Proteins

Tandem Repeat (TR) proteins convey the least complicated relationship between a sequence and the corresponding three-dimensional structure. Indeed, they consist of repetitive sequence stretches ranging from less than 5 to more than 60 amino acids (79). These give rise to a modular protein structure composed by the repetition of the same structural unit. A repeat "unit" is defined as the smallest structural building block forming the repeat region. The repeat region may include insertions, i.e. non-repeated segments occurring either inside a single repeat unit or between consecutive repeats (80). The protein repeat sequences can be described by two parameters: period

1. INTRODUCTION

„Linear” or „open” solenoid domain

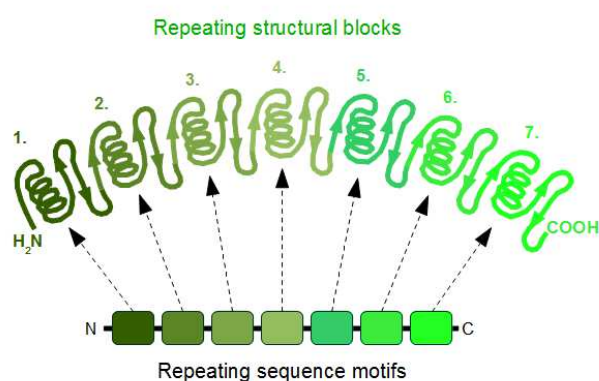


Figure 1.5: Schematic of a linear (or open) solenoid protein domain. License: Creative Commons Attribution 3.0 Unported. Author: Bubus12.

and number of units/repetitions (see Figure 1.6). The period (or repeat length) is the number of amino acids contained in each repetition. In addition, from the multiple alignment between the unit sequence, it is possible to derive the consensus sequence, which is the representative model of the repeated sequences.

1.4.1.1 Solenoids

Scientific literature about repeat proteins is dominated by a specific type of repeats: the solenoids (81, 82). Even if the complete classification of TRPs will be presented lately in the present thesis, here the specific class of solenoids is presented as a case-study of repeat structure, mechanism of function and evolution. The simplest types of solenoids contain hairpins with two elements of secondary structure (α/α , β/α , β/β), one flanking the other in such a way that the start and end of the unit fall on the same axis and the succession of units can continue along that axis (81). More complex structures include units with three or four elements of secondary structure, or curved axis along which the superhelix elongates. Figure 1.5 shows the schema of a linear solenoid, with repeated β - α - β units. The resulting shapes allow the formation of diverse interfaces for interaction, as well as cooperative multivalent interactions. In addition, the overall

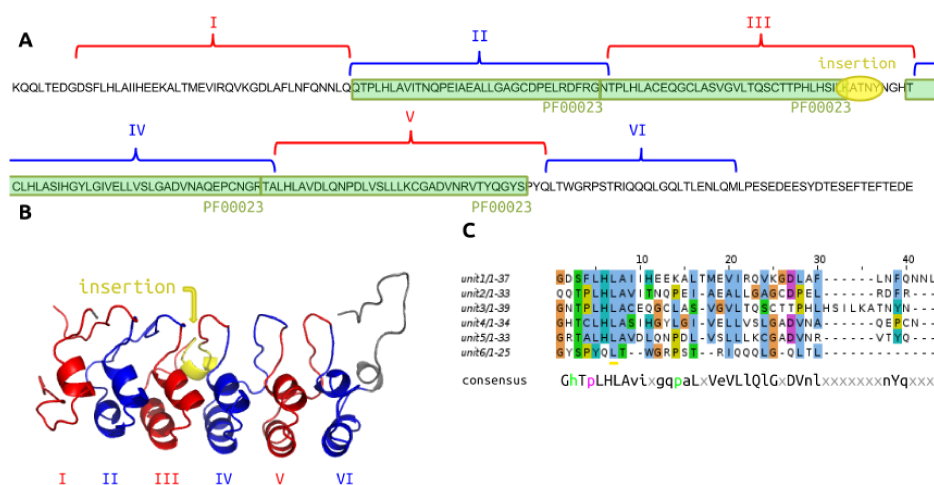


Figure 1.6: The $I\kappa B\alpha$ protein (PDB code: 1lkn) containing six ankyrin repeats is used as an example. (A) The units, the insertions and the Pfam domains are mapped on the 3D structure sequence. (B) The structure is colored to highlight the units in red and blue. The insertion is showed in yellow. (C) Sequence alignment of the units with the derived consensus sequence. (87)

curvature influence the binding properties as superhelices with a large curvature can host other macromolecules (83) while extended superhelices with flat axis can function as platforms for multiple interactors or accomodate extended peptides (84). Thanks to this binding properties this simple arrangement was widely used by evolution in a number of pathways (85), especially when rapid evolution yet specialization is implied (86).

1.4.1.2 Folding

TRPs are stabilized by a regular patter of intra- and inter-unit interactions, so mainly local contacts. By contrast, the stability of globular proteins originates from the high cooperativity between sequence-distant interactions and the burial of a large hydrophobic surface area. Very small repeat proteins (about 100 amino acids) show an almost direct unfolded to folded transition, such as globular proteins of similar sizes (88, 89), however longer repeats are characterized by a different folding pathway than the typical globular polypeptides. They usually fold slower, with a vast population of folding intermediates, and following a linear path where each unit may drive the folding of the next one (90), although this effect acts locally. The low cooperativity of TRPs

1. INTRODUCTION

folding is often exploited for their function. A striking example is the interaction between the transcription factor NF- κ B and the 6-ankyrin-repeat protein I κ B α (in Figure 1.6), which regulates NF- κ B by sequestering it in the cytoplasm. This solenoid repeat completes the folding of the two C-terminal units only upon binding to NF- κ B, in a process which was shown to be critical for high-affinity binding. In addition, the different stability of I κ B α when it is bounded or not to NF- κ B was demonstrated to play a role in transcriptional regulation (91). As exemplified by the case of I κ B α , the specific properties of TRPs structure (discussed in section 4.1.2) are due to the mechanism of their folding and motivate their widespread distribution among organisms.

1.4.1.3 Organism distribution

TRPs are prevalent in eukaryotes, but also present in Bacteria and Archaea (79). The higher number in eukaryotes suggests that TR protein development has been an important process during evolution of multicellular organisms (92, 93). It has been pointed out (73) that most eukaryotic repeat proteins have few similarities with prokaryotic ones, suggesting that they arose after the two lineages diverged. While prokaryotic TRs usually intervene in specialized secretion systems and pathogen virulence factors (73), the most frequent classes of repeat proteins perform functions unique to eukaryotes. In particular, eukaryotic TRs are often hubs in the PPI networks, i.e. they are characterized by high number of interactors. This concept is extensively discussed in this thesis in section 4.1.2. The reason can be summarized as follows: the repetition of the same architectural domain is the most common evolutionary strategy for the design of long and extended structural proteins, which are often exploited as PPI platforms. However, the importance of repeat proteins relates to several other biological functions. The elasticity and antigenicity stand out as functional properties. Repeat proteins are involved in processes such as biomineralization (94), adhesion (95), ice crystallization (96), and pathogenesis (97, 98). Some repeat structures are incorporated in membranes as pores (99). Different structural properties and, consequently, function are observed based on the repeat length. For this reason, TRPs have been classified basing on this parameter (79).

1.4.1.4 Identification

A large-scale analysis of repeat protein features would require a large set of protein sequences, or structures, identified as repeated and classified. However, the identification of TRPs represents an issue both for sequence-based and for structure-based methods. TRPs evolve quickly while maintaining their fold, hampering detection by traditional methods for sequence analysis. The same holds for modeling and functional characterization, which usually relies on well-conserved sequence features. As a result, specialized methods were built for the identification of repeat proteins (100). Sequence-based strategies, based on homology search (101) or domain assignment (102, 103), mostly underestimate TRs due to the presence of highly degenerate repeat units (77). A recent study to understand and improve Pfam coverage of the human proteome (77) showed that five among the ten largest sequence clusters not annotated with Pfam are repeat regions. Alternatively, methods requiring no prior knowledge for the detection of repeated substrings can be based on self-comparison (104, 105), clustering (106, 107) or hidden Markov models (108, 109). Some others rely on complexity measurements (100) or take advantage of meta searches to combine outputs from different sources (110, 111). Methods recognizing TR proteins based on the modularity of their 3D structure have also been developed (112, 113, 114, 115, 116).

1.4.1.5 Prediction of repeat units

Existing methods for TR protein identification do not deal with the TR structures classification problem. TR classification was originally described in (79) by manual inspection and based on the unit length and unit structural features, since these features determine the global arrangement of the repeat region. The identification of single repeat units has so far been addressed by few automatic methods, including ConSole (115), which exploits the modularity of protein contact maps, and TAPO (116), evaluating the periodicities of atomic coordinates and other types of structural representation. Both are available through a Web server interface that allows the user to evaluate one protein at a time, while the automatic identification of units inside a TR protein structure allows to scale up this type of information. This data could be a powerful tool not only to classify TR structures, but also to understand TR evolution

1. INTRODUCTION

and assess conservation at the sequence level, since the repeat unit is the TR evolutionary module. Moreover, the collection of an alphabet of TR units can also be useful for protein engineering applications (117). RAPHAEL (114) is a support vector machine classifier for the identification and classification of repeat structures. It reliably solves three problems of increasing difficulty: (1) recognition of repeat domains, (2) determination of their periodicity and (3) assignment of insertions. RepeatsDB (80), database of repeat protein entries, was built through systematic annotation of the PDB by RAPHAEL. The database entries were annotated with units and classified based on repeat length and structural features, as proposed by Kajava (79), through manual curation. RepeatsDB classification will be described in section 4.2.

1.4.1.6 Evolution

Also due to the difficulties in TRPs detection, the characterization of their evolution is still an open problem. The mechanisms behind the expansion of internal repeat duplication and level at which these duplications occur are not well understood. Different mechanisms appear to be involved in the origin of different repeat types (92). Moreover, right after the generation of a repetition, two properties come into play and influence TRP evolution. Repeat segments have an intrinsic tendency of to self-propagate at the DNA level, generating further repetitions (118). This property characterizes both coding and non-coding repeat sequences and was demonstrated to accelerate the evolution of a functional genome (119). On the other hand, arrays of perfect tandem repeats falling into coding sequences are correlated with disorder and aggregation and found to be counter-selected (120). Jorda et al. suggested that immediately after the duplication, consecutive sequence repetitions are prone to differentiate, in order to avoid erratic pairing and ensuing structural misfolding. The tendency to a diversification process would explain TRPs sequence diversity versus structural conservation (79). Indeed, the symmetry exhibited by repeat protein structures is encoded by symmetric signals hidden in irregular sequences. These mechanisms are considered sources of hypermutability and have given rise to a high polymorphism rate compared with the background rate of point mutations (79, 121). The combination of the two discussed properties makes TRPs them a vast source of genome variability (122), and explains why TRPs have been especially exploited for functions which require quick adaptability.

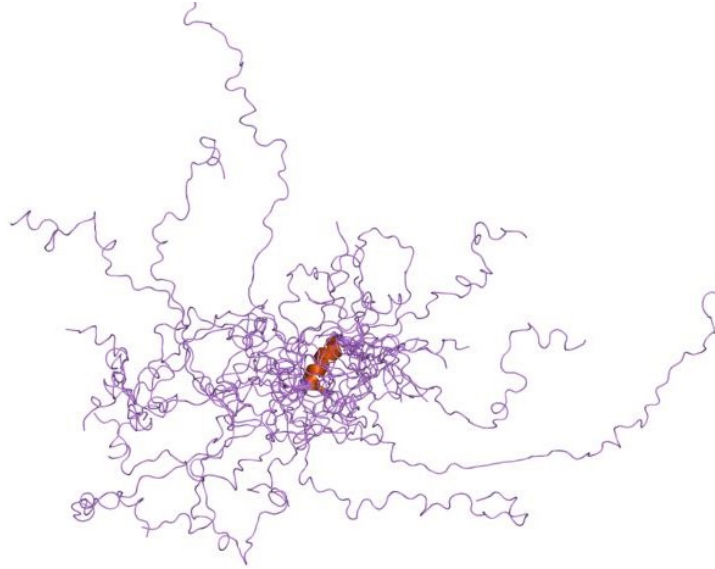


Figure 1.7: An ensemble of NMR structures of the Thylakoid soluble phosphoprotein TSP9, which shows an intrinsically disordered protein chain. Licence: Public domain. Source: <http://www.ebi.ac.uk/>.

1.4.2 Disordered proteins

Intrinsically disordered proteins (IDPs) and regions (IDRs) are devoid of order in their native state (123, 124) (see Figure 1.7). More specifically, IDPs cover a spectrum of phenomena from fully unstructured to partially structured including random coils (fully unstructured), (pre-)molten globules, and flexible linkers connecting multi-domain proteins. Intrinsic disorder is prevalent in the human proteome (125), appears to play important signaling and regulatory roles (126) and is frequently involved in disease (127). The discovery of intrinsic disorder and its prevalence and functional importance is transforming the field of molecular biology, as it demonstrated the insufficiency of the assumption that proteins become functional by assuming a well-defined structure.

1.4.2.1 Sequences

IDPs properties are encoded in their peculiar sequences. IDRs are characterized by high content of polar and charged amino acids, thus they are usually depleted in hydrophobic residues (127). Thanks to this specific composition they are prone to interact with the solvent, differently from globular domains which fold driven by the hydrophobic

1. INTRODUCTION

effect. Furthermore, their internal charges promote disorder because of electrostatic repulsion resulting from equally charged residues (128). Two independent studies to assess sequence composition of intrinsically disordered proteins, focusing respectively on the amino acids preferred at the surface of globular proteins or on those found less frequently in secondary structures (129, 130), identified three disorder-promoting amino acids, namely Glycine, Serine and Proline. The latter might be surprising, due to its hydrophobic side chain. However, Proline is a very atypical amino acid since the backbone has hydrogen bond acceptors but no donor, and for this reason it is costly from an energetic point of view to sequester it from the solvent. Proline has a role in disorder as secondary structure breaker, to prevent amyloid-like aggregation, assembling the peculiar polyproline type II helices and as a fundamental component of elastomeric proteins (131). The role of other specific residues in protein disorder, namely Glutamic acid (132) and Serine (133).

1.4.2.2 Function

The early emphasis in the field was on proteins that are mostly or fully disordered, namely MAP2 (134), tau (135), Myelin basic protein and α -synuclein (136). These proteins escaped the characterization by the dominant experimental methods such as X-ray crystallography, and the experimental challenges related to them attracted attention to the phenomenon. However, the protein universe is far more complex than the simplistic semantic separation between "structure" and "disorder". All proteins have some movements, and no protein is completely chaotic (137). The term "disorder" may come into play when a protein lacks tertiary structure, but it may include regions without any regular secondary structure or regions with transient secondary structure elements which can switch from one type to the other (e.g. α -helix or β -sheets to coiled coils). Furthermore, the definition of a protein as intrinsically disordered is largely supported by its mechanism of function. IDRs functional advantages are provided by (137):

- **High entropy.** IDRs are characterized by inherent dynamic movement, which create a less restricted space.
- **Accessibility.** Site accessibility is essential in binding of other molecules and for the post-translational modification (PTM) of the protein.

- **Plasticity.** IDRs may respond to reaction with other molecules by changing shape, even becoming more ordered, and triggering other reactions.

Thanks to these properties, disorder may provide the necessary mobility to a flexible linker connecting domains (129), facilitate the different conformational requirements for binding the modifying enzymes as well as their receptors in PTMs (138), undergo transitions to more ordered states upon binding to their targets (139) and host short linear motifs (SLiMs). SLiMs are short disordered segments of proteins that mediate functional interactions with other proteins or other biomolecules (RNA, DNA, sugars etc.). Depending on the partner proteins that recognize them, these sites can facilitate a diverse set of functions including targeting a protein to a specific subcellular location, determining the modification state of a protein, controlling the stability of a protein, and regulating the context-dependent activity of a protein (140). The biophysical mechanisms exploited by IDRs are of fundamental importance in protein signaling and regulation (137). Signaling pathways consist in a cascade of interactions that trigger some activity in the cell, usually characterized by high speed and precision (141). Regulation at the cellular level involve IDRs in gene transcription (142), protein degradation (143) as well as through allosteric effects or PTMs which may result in the fast masking and unmasking of interaction sites. The speed and precision of IDRs binding derive from their

- **Low affinity.** The entropic cost of IDRs binding, which restricts their degree of freedom, is usually high and makes the interaction transient.
- **High specificity.** The binding surface is usually high with respect to the IDR volume, and it responds to the binding by shaping in such a way that the interaction becomes highly specific.

Furthermore, several newly recognized functional mechanisms (144) have been added to IDRs functional classification. For example, the central role of intrinsic disorder in the formation of membraneless organelles, such as nucleoli and stress granules, by liquid-liquid phase separation has been characterized recently (145, 146, 147, 148).

1. INTRODUCTION

1.4.2.3 Detection

A wide range of experimental observations on the structure-function relationship of IDPs/IDRs is furthering our understanding of disordered states and of the manners in which they function (149, 150, 151). As intrinsic disorder is emerging as a general phenomenon, several different methods have been developed for the detection of intrinsic disorder. Experimental approaches for the study of IDPs have been collected and discussed (152, 153). In practice, multiple techniques are usually employed to aggregate evidences intrinsic disorder. In addition, a number of bioinformatics methods have been developed to extract disorder information from protein sequences taking advantage of their peculiar features (154, 155). Databases are collecting and presenting disorder related data in a systematic manner. One of the major repositories for experimentally determined disorder is the DisProt database (156), containing manually curated information on IDPs from the literature. Although invaluable as a gold standard, DisProt being manually annotated represents only a fraction of the known protein sequences posing a bottleneck for large-scale analysis of intrinsic protein disorder. Several disorder prediction methods where designed to overcome this limitation. A comprehensive resource that collects several sources of disorder annotation is MobiDB (157), providing consensus predictions and functional annotations for all UniProt proteins.

1.4.3 The relationships between repeats and disorder

Even if for different reasons, TRPs and IDPs share the binding character. Both the protein species are hubs in the protein-protein interaction network, and their high connectivity makes them essential as well as potentially lethal in case of deletion, a phenomenon known as the centrality-lethality rule (158). This is the reason why the description of their mechanism of action and inherent properties is extremely important not only to characterize essential functions in organisms but also to understand their molecular physiopathology. On this line, it is important to note that the typical TRP and the typical IDP are hubs in a different way. An extended molecule such as the spectrin repeat, which serves as a platform for cytoskeletal protein assemblies (159), is characterized by multiple interactions at the same time and this is an essential aspect of its function. Molecules like spectrin are termed "party hubs", or static hubs, because their interactions are not exclusive. On the other side, a highly promiscuous protein

such as the tumor suppressor p53 (160), with disordered tails, has an incredibly high number of interactors considered its limited length. The number of partners is justified from its disorder properties, so it has the low affinity but high specificity necessary to easily adapt to a new partner, quickly and efficiently. This is a "date hub", or dynamic hub, characterized by a number of alternative interactions, influencing each other in a dynamic way. The link between repeats and disorder is not limited to their binding properties. As already presented in the example of NF- κ B (91), the peculiar folding landscape of TRPs is characterized by cases where some units are unfolded in the native form. With a number of semi-stable intermediates and a linear pathway where each unit stabilizes the following one, the folding funnel of repeats can be represented as much shorter and wider than the globular one in terms of minimization of the free energy and entropy (161), therefore it is more similar to the one of disordered proteins (162). The yet not fully characterized overlapping between the two phenomena has been observed in correlating perfect repeats with the tendency to be unstructured (120), in observing how amino acid repeats accumulate in disordered regions of proteins (163) and how they are at the basis of their evolution (124).

1.4.4 Low complexity

The difference between globular and non-globular polypeptides has its origins at the sequence level. The easiest difference to identify is the sequence composition. NGP composition in terms of amino acids is indeed biased towards specific properties, which influence their folding and stability, e.g. the previously discussed case of IDPs. The regions enriched in a specific type of amino acid are compositionally biased, or characterized by Low Complexity (LC). According to conservative estimates, Low Complexity Regions (LCRs) represent the 20% and 8% of all known sequences of eukaryotes and non-eukaryotes, respectively (164). However there has been an early reluctance to consider these regions for biological studies, mainly due to their "annoying" statistical features. Due to the high redundancy of these sequences, tracing their homology is a very difficult task. Indeed, in homology-based database searches low-complexity stretches are often masked to avoid spurious alignments (165). Only recently, there is an intensification of research on LCRs - e.g. (126, 127, 166, 167), reminiscent of the paradigm shift that brought non-coding RNAs to the forefront of genomics research in the recent past. In the definition of LCRs, multiple concepts related to sequence

1. INTRODUCTION

composition, periodicity and structure have been used (Table 1.1). Regarding amino acid composition, there is no consensus about which metrics are the most appropriate to measure the different related phenomena. An extended discussion about LCR detection and related properties is included in this thesis as section 4.4.

Term	Definiton
Definition based on amino acid composition	
Low complexity region (LCR)	Regions with a skewed amino acid composition
Compositionally biased region (CBR)	
Definition based on amino acid periodicity	
Repeat motif	Reiteration of residues: $(\dots)_n$
Homorepeat (polyX)	Consecutive runs of a single residue: $(X)_n$
Direpeat	Consecutive runs of two ordered different residues: $(XY)_n$
Tandem repeat	Pattern of residues which are directly adjacent to each other: $(XYZ)_n$
Cryptic repeat	Scrambled arrangements of repetitive motifs
Imperfect repeat	Regions in which the repeat units are not the same
Definition based on structure	
Intrinsically disordered protein (IDP)	Protein that lacks a fixed or ordered 3D-structure
Coiled coil (CC)	Structural motif characterized by a seven-residue sequence repeat in which alpha-helices are coiled together to form an extended rope-like structure: $(a-b-c-d-e-f-g)_n$
(Charged) single alpha-helix ([C]SAH)	A segment forming stable monomeric alpha-helix in aqueous solution, typically rich in Arg/Lys/Glu forming an alternating pattern of short runs of oppositely charged residues
Protein flexibility	Ability of a protein to fold into multiple stable 3D-structures
Amyloid fibrils	Stable insoluble protein assemblies composed predominantly of beta-sheet structures in a cross-beta conformation

Table 1.1: Overview of complexity terms and their definitions.

1.4.4.1 Structural features

Firstly, the concept of LCR is intermingled with the concept of sequence repeats. While measuring the complexity of a sequence, if the considered window includes multiple copies of a repeat the sequence will result redundant, and thus low complexity. Shorter repeats will be more easily detected as low complexity, an extreme case of minimal complexity is represented by tracts of a single repeated residue, known as homorepeats (120). Regarding protein structure, LCRs mostly have a non-globular conformation (168). Factors such as the sequence context (features present in the flanking regions) and the molecular context of the protein (e.g. interacting proteins, cell tissue or state when it is expressed) can influence their structural state. This landscape is complemented by emerging concepts such as structural repeats (discussed in section 1.4.1),

intrinsic disorder (discussed in section [1.4.2](#)) and aggregation or protein phase separation, all formalized in the literature (see e.g. ([169](#), [170](#), [171](#), [172](#))).

1. INTRODUCTION

Personal Contribution and Thesis outline

The leading thread of my thesis consists in the computational description of structure and function of non-globular proteins (NGPs). Several state-of-the-art resources for the collection of NGP sequences and structures were built in the Biocomputing UP Lab, where I carried out my research. These resources include RepeatsDB (173), a database of tandem repeat protein structures, and MobiDB (174), a comprehensive resource for protein disorder annotation. During my PhD, I contributed to the development of these resources and related tools and methods. The main focus of my research is tandem repeat proteins (TRPs), although I also exploited the knowledge acquired in the field to contribute to the study of other NGPs. Chapters 3 and 4 are mostly based on my PhD publications (see "List of publications").

2.1 Tandem repeats

Tandem repeat regions in proteins represent the first focus of my PhD research. In order to characterize TRP properties in terms of function, role in protein-protein interaction (PPI) networks and association to diseases, I firstly performed a computational analysis of extensive datasets of TRPs derived from RepeatsDB, presented in section 4.1.2. I showed that there is a significant association between TRPs and human diseases and that it can be explained by their role as hubs in PPI networks. Indeed,

2. PERSONAL CONTRIBUTION AND THESIS OUTLINE

TRPs extended and modular structure makes them perfect candidates to serve as PPI platforms. As a case-study of this phenomenon, I dissected the interactome of Collagen V, a repeat protein associated to Ehlers-Danlos syndrome (EDS), in order to identify genotype-phenotype correlations in relation to the interaction network model (section 4.1.3). This work was published in (175). I personally contributed to the project by collecting and analyzing the data. Moving from a single example to the characterization of the different repeat types, I analyzed the classification of TR included in RepeatsDB. RepeatsDB defines five main classes, mainly based on repeat unit length, with subclasses representing specific structural arrangements. In section 4.1.1 I compared this data to the one in Pfam database, which provides an alternative classification based on evolutionary conserved repeat families. The comparison was published in (87). This work highlighted that the goal to completely characterize all human proteins in terms of their domains cannot be reached without an effort to improve TR recognition and classification. Starting from this observation, I moved to the curation and improvement of RepeatsDB database. A detailed structural characterization of repetitive elements was largely missing, as repeat unit annotation in RepeatsDB was manually curated and covered only 3% of bona fide TRPs at the time. This is the reason why we developed Repeat Protein Unit Predictor (ReUPred, algorithm described in section 3.1.1, results described in section 4.2.1), a novel method for the fast automatic prediction of repeat units and repeat classification using an extensive repeat unit library derived from curated data in RepeatsDB. ReUPred, published in (176), uses an iterative structural search against the library to find repetitive units on target structures. My contributions to the development of the predictor included the identification of challenging cases to test and improve ReUPred performances, as well as the tool benchmarking. As a following step, we published the second release of RepeatsDB database (173). RepeatsDB 2.0, discussed here in section 4.2.2, features information on start and end positions for the repeat regions and units for all entries, a substantial growth of repeat unit characterization that was possible by applying the ReUPred algorithm over the entire Protein Data Bank (PDB). RepeatsDB is continuously updated, and therefore requires a continuous effort in the manual curation. To facilitate this process we designed RepeatsDB-lite, web server for the prediction and refinement of TR in protein structure (algorithm described in section 3.1.2, results in section 4.2.3), published in (177). It takes advantage of ReUPred algorithm and an extended library that covers

all different TR classes. The web server allows an intuitive revision of the prediction and submission of reviewed entries to RepeatsDB database. It represents a platform to harness community annotation efforts, which have been proven to be effective in RepeatsDB experience. Both in the case of RepeatsDB and RepeatsDB-lite, my contributions include the conception, design and implementation of the data structure, data management server and user-friendly web interface. RepeatsDB project aims at the definition of best-practices in the annotation of repeat proteins, and represents a powerful resource both in terms of structure and sequence data. We recently established a collaboration with Pfam authors, aimed at the improvement of existing Pfam domains and the creation of accurate repeat models based on structural information. At the moment of writing, I am involved in the project as a curator of Pfam models and I am collecting guidelines for the specific definition of repeat entries, which I included in section [4.2.4](#).

2.2 Intrinsic disorder

Another database that represents a central resource for the scientific community working in the field of NGPs is MobiDB. MobiDB is a database of protein disorder and mobility annotations that describes several aspects of NGPs structure and mechanism of function, which has provided a major contribution to the field by providing consensus predictions and disorder annotation for all UniProt proteins and it is also linked from the UniProt entry page. I contributed to the development of the new release of the database (section [4.3.1](#)), which provides both disorder type and quality of the disorder evidence, derived from the annotation source. Indeed, the data in MobiDB is based on the source quality, comprising manually curated data, annotations derived from experiments and annotations derived from predictions. The main disorder information in MobiDB is provided by a consensus combining all available sources prioritizing curated and indirect evidences over predictions. Predictions have been expanded to provide new types of annotation on backbone rigidity, secondary structure preference and disordered binding regions. MobiDB 3.0 was published in ([174](#)).

2.3 Low complexity sequences

A common feature of TRPs, IDPs and other NGPs is that they are characterized by a peculiar sequence which hampers their detection and analysis. In particular, several NGPs are characterized by low complexity (LC) sequences. I contributed to a critical review focusing on the definition of sequence features of LC regions and their connection with structure (section 4.4.1). At the moment of writing, the manuscript was submitted and accepted for review. We presented statistics and methodological approaches that measure low complexity and related sequence properties. We illustrated the dichotomy between low complexity in structural repeats and unstructured regions, and more generally the overlaps between different properties related to LC regions, providing meaningful examples. In this work, I curated the analysis of LC regions structural properties.

2.4 Solubility

Finally, I exploited the knowledge acquired in my studies regarding NGPs to build one of the first sequence-based methods for the prediction of protein solubility, SODA (algorithm in section 3.2.1, results in section 4.5.1). Solubility is an important, albeit not well understood, feature determining protein behavior. It is of high interest to the field of protein engineering, where similar folded proteins may behave in very different ways in solution. SODA (published in (178)) uses the aggregation propensity of the protein sequence as well as intrinsic disorder, plus hydrophobicity and secondary structure preferences derived from sequence features and complexity. SODA is able to evaluate solubility changes introduced by a mutation by comparing the profiles of the wild type and mutated sequences, and it is compatible with different types of variation including point mutations, deletions and insertions. The predictor is based on sequence features and allows therefore the large-scale screening of protein mutations. I contributed conceiving and implementing the web server, designed to allow large-scale annotation, and the user interface, which provides an intuitive form to guide detailed selection of mutations based on sequence solubility plot and, if the protein structure is given, residues accessibility to solvent.

3.1 Prediction of tandem repeat units in structures

The main focus of my PhD research was to identify, annotate and classify repeat protein structures. To achieve these goals we developed ReUPred, a predictor of repeat units in protein structures which results are described in section 4.2.1. ReUPred algorithm (176) allowed us to design a semi-automated detection and annotation pipeline of TR proteins, which steadily increased the number of TRPs annotated in our dedicated database, RepeatsDB (173), described in section 4.2.2. As the semi-automated annotation experience proved to be efficient in the classification and description of TRPs, we designed an easy-to-use web server for the prediction of TRs, RepeatsDB-lite (177). RepeatsDB-lite (results described in section 4.2.3) extends the ReUPred algorithm to all TR types and strongly improves the performance both in terms of computational time and accuracy. The following section describes the implementation details of the TR predictors.

3.1.1 ReUPred

ReUPred is a predictor for the classification of tandem repeat proteins and identification of the composing repeat units. The inputs are a target protein structure and the structural repeat unit library (SRUL). The output is a list of fragments corresponding

3. MATERIALS & METHODS

to the predicted unit positions in the structure and the class assignment according to RepeatsDB definition (80). The iterative algorithm decomposes the input structure using a template library. A divide and conquer strategy is used to improve both accuracy and speed, so the average calculation requires ca. 2 min on a standard laptop. ReUPred was optimized by filtering the SRUL and fine-tuning parameters in order to choose the best alignment and detect insertions between units. Each step is described in the following section. ReUPred is implemented in Python for Linux. The source code is distributed under the GPL license and freely available from the URL: <http://protein.bio.unipd.it/reupred/>.

ReUPred algorithm The algorithm exploits the evolutionary history of tandem repeat proteins. Solenoid units have been demonstrated to evolve from a single representative unit to multiple copies through repeated duplications (92). Units of a solenoid protein show a different degree of similarity, which is strongly correlated to the distance from the middle of the repeat region. This is consistent with the observation that units at the edges are more degenerated (73). ReUPred exploits this knowledge and tries to mimic evolution. The objective is to predict adjacent units, i.e., to minimize the number of residues between predicted flanking units, and obtain at least three repeated elements. This is important since in known RepeatsDB solenoid structures, insertions of non-repeat fragments are rare and mostly observed inside and not between units. See Figure 3.1 for a schematic description. ReUPred uses an iterative divide and conquer approach. Each iteration corresponds to a structural search, i.e., structural alignment of the query structure against all SRUL elements to identify a unit. The predicted unit corresponds to the aligned region in the query. At each cycle the algorithm forks (divides). Two new input structures are created, corresponding to the N- and C-terminal flanking fragments of the predicted unit and two new cycles (structural searches) are performed. After the first cycle, i.e., after the "master" unit is found, SRUL is no longer used. Instead, a new ad hoc library is created on the fly. At the beginning of the second cycle, only the "master" unit populates the ad hoc library and all newly predicted units are included for search in the following cycles. The algorithm stops when the entire input protein is consumed, i.e., new input fragments are too short, or the structural search does not provide any new valid alignment. The predicted units are then collected and evaluated together (conquer). If the result does not satisfy a

3.1 Prediction of tandem repeat units in structures

set of rules, the structural alignment filters for the "master" unit are relaxed and the entire iterative part is repeated from the beginning for up to four increasingly relaxed iterations. This strategy allows to predict both easy and difficult cases automatically. A valid solution for ReUPred is obtained when at least three units are found and their proximity in sequence is ensured by at least one of two simple rules to measure unit proximity: (i) the total number of gaps between units is less than 40 residues, (ii) the number of non-adjacent units divided by the total number of predicted units is less or equal to 0.25. Replacing the original SRUL with an ad hoc library from the second cycle onward improves both computational cost and accuracy. SRUL is quite large, with 997 unit templates. Instead, the ad hoc library reaches the maximum size at the end of the algorithm and corresponds to the number of predicted units, drastically reducing the number of structural alignments. On the other hand, using only units from the query structure itself increases the accuracy as these are structurally more similar to each other than units from other proteins. The class assignment is provided by simply reporting the classification assigned to the first "master" unit identified from SRUL. ReUPred accuracy strongly depends on the quality of the structural alignments at each cycle. In particular, it is very important to correctly predict the first "master" unit because errors propagate. Alignments have to abide a set of rules and constraints that are much more stringent for the "master" search compared to successive cycles. Structural alignments are calculated using TM-Align (179), filtering by TM-Score, RMSD, alignment length and number of gaps.

Iteration	TM-Score	RMSD (A)	Alignment (aa)	Unit gaps (%)
1	≥ 0.52	≤ 1.6	> 21	< 10
2	≥ 0.47	≤ 1.9	> 17	< 20
3	≥ 0.30	≤ 2.5	> 16	< 50
4	≥ 0.23	≤ 3.0	> 14	< 50

Table 3.1: Structural alignment constraints for the "master" unit. TM-Score and RMSD are the same provided by TM-Align. Coverage and gap are calculated as described in the manuscript. Different columns correspond to different algorithm runs that are performed on cascade until a valid solution is found.

Tables 3.1 and 3.2 list all cutoff values for the cascaded four runs used to select valid alignments for the "master" and "secondary" units, executed on cascade until a valid solution is found. The parameters for structural alignments have been optimized manually on the training set to maximize the number of repeat proteins, for which a valid

3. MATERIALS & METHODS

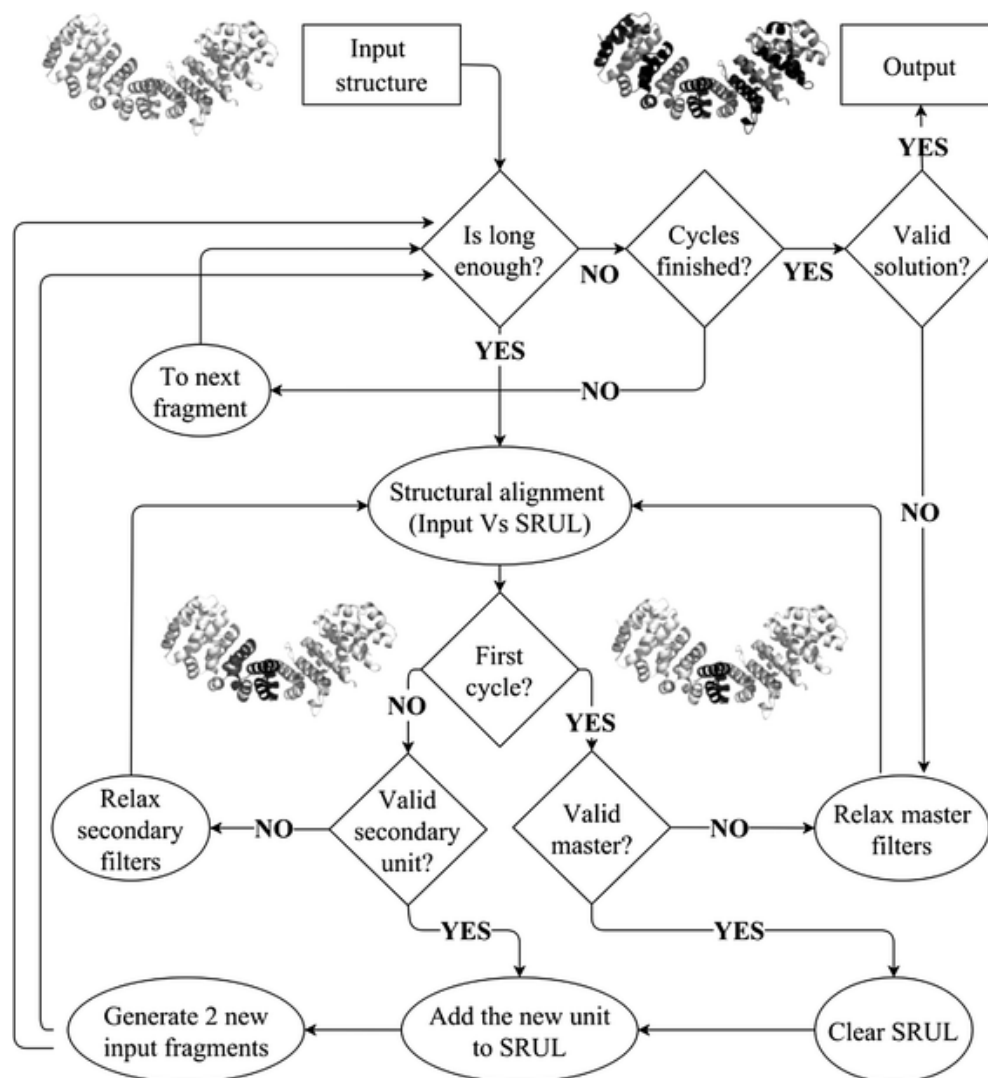


Figure 3.1: Schematic description of the ReUPred algorithm. The input structure (PDB 1IQ1, chain C) is processed iteratively until a valid solution is provided and no new fragments (subproblems) are generated. (176)

Iteration	TM-Score	RMSD (Å)	Alignment (aa)	Unit gaps (%)	Length ratio (%)
1	≥ 0.35	≤ 1.8	≤ 1.20	< 40	≥ 70
2	≥ 0.30	≤ 2.0	≤ 1.15	< 40	≥ 70
3	≥ 0.30	≤ 2.5	≤ 1.15	< 40	≥ 70
4	≥ 0.30	≤ 3.0	≤ 1.10	< 50	≥ 70

Table 3.2: Structural alignment constraints for the "secondary" units. Columns are as in Table 3.1. The length ratio is calculated as the unit length divided by the length of the first "master" unit.

3.1 Prediction of tandem repeat units in structures

Class	Units	Detailed	Classified	Predicted
β	367	41	128	-
α/β	180	19	70	-
α	388	48	875	-
Total	935	108	1073	7948

Table 3.3: RepeatsDB 1.0 solenoid dataset used in ReUPred benchmarking. Units list the number of single defined repeat units. Detailed proteins have the unit position identified manually. Those protein for which the subclass assignment is known are classified, including "manually" and "by similarity". The predicted proteins are not classified.

output is provided, and prediction accuracy, i.e., correct unit position assignment.

Repeat unit library and datasets The SRUL constitutes a fundamental part of the ReUPred input and represents the conformational space and diversity of bona fide repeat units. It has been generated by extracting all structural unit fragments from the "detailed" solenoid proteins in RepeatsDB 1.0 (see Table 3.3 for statistics). After filtering units shorter than 10 residues and larger than 90, the solenoid SRUL is composed of 916 structural unit fragments from 108 different proteins non-redundant at the sequence level. After clustering the sequences with CD-HIT (180) at 40% identity, 531 clusters are obtained. The largest cluster contains 17 units from 5 proteins and the others have less than 10 units each. From the structural point of view, SRUL is biased toward α -helical units. All-against-all structure similarity was measured by TM-Align (179). Clustering at 0.6 TM-score generates 362 clusters, where the majority of α units (319) fall inside a single cluster. Three different datasets have been used throughout this work. The training set has been generated from the "detailed" RepeatsDB 1.0 entries (108 proteins) and represents the reference for unit prediction evaluation. Since SRUL was generated from the same protein set, to benchmark ReUPred, all units coming from the target itself and all similar units (>30% sequence identity) were removed from SRUL at each benchmarking step. Another set with all "classified" and "by similarity" entries (1075 proteins) was used to test the ability to automatically classify repeat proteins and compare unit length prediction with RAPHAEL (114). Finally, the dataset to test the detection of repeat proteins is taken from the same paper, i.e., 105 solenoid and 247 non-solenoid proteins with different topologies and no detectable sequence similarity.

3. MATERIALS & METHODS

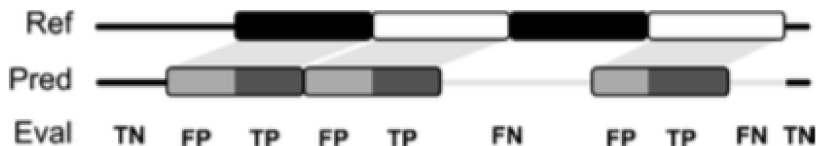


Figure 3.2: Evaluation of repeat unit predictions. The evaluation (Eval) of a prediction (Pred) against a manually curated reference (Ref) is shown, with repeat units as rounded rectangles. The reference and predicted units are paired for maximal overlap. Residues are then categorized as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) depending on repeat unit overlap. (176)

Performance evaluation For the unit-centric evaluation, a new strategy was implemented to take into consideration both the unit phase (position shift) and size. Predicted units are paired against the reference before defining the confusion matrix. Only one predicted unit is matched for each reference unit. When multiple predicted units overlap a single reference unit, the predicted unit with maximum overlap is selected. Figure 3.2 shows how a prediction is evaluated and how true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are calculated. For classification, given a solenoid subclass, TP is the number of proteins with correct assignment, FN are proteins assigned the wrong class and FP are class assignments to wrong targets. TN is always zero since the test set contains only classified proteins. For all evaluations, the measures recall (or sensitivity, equation 3.1), precision (equation 3.2) and accuracy (equation 3.3) are used. ReUPred is compared to the TAPO (116) and ConSole methods (115). TAPO predictions have been generated from the web server (default parameters) considering only the first solution. ConSole predictions were generated locally by the stand-alone software (default parameters). The RAPHAEL period is provided in the RepeatsDB entry metadata. For all evaluations, ReUPred has been benchmarked after removing from SRUL units coming from the test protein or structurally similar units. The comparison with TAPO and ConSole was performed on a set of proteins for which all methods predict at least one unit, i.e., 89 out of 108 proteins. Results of the performance evaluation are presented in section 4.2.1.

$$Sn = \frac{TP}{TP + FN} \quad (3.1)$$

3.1 Prediction of tandem repeat units in structures

$$Pr = \frac{TP}{TP + FP} \quad (3.2)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.3)$$

3.1.2 RepeatsDB-lite

RepeatsDB-lite is a web server designed for the prediction, visualization and analysis of repeated regions in protein structures. It is based on an improved ReUPred algorithm (176) using several checks to minimize errors in the unit detection step and speeding up the calculation. A refactoring of the TR unit library allows it to cover all RepeatsDB classes. Its ability of predicting unit position is evaluated against all manually curated RepeatsDB entries (173). A comparison with existing methods is provided for a limited set of solenoid examples for which predictions are available (176).

Alignment type	Class	Min coverage	Max RMSD	Min TM-Score
Repeat unit library	III-Elongated	0.8	2.3	0.35
	IV-Toroid	0.6	2.8	0.4
	V-Beads on a string	0.6	3.5	0.4
Intra-protein	III-Elongated	-	2	0.25
	IV-Toroid	-	4	0.28
	V-Beads on a string	-	4	0.28

Table 3.4: Validation rules for RepeatsDB-lite predictions. The criteria used to include structural alignments in RepeatsDB-lite predictions are shown per repeat class. Coverage is the fraction of residues covered compared to the reference structure. RMSD is the root mean square deviation. The TM-score method is used to calculate the RMSD and TM-score values.

RepeatsDB-lite algorithm RepeatsDB-lite is the evolution of the ReUPred method. As in ReUPred, the inputs are a target structure and the TR unit library, which represents the conformational space and diversity of bona fide repeat units. The algorithm exploits the library by aligning it against the target structure, using the same divide and conquer approach described in section 3.1.1. Once the best unit is identified by structural similarity with the library (called Master unit), the unit is fixed and the algorithm forks (divides), propagating the search of the Master unit at the N- and C-termini. Only alignments satisfying the similarity criteria described in Table 3.4 are considered valid. Coverage, RMSD and TM-score thresholds were calculated from a similarity network analysis and guarantee separation between subclasses. The units

3. MATERIALS & METHODS

predicted in the target structure are then collected and evaluated together (conquer). In this phase, fragments included in the region but deviating from the classical unit structure are annotated as insertions. At the end, if the region has fewer than three units, the next potential unit from the library is used as Master and the entire iterative part is repeated from the beginning for up to four increasingly relaxed iterations. Compared to ReUPred, the new algorithm discards structural alignments where unit boundaries break (fall inside) secondary structure elements and is able to detect multiple regions of different repeat classes inside the same chain. For α -solenoids and β -hairpins it also provides a finer classification that describes the unit conformational type or fold, often corresponding to the protein family. The average execution time for a single chain is some minutes but varies depending on the class of the master unit.

Repeat unit library The RepeatsDB-lite TR unit library represents all known repeat conformations, including elongated, closed and beads-on-a-string repeats, for a total of 20 subclasses. To increase predictor speed the unit library has been structured hierarchically in three layers. The search for the Master unit starts from the reduced library and then propagates to other layers considering only related units, i.e. belonging to the same cluster as the previous layer. The bottom layer of the unit library is built considering all units of manually curated RepeatsDB entries. A strong reduction is performed by excluding those with insertions (3,401 units) and with missing (non-crystallized) residues (530 units). Units diverging from the subclass average length and redundant units at 70% sequence identity, calculated with CD-HIT (180), are also discarded. At the end, the bottom layer counts a total of 2591 TR units. The other two layers are generated by reducing the structural similarity. Units are clustered at 0.5 TM-score and 80% overlap (coverage) in the middle layer (1160 units) and at 0.3 TM-score and 80% overlap for the SRUL core (top layer, 536 units).

Analysis of the prediction The RepeatsDB-lite software includes additional modules to analyze TR unit predictions. The first is a multiple structure alignment of the units calculated with Mustang (181) useful to highlight overall unit conformation, insertions, diverging units and prediction errors. Another output is a matrix representing the structural similarity between unit pairs. It is calculated by performing an

3.1 Prediction of tandem repeat units in structures

all-against-all pairwise structure alignment with TM-align (179). The units in the matrix are reported from N- to C-terminus and cells are colored based on the observed sequence similarity calculated upon structural alignment and normalized by the length of the shortest unit. From the matrix it is possible to identify patterns of similarity useful to trace the evolutionary history of duplication events.

Benchmarking unit prediction In order to characterize TR proteins it is necessary to identify the position of the repetitive structural elements (units). Assuming TR units are structurally similar inside the same protein, the problem would be reduced to the identification of the unit phase and length. Since in reality units are not homogeneous but often include insertions and structural variation, the evaluation has to be performed unit by unit. We considered manually curated RepeatsDB (version 2017.10.25) entries as source of real unit annotation for benchmarking. TAPO and Console (115, 116) were also compared on the solenoids class. The dataset is the same used in the ReUPred paper (same proteins, described in section 3.1.1) with updated unit annotation according to RepeatsDB version 2017.10.25. Unit prediction performance is measured adopting a strategy similar to the ReUPred paper (176). To obtain a fairer evaluation and assess the effect of incomplete data, RepeatsDB-lite was also benchmarked removing units in the library with over 40%, 60% or 80% sequence identity with dataset proteins. Each reference unit is paired with the predicted unit with maximum symmetric coverage (if any). True positives (TP) are matching residues, false positives (FP) are all predicted unit residues outside reference units, false negatives (FN) are reference unit residues not overlapping with any matching unit and true negatives (TN) are all residues correctly predicted as not repeated. Insertion residues in the reference are masked, i.e. not considered for the calculation of the confusion matrix. Predicted insertions are considered negative predictions and overwrite overlapping unit predictions. When the predictor does not identify any unit or the returned file is empty it is evaluated as a fully negative prediction. When the reference protein contains multiple TR regions the entire sequence is split along the middle point between regions. This is necessary to distribute negative residues equally between regions and to perform region and class based statistics accurately. Sensitivity (equation 3.1), specificity (equation 3.4), precision (equation 3.2), balanced accuracy (equation 3.5) and F-measure (equation 3.6)

3. MATERIALS & METHODS

are calculated. Results of the comparison with other methods are presented in section 4.2.3.

$$Sp = \frac{TN}{TN + FP} \quad (3.4)$$

$$Acc = \frac{Sn + Sp}{2} \quad (3.5)$$

$$F = \frac{2 * Pr * Sn}{Pr + Sn} \quad (3.6)$$

3.2 Prediction of changes in protein solubility

Taking advantage of the knowledge derived from the study of protein folding, in particular in non-globular proteins, we developed a tool for the prediction of protein solubility, SODA (178). SODA results are described in section 4.5.1; the algorithm predicts the changes introduced by a mutation in the "solubility profile" of a protein, based on its aggregation, disorder and secondary structure propensities. The following section describes the algorithm.

3.2.1 SODA

SODA predicts solubility changes introduced by a mutation by comparing the profiles of the wild type (WT) and mutated sequences. SODA is able to evaluate difficult types of variation including point mutations, deletions and insertions. The predictor is entirely based on sequence features. The PASTA (182) aggregation propensity and ESpritz (183) intrinsic disorder scores are combined with a Kyte-Doolittle hydrophobicity profile (184) and secondary structure propensities for α -helix and β -strand estimated with FESS (185). When available, a protein structure can be used to improve the prediction by masking buried residues from the solubility prediction.

Algorithm SODA prediction is based on five individual component scores (calculated with default parameters): PASTA aggregation energy with 90% cut-off specificity (182), ESpritz disorder propensity in X-ray prediction mode (183), the negative Kyte-Doolittle hydrophobicity profile (184) and the two secondary structure propensities for α -helix

3.2 Prediction of changes in protein solubility

and β -strand calculated with FESS (185). Each score difference ΔS is summed and normalized for the full sequence using the following formula:

$$\Delta S = \frac{\sum_{j=1}^n s_j^{mut}}{n} - \frac{\sum_{j=1}^m s_j^{wt}}{m} \quad (3.7)$$

where s_j^{mut} and s_j^{wt} are the scores of the mutated and wild-type residue j in the sequences and n and m are the respective sequence lengths. Note that the two sequences may be of different length as SODA also supports insertions and deletions. When a structure is available, the ΔS value for residues with less than 20% solvent accessible sidechain area (calculated with DSSP) are set to 0. The final SODA score, $\Delta S_{Solubility}$, is the weighted sum of the partial scores:

$$\begin{aligned} \Delta S_{Solubility} = & \Delta S_{Aggregation} + w_1 * \Delta S_{Disorder} \\ & + w_2 * \Delta S_{Hydrophobicity} + w_3 * \Delta S_{Helix} + w_4 * \Delta S_{Strand} \end{aligned} \quad (3.8)$$

where w_1, w_4 are weighting parameters set to optimize the SODA score on the PON-Sol dataset. Their optimized values are 2, -50, 2 and 2, respectively. When the difference ($\Delta S_{Aggregation}$) is positive, the mutated protein is more soluble (lower aggregation energy) than the WT. Similarly when $\Delta S_{Disorder}$ is positive, the mutated protein gains solubility because it is more disordered. Likewise, hydrophilic (charged/polar) residue content increases solubility.

Training and evaluation SODA is trained using 5-fold cross-validation on a filtered version of the PON-Sol dataset (186). Weights for the parameters are chosen from a grid search on the interval [-100,...,+100], selecting the first weight optimizing the PON-Sol prediction for each term. All variants without any solubility effect as well as ambiguous examples from the original dataset were discarded. These are cases where it is not possible to obtain the original sequence or containing a mismatch between mutation and original sequence. Moreover, in order to make the benchmarking fair, a maximum pairwise sequence identity of <30% was imposed against the CamSol dataset (see below). A total of 142 variants classified as increasing (positive values) or decreasing (negative values) solubility from 49 proteins were used for training. Table 3.5 shows the performance of SODA and its components on the PON-Sol training set. Among the single component scores, PASTA and hydrophobicity stand out for opposite reasons,

3. MATERIALS & METHODS

with good performance for positive and negative cases respectively. SODA reaches an accuracy of 59% overall (84 correct predictions). On the restricted dataset, including only mutations classified in PON-Sol dataset as having stronger effect on solubility, the accuracy is 67% (35 / 52 correct predictions, data not shown). Mutations in the PON-Sol dataset are manually classified based on experimental evidence from the literature. Notably, SODA is very good at predicting solubility decrease. The specificity, i.e. fraction of true positives over all positive predictions, is 72% and 100% in the full (Table 3.5) and restricted (not shown) training sets respectively. This is somewhat expected, as SODA uses the PASTA energy, which is known to be highly specific for aggregation prediction. SODA is compared to other solubility predictors in section 4.5.1. The dataset is the same used in the recent CamSol paper (187) and includes 19 proteins and 56 variants from four publications: Trevino (188), Miklos (189), Tan (190) and Dudgeon (191). All proteins have less than 30% pairwise sequence identity to the training set and represent a real blind test.

	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Strand	21	<u>45</u>	40	<u>36</u>	36.8	52.9	46.5
Helix	35	35	26	46	43.2	57.4	49.3
Hydrophobicity	35	46	26	35	<u>50.0</u>	63.9	<u>57.0</u>
ESpritz	39	41	22	40	49.4	65.1	56.3
PASTA	47	31	14	50	48.5	<u>68.9</u>	54.9
SODA	<u>46</u>	38	<u>15</u>	43	51.7	71.7	59.2

Table 3.5: Evaluation on the PON-Sol training set. True positives (TP), true negatives (TN), false positives (FP), false negatives (FN), and sensitivity ($TP/(TP+TN)$), specificity ($TN/(TN+FP)$) and accuracy ($(TP+TN)/(TP+TN+FP+FN)$) values are reported as percentages. The best value is in bold and the second best underlined.

3.3 Web resources implementation protocols

The present thesis describes two databases and two web servers. The former are RepeatsDB, database of TR structures described in section 4.2.2, and MobiDB, a comprehensive resource for the annotation of protein disorder described in section 4.3.1. The web server were designed for TR prediction (RepeatsDB-lite, in section 4.2.3) and solubility prediction (SODA, in section 4.5.1). The following section goes through the detail of their implementation with a multi-tier architecture, using separate modules

for data management, data processing and presentation functions. Figure 3.3 presents a schema of the technologies used in websites implementation.

3.3.1 Databases

Data are stored through MongoDB, a NoSQL document oriented database management system. "NoSQL" means non-relational database, i.e. a database with non-fixed structure with a key-value storage type, document storage type and so on. MongoDB provides a schema-free, document oriented framework where documents are stored in collections and collections in databases. Collections can contain records with different schema documents, i.e. different number of attributes. For this reason it fits our type of data, since we have different annotation available for different entries. In addition, MongoDB provides a straightforward framework for data extraction and re-modelling, that is the Aggregation Pipeline. Documents enter a multi-stage pipeline that transforms the documents into aggregated results, allowing fast data processing. The Aggregation Pipeline was extensively used both for the calculation of statistics in the characterization of NGPs and for the visualization of aggregated data provided by the websites on the fly.

3.3.2 Interfaces

To simplify website development and maintenance, all tiers handle the JSON (JavaScript Object Notation) format, which is the format of documents in MongoDB, thereby eliminating the need for data conversion.

Back-end The "Back-end", or server-side, refer to the data access layer that connects the data (stored, in our case, in MongoDB databases) to the presentation layer (the web interface that the user navigates). In addition, this is the layer that manages user sessions and data processing. In the case of a web-server such as RepeatsDB-lite, the "Back-end" is the architecture that allows the interpretation of the user request, that submits it to the predictor software and that check and finally processes the predictor response in order to make it available to the "Front-end". All our web servers exploit the Node.js functionality, which supports the implementation of REST (Representational State Transfer) architectures. The REST architecture allows access from a web-based

3. MATERIALS & METHODS

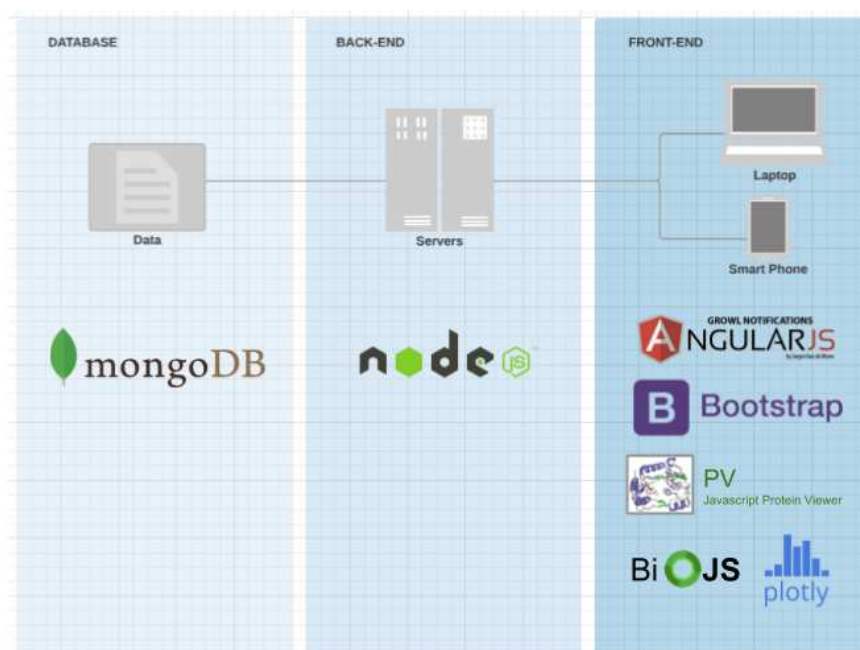


Figure 3.3: Schematic description of the technologies used in the implementation of database, front-end and back-end. Licence: Creative Commons Attribution 3.0. Author: Lisanna Paladin.

user interface as well as programmatically through external APIs.

Front-end The "Front-end" represents the web interface that the user is presented to when navigating a website. Our web interfaces were developed using the AngularJS or Angular2 (in MongoDB 3.0) framework and Bootstrap CSS style sheets. AngularJS to Bootstrap integration is available through the angular-ui project. Angular meets huge data requirements and allows the high speed rendering of web pages across different platforms. Bootstrap is an open source toolkit for developing with HTML, CSS and JS and is used to build responsive and mobile-friendly projects. These technologies were selected to provide the overall uniform look-and-feel. The dynamic and interactive views of biological data, i.e. the protein sequence, structure and features viewers, are developed using state-of-the-art technologies. PV (<https://biasmv.github.io/pv/>) and WebGL (<http://nglviewer.org/ngl/>) are used for structure visualization and the specialized libraries from BioJS (<https://biojs.net/>) for sequence and sequence alignment viewer. Graphs are designed using the Plotly.js library (<https://plot.ly/javascript/>). Finally, protein sequence features are visualized using a D3 library built in-house on the

3.3 Web resources implementation protocols

model of the neXtProt project of SIB CALIPHO group (<https://github.com/caliphosib/feature-viewer>). Major innovations were introduced to the Feature Viewer (FV) project to meet the requirements of our data visualisations: (i) the FV project was entirely converted from Javascript to Typescript; (ii) the new FV allows subfeatures visualization, i.e. selected features are clickable to show details as new tracks (placed under the "parent" feature); (iii) on the right side of each feature, developers can now customize specific buttons, a default button is provided to show a percentage (e.g. disorder coverage in MobiDB 3.0); (iv) the FV features customizable tooltips and styles (e.g. stroke and opacity of the boxes). The new FV was specifically designed for MobiDB multi-layered data (described in section 4.3.1), allowing users to visualize entry annotation at the desired level of detail.

3. MATERIALS & METHODS

4.1 Protein tandem repeats characterization

Tandem repeat (TR) regions in proteins are characterized by a repeated sequence which codes for a modular architecture, where structural modules are called units (Figure 1.6). TRPs structures are non-globular in the sense that, to stabilize the folded state, instead of relying on a hydrophobic core of amino acid chains buried from water they rather show an hydrophobic axis of stacking interactions between each unit and the flanking ones. As a consequence, they show elongated shapes and allow for a higher flexibility. Exploiting these properties, they carry out fundamental functions in all kinds of organisms. The following chapter includes three sections. The first, section 4.1.1, presents the classification of TR included in RepeatsDB. Basing on the original classification proposed in (79), RepeatsDB defines five main classes, mainly based on repeat unit length, with subclasses representing specific structural arrangements. We compared this data to the one in Pfam database of protein families, which provides and alternative classification based on evolutionary conservation of protein sequences. Most instances are found to map one-to-one between structure- and sequence-based schema. Some notable exceptions are discussed. The following, section 4.1.2, describes the large-scale characterization of a dataset of human TRPs in terms of function, role in protein-protein interaction (PPI) networks and association to diseases. The compu-

4. RESULTS & DISCUSSION

tational analysis highlights a significant association between TRPs and human diseases explained by their role as hubs in PPI networks. Indeed, TRPs extended and modular structure makes them perfect candidates to serve as PPI platforms. As a case-study of this framework, section 4.1.3 presents the interactome of Collagen V (a repeat protein associated to Ehlers-Danlos syndrome, EDS) in order to identify genotype-phenotype correlations in relation to the interaction network model. It shows that heterogeneous classical EDS manifestations may be explained by the involvement in different extra-cellular matrix pathways.

4.1.1 Comparison of protein repeat classifications based on structure and sequence families

RepeatsDB 1.0 (80) contained a set of detailed proteins, manually annotated with the exact location of structural units and regions, allowing a comparison between structural conformation and repetitive Pfam sequence families. This comparison may be useful to explore the hypothesis of repeat evolution through sequence duplication as it assesses the conservation of Pfam families inside each repeat subclass. Figure 4.1 summarizes the most frequent domains/clans detectable inside the subclasses. In the following section, we briefly review the results by RepeatsDB class.

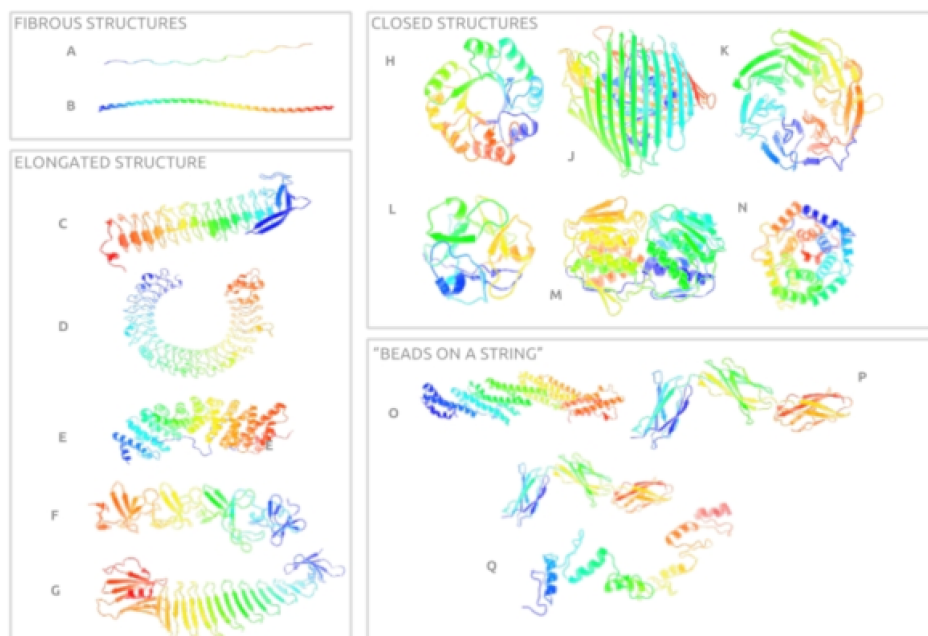
4.1.1.1 Class I: Crystalline aggregates of unlimited size

Class I includes proteins with 1 or 2 residue-long repeats. These sequences arrange in crystallites which are usually transported outside the organism. The structures are characterized by unlimited size and high stability but do not correspond to biological molecules since their function is not required inside the organisms, hence no PDB structure is associated to this class (79).

4.1.1.2 Class II: Fibrous structures

The repeat proteins belonging to class II are fibrous structures characterized by very short repeat length: collagens and α -helical coiled coils. The sequence repeat in these cases corresponds to one turn of the fibrous structures. Two subclasses can be distinguished. Collagens (subclass II.1) are chains that assemble into super molecular triple helices. A limited number of collagen structures is available in the PDB due to

4.1 Protein tandem repeats characterization



Class	Subclass	Letter	Entries	Pfam Clan
I	I.1: Poly-alanine β structures	-	0	
II	II.1: collagen triple-helix	A	5	
II	II.2: α helical coiled coil	B	128	
III	III.1: β -solenoid	C	170	hexapeptide repeats (CL0536) Hemolysin (PF00353) Pectate lyases (CL0268) LRRs (CL0022)
III	III.2: α/β solenoid	D	89	LRRs (CL0022)
III	III.3: α -solenoid	E	923	ANKs (CL0465) Tetratricopeptide repeats (CL0020) Cell wall binding domain (PF01473)
III	III.4: trimer of β spirals	F	20	
III	III.5: single layer anti-parallel β	G	7	
IV	IV.1: TIM barrel	H	827	Amidohydrolases (CL0034) Tim barrels (CL0036)
IV	IV.2: β -barrel	J	17	
IV	IV.3: β -trefoil	K	44	Beta trefoils (CL0066)
IV	IV.4: β -propeller	L	433	Beta propellers (CL0186) Hemopexins (PF00045)
IV	IV.5: α/β prism	M	14	
IV	IV.6: α -barrel	N	5	Six hairpins (CL0059)
V	V.1: α -beads	O	3	
V	V.2: β -beads	P	112	Ig-like fold (CL0159) Immunoglobulins (CL0011)
V	V.3: α/β -beads	Q	7	
V	V.Other	R	7	

Figure 4.1: (TOP) The typical structures of the repeat subclasses are shown. The reference letters are reported in the table below. (BOTTOM) The first two columns report the RepeatsDB 1.0 structural classification of repeat proteins, followed by the reference letter for figure and the number of entries. The last column shows the Pfam clans associated to the structural subclasses. (87)

4. RESULTS & DISCUSSION

limitations in crystallization. The collagen repeat is the tripeptide Gly-X-Y (192), with frequent prolines and hydroxyprolines in the X and Y positions. As their sequence is usually highly degenerated and a limited number of structures is available, it is difficult to assess the quality of the collagen Pfam domain (PF01391). The α -helical coiled coil (subclass II.2) is characterized by a (abcdefg)_n repeat, with the a and d positions occupied by hydrophobic and the remaining by polar residues (193). They fold into alpha-helices winding around the axis of the coiled coil structure. A limited number of Pfam domains is associated to this subclass. The association is usually case-specific, with each Pfam domain corresponding to one entry. Some examples are the Rabaptin family (PF03528) and the ATP synthase (E/31 kDa) subunit family (PF01991). These domains are usually defined by the function of the whole protein containing the repeat region. Hence, domains retrieved from the analysis of this family are too specific to group them into a coiled coils family.

4.1.1.3 Class III: Elongated structures

Class III is characterized by repeats forming elongated structures that can vary in length from 5 to ca. 45 amino acids. The typical feature of this class is that the repetitive structural units require one another to maintain structure. They split into two categories: solenoid or non-solenoid structures. Solenoid repeats are composed by solenoid windings of the polypeptide chain that can be made exclusively by α -helices, only β -sheets or a mixture of the two secondary structure elements (83). Non-solenoid structures of class III have been described more recently. They comprise the trimer of β spirals, the single layer anti-parallel β structure and some others, including the antiparallel β structure folded along the longest axis and the spiral β -hairpin staircase fold. Beta-solenoids (subclass III.1) include four different manually annotated clans. The most represented are CL0268 (the pectate lyase clan) and the Leucine Rich clan. The hexapeptide repeat and hemolysin clans are also associated to this subclass. Some other frequent domains are not grouped in a clan and do not show a homogeneity of functions. Alpha-beta solenoids (subclass III.2) mainly contain LRR domains, while alpha solenoids (subclass III.3) show an extreme sequence divergence. The Pfam domains associated to this subclass are the most common types of sequence repeats. A recent review explored alpha-solenoid diversity, identifying solenoid sequences containing HEAT, Armadillo, Pumilio, Ankyrin (ANK), LRR and tetratricopeptide (TPR)

4.1 Protein tandem repeats characterization

repeats (194). They analyzed the phylogenetic distribution of these repeat families and concluded that they probably emerged independently several times during evolution. This hypothesis is supported by the widespread distribution of protein functions across the subclass. Pfam mainly collects the different repeat types in two clans: the ANKs (clan CL0465), and the tetratricopeptide repeat clan that groups together the other repeat types (Armadillo, Pumilio, TPRs, HEAT, Sel1, LRRs and others; CL0020). This division suggests that ANK-containing proteins are a functionally distinct entity inside alpha-solenoids, and indeed they appear to have a more complex structure. A recent study (194) assessed the distribution across organisms of three of these most common families: the armadillo (ARM), tetratricopeptide (TPR) and ankyrin (ANK) domains. They analyzed the typical structural arrangement of these domains, and showed that the ANK repeat folds in a unit composed by a β -turn, two antiparallel α -helices and a loop while the TPR repeat is composed of an α -helix-turn- α -helix motif. The ARM superfamily folds into a module with three α -helices, with the exception of the HEAT family, composed by two α -helices. Hence, in α -solenoids the existence of different structural arrangements is becoming evident, corresponding to distinct functional and evolutionarily related families. Regarding the non-solenoid class III structures, only subclass III.4 shows a clear association with the putative cell wall binding repeat Pfam domain (PF01473) folding into a trimer of β spirals. An interesting distribution of Pfam domains is observed within elongated structures, considering how Leucine-rich repeat (LRR) domains are present in all three solenoid subclasses. The structure of LRRs has been early defined as an arc or horseshoe shape, with a concave face consisting of parallel beta-strands and a convex face composed by variable secondary structure elements (83, 195). Later, this structure was identified as an alpha/beta solenoid by Kajava (120). LRRs are involved in receptor-based recognition, but beyond innate immunity are associated to a widespread range of functions and functional sequence motifs (195). Their sequences have been classified into seven classes and the corresponding structures analyzed (83). LRR structural conformations cover a wide range of the variability of subclasses III.1 and III.2 (beta- and alpha/beta-solenoids). Their association with alpha-solenoids still needs to be investigated further.

4. RESULTS & DISCUSSION

4.1.1.4 Class IV: Closed structures

In the repetitive "closed" structures, the last unit interacts with the first one and all are flanked one to the other building up a torus with a fixed number of repeats. This is the most represented class in RepeatsDB 1.0, with a total of 3,828 structures. The repeat length overlaps with both classes III and V, ranging from ca. 30 to 60 amino acids and the subclasses are the TIM-barrels, β -barrels, β -trefoils, β -propellers, α/β prism and α -barrels. TIM-barrels (subclass IV.1) are associated to the Pfam Glyco_hydro_tim clan (CL0036), which groups a series of TIM-barrel families defined by their enzymatic function (aldolases, isomerases, DNAses and others). A distinct clan associated to subclass IV.1 are the amidohydrolases (CL0034). β -barrel structures (subclass IV.2) are far less numerous, with only 17 examples in RepeatsDB 1.0. Domains in this subclass associated to single entries, mostly porines, ubiquitin or, in some cases, the Pfam MBB clan (CL0193). The latter groups together a set of β -barrels with significant sequence similarity and different numbers of β -strands. Subclass IV.3 is mainly associated to β -trefoil clan (CL0066). Somewhat surprisingly, the PF00331 family associated to this subclass is not part of the β -trefoil clan, but belongs to the TIM-barrel one instead. Almost all examples of β -propellers (subclass IV.4) are associated to at least one domain in the Pfam Beta_propeller clan (CL0186) with the sole exception of the hemopexin domain (PF00045). The WD40 domain (PF00400) is the most important result, since it is recognized in correspondence with about 60% of the subclass IV.4 units while exhibiting functional diversity. The functional variability and evolutionary origin of WD repeats was explored proposed to functionally cluster by surface similarity with the purpose to identify common interaction partners (196). Indeed, this β -propeller domain seems to be one of the cell's most pervasive scaffolding and interaction domains (197). The last clan associated to elongated repeats, the α -barrel subclass, is the 6_Hairpin clan (CL0059) where the six helical hairpins characterizing these domains correspond to the six alpha barrel units.

4.1.1.5 Class V: Beads on a string

The class V structures have repeat units of more than 50 residues forming globular domains connected as beads on a string, through either flexible or rigid linkers. The classification comprises α -beads, β -beads, α/β -beads and other as a catch-all for the

4.1 Protein tandem repeats characterization

remaining types of beads on a string. The low number of structure examples in the database limits the analysis of the associated domains and suggests that further improvements to the RepeatsDB classification procedure are warranted. Subclass V.2 (β -beads) shows a clear association to two main Pfam clans: the Immunoglobulin superfamily (CL0011) and the Ig-like fold superfamily (CL0159), which includes fibronectins, filamins, integrins and other types of Ig-like binding domains. In addition, the Sushi domain (PF00084, no clan) seems to be associated to β -bead structures.

4.1.2 Tandem Repeat proteins at a glance: functions, diseases and role in protein-protein interaction network

We describe here the role TPRs in the human organism, by analysing their function, localization, position in protein-protein interaction network, highlighting their ubiquity in tissues, subcellular districts and functional pathways. TPRs are exploited as hubs of protein-protein interactions, being critical players in the cell and probable disease targets. The present analysis is based on UniProt annotation and derived from the current versions of UniProt and RepeatsDB, 2018_01 and 2017.10.25 respectively. The subset of UniProt entries retrieved in RepeatsDB is collected in the "RepeatsDB" dataset. An additional and more extended dataset ("Repeats" dataset) of repeat proteins is collected using the "repeat" tag in UniProt, assigned through sequence-based methods. Other datasets collected to test different hypothesis are the "Hubs" and the "Disease", i.e. proteins with more than 50 interactors and proteins involved in at least one disease, respectively. Dataset sizes are reported in Table 4.1.

	Repeats	Disease	Hubs	Total
RepeatsDB	164	88	4	273
Repeats		471	17	3206
Disease			40	4070
Hubs				135
All				73112

Table 4.1: Dataset sizes in human proteome analysis and intersections.

4. RESULTS & DISCUSSION

4.1.2.1 A structure of success

The key of TR evolutionary success may be accounted for the peculiar properties of a modular structure (198). The assembly of several similar building blocks is the result of a number of repetitive short-range interactions established between the contacting unit interfaces, building an elongated central hydrophobic protein axis (79). Escaping the globular protein paradigm of folding as a well-defined route to the native 3D structure, the typical TR proteins assembly is rather a cluster of different parallel folding pathways (199), with each unit contributing to the folding and stability of the flanking ones. This is especially true for solenoids (81), but also for other types of TR proteins: e.g. fibrous structures are stabilized by short-range and regularized inter-chain interactions (200), and toroids are formed by the cooperative stabilization of several simple super-secondary modules that assemble in a closed structure, with the first unit contacting the last (201). The main player in TR regions stability is therefore the central hydrophobic axis, rather than core. Once it is conserved, the other residues are not only able, but prone to diverge, as demonstrated by (120) where the perfection of repeats was correlated with the tendency to be unstructured. The extended and quickly evolving surface of TR regions have been exploited for the specialization in protein binding (81, 202). Further tunable factors that confer specificity to the binding are the overall shape of the molecule (twist and curvature) as well as its flexibility. Relatively little rearrangements at the level of the module sequence, such as deletion, insertions or substitutions, lead to the building of units that show different angle of inclination with respect to the protein axis. Insertions and combinations of these junction-modules account for the global shape of the scaffold, and have been exploited for the building of unique curvatures in repeat protein design (203). All these properties are related to the mechanism of their evolution. Repeated stretches at the DNA level are prone to self-expansion via tandem duplication, and the peculiar elongated arrangement of the protein product can easily tolerate the insertion of a new structural unit (204). As a consequence, the evolution of TR proteins usually includes a complex pattern of insertion, deletion and rearrangement of units (205). All the described features dictate the perfect recipe to build a binder, giving a possible explanation evolutionary advantage of repeat structures.

4.1 Protein tandem repeats characterization

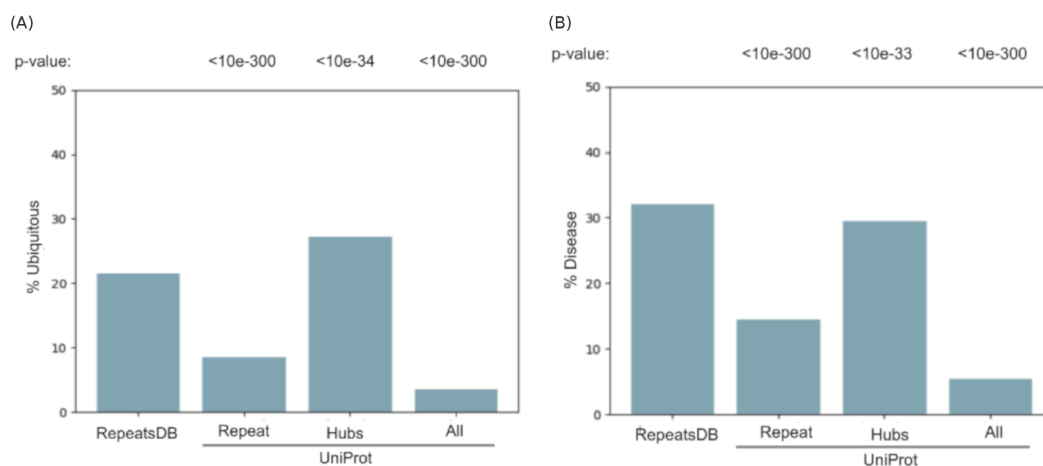


Figure 4.2: Percentage association to diseases and ubiquity. The association to diseases and ubiquity ("ubiquitous" tag in UniProt "Tissue specificity" text) in tissues is tested in comparison to the "RepeatsDB" dataset in "Repeat", "Hubs" and "All" through a Chi-Square Test. License: Attribution-NonCommercial-NoDerivatives 4.0 International. Author: Lisanna Paladin.

4.1.2.2 The perfect binder

TRs are characterized by conserved hydrophobic axis and hypervariable positions usually in contact with the ligand (206), evolved to be specific for that interaction. They show variable specialized shape, structural arrangement, number of units. It is no coincidence that TR proteins are largely involved in the immune system and pathogen recognition in vertebrates (207), functions that require fast evolution but also high specialization of the interaction surfaces. Thanks to the tendency to duplicate and tolerate new unit insertions, TRs evolved specificity for more than one partner at the same time, bringing them together into a functional complex (208). This tendency is confirmed by the average of interactors number in the dataset of human TR protein collected from RepeatsDB, higher than the UniProt background (Figure 4.5A). By further investigating the features of this census of TR proteins, we evidenced other traits of this phenomenon.

4.1.2.3 A protein population that colonized every organism district

At the level of organism tissues, TR proteins appear to be ubiquitous. Figure 4.2A depicts the distribution of the "ubiquitous" tag of UniProt database, which character-

4. RESULTS & DISCUSSION

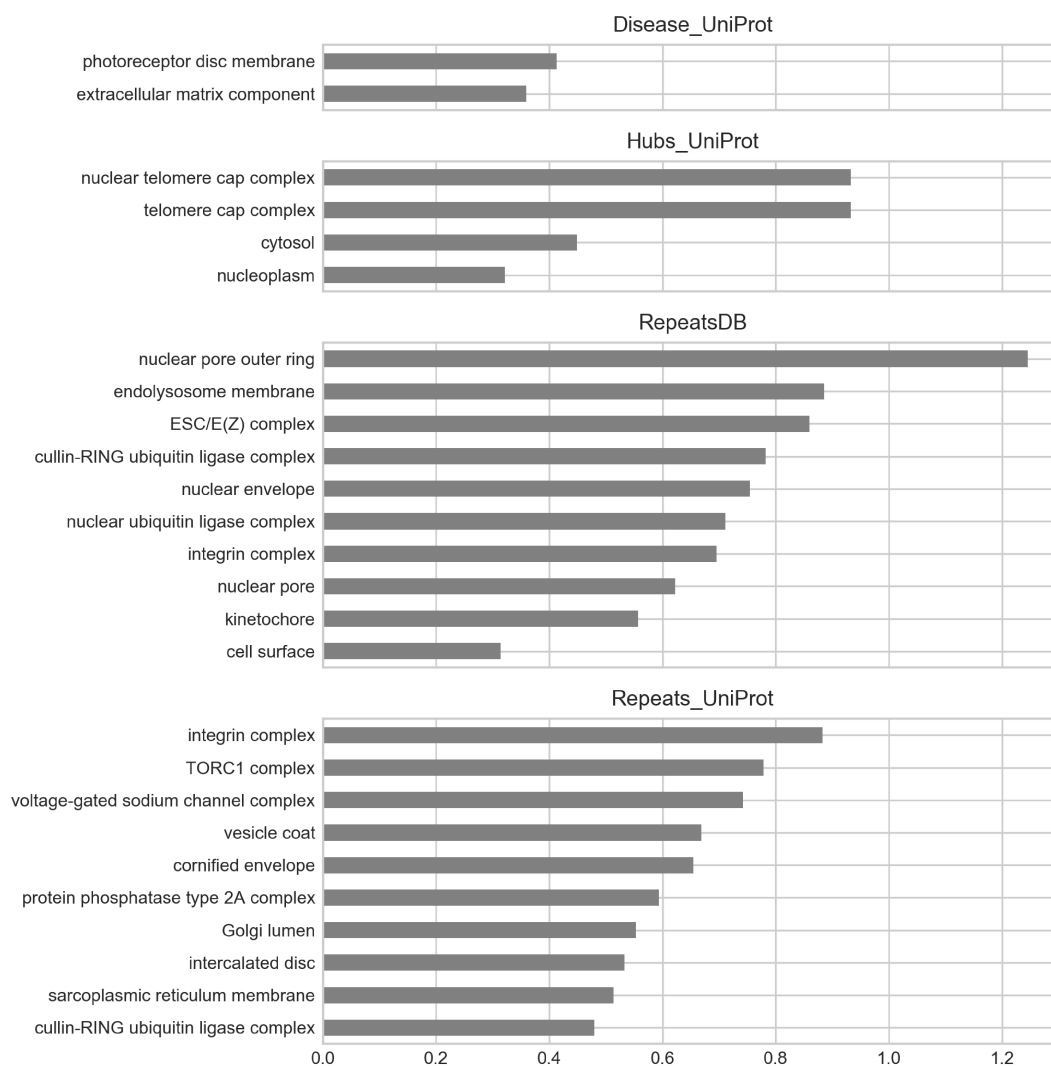


Figure 4.3: The GO terms enrichment was computed for the four datasets "Disease", "Hubs", "RepeatsDB" and "Repeats". The GO terms enrichment is based on the frequency a term is found in the datasets compared to the control group (all proteins in UniProt but not in dataset). The null hypothesis is that the GO term us associated to proteins in the test group by chance. In particular, terms are kept if their P-value is lower than 5×10^{-7} (corrected using Bonferroni) and the enrichment is computed as follows: $E = \log\left(\frac{\#GO_{i \text{ in } P_{\text{Dataset}}}}{\#GO_{i \text{ in } P_{\text{Dataset}}}} - \frac{\#GO_{i \text{ in } P_{\text{UniProt}}}}{\#GO_{i \text{ in } P_{\text{UniProt}}}}\right)$ where # denotes the number of proteins belonging to a given class. Only GO-terms that are one step away from the root are considered, so that more general features can emerge. Plots show the ten (or less, if less terms are significant) most enriched terms for each dataset. Here only the enrichments discussed in the text are reported, i.e. the Cellular Component enrichment. License: Attribution-NonCommercial-NoDerivatives 4.0 International. Author: Lisanna Paladin.

4.1 Protein tandem repeats characterization

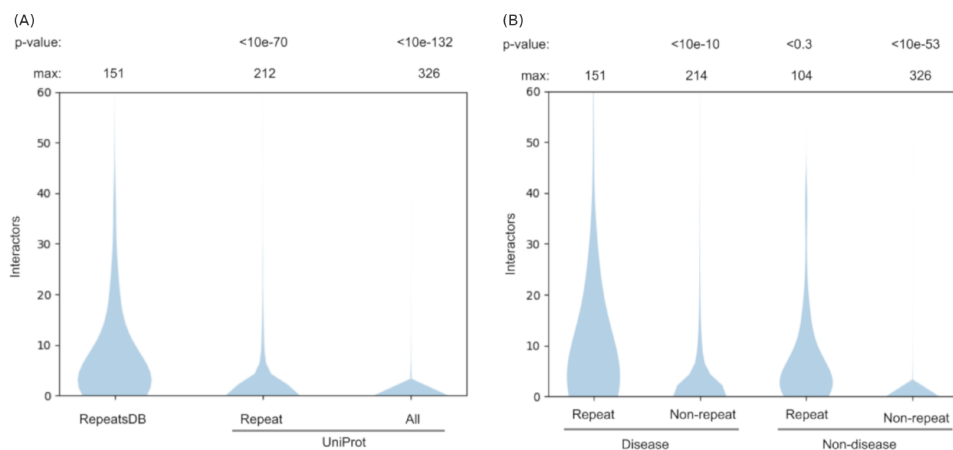


Figure 4.4: (A) Interactor number of UniProt and repeat proteins. Interactors number violin plot of human UniProt proteome and two repeat datasets ("Repeats" from UniProt, sequence-based, and "RepeatsDB" structure-based). Few extreme values (proteins with more than 60 interactors) are cut for a better visualization of the distributions. The significance of number of interactors of "RepeatsDB" versus "Repeats" and "All" is tested through a Chi-Square Test. (B) The four datasets are subset of the human proteome associated or not to diseases and containing or not a repeat (according to RepeatsDB). The significance of number of interactors of "Repeat Disease" (RepeatsDB entries associated to a disease) versus the other datasets is tested through a Chi-Square Test.

izes proteins present in all human tissues. Repeat proteins from RepeatsDB, and hub proteins, are much more diffused than the human proteome background. At the level of the cell, TR proteins are widespread in all subcellular compartments. Figure 4.3 shows the GO Cellular Component enrichment of TRPs and the other datasets. It evidences that TRPs are particularly abundant in pores, envelopes and one the cell surface, so mainly in non-organelles district and organelle "gates". These compartments are large spaces devoted to cell diffusion where a number of essential cell processes occur, from the assembly of transfer RNA, messenger RNA and ribosomes (in the nucleus) to the translocation into the cytoplasm and subsequent translation of them into proteins and their post-translational modification. This kind of district are enriched in protein complexes (209) and consequently in proteins devoted to binding functions (85, 210), a probable reason for the preferential localization of repeat proteins in those.

4.1.2.4 Probable candidates of disease-association

The central role of TR proteins in protein-protein interaction network suggests their importance in cell pathways and therefore their possible association to disease. Ex-

4. RESULTS & DISCUSSION

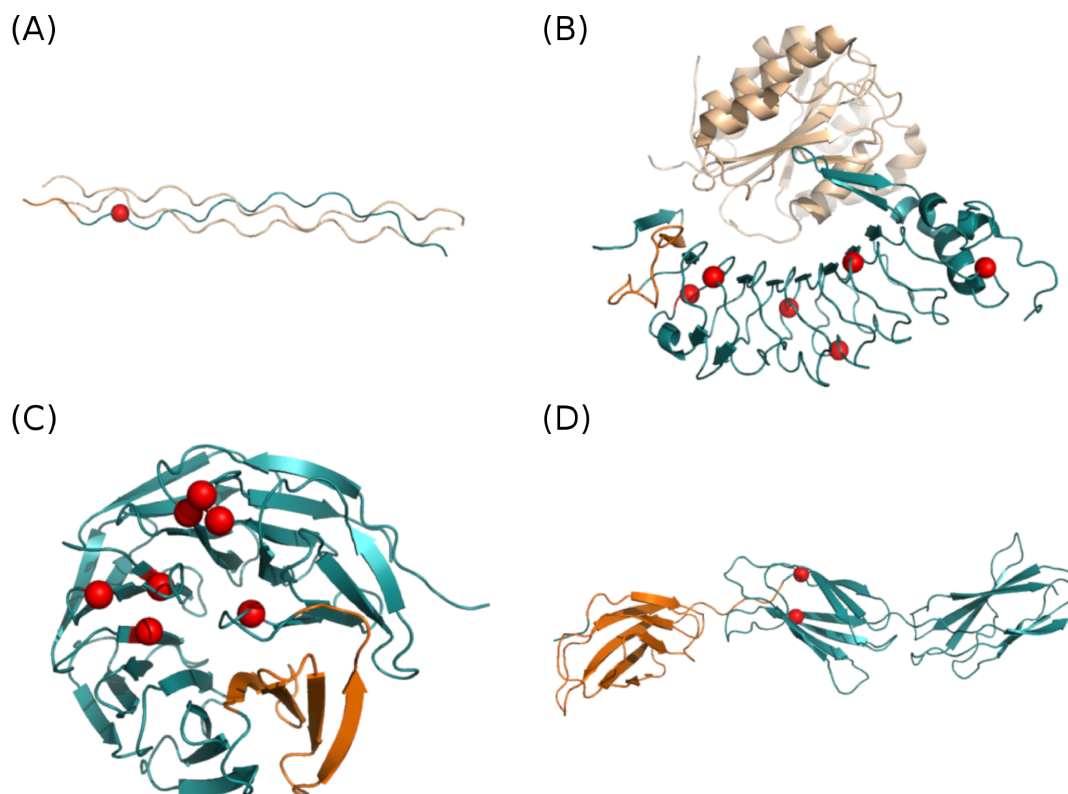


Figure 4.5: Repeat proteins and disease-related mutations. Four examples of repeat protein structures, belonging to different classes, are shown. The repeat region is coloured in blue and inside it a single unit is highlighted in orange. Red spheres indicate the position of pathogenic mutations mapped on the wild type structure. License: Attribution-NonCommercial-NoDerivatives 4.0 International. Author: Lisanna Paladin.

amples can be selected from all repeat classes, namely fibrous, elongated, closed and beads-on-a-string repeats (Figure 4.5D).

- Collagenopathies is a whole group of diseases related to collagen-like structures, where usually the disruptive mutagenesis-site falls into the first residue of the three amino-acid repeat, a Glycine essential to maintain the triple helix structure (211). In figure 4.5A, a collagen structure (UniProt ACC: P02461, PDB ID: 4GYX) is shown and a glycine substitutions is mapped on the structure.
- Glycoprotein Iba is a surface membrane protein of platelets, including a solenoidal region. This protein (UniProt ACC: P07359, Figure 4.5B) contains indeed a

4.1 Protein tandem repeats characterization

Leucine Rich Repeat (LRR) domain, with multiple disease-related missense mutations localized at the interface with its interactor, the plasma glycoprotein von Willebrand factor (the complex is crystallized as PDB code: 1M10).

- Falling into the group of "closed repeats", or toroids, the WD repeat-containing protein 5 (UniProt ACC: P61964, PDB ID: 2CNX, Figure 4.5C) shows a number of mutations linked to a reduced histone H3 binding.
- In Figure 4.5D two pathogenic substitutions are mapped in the repeat domain of Fibronectin (UniProt ACC: P02751, PDB ID: 1FNH), a beads-on-a-string repeat.

The importance of repeat domains contribute in the disease insurgence is confirmed by the prevalence of repeat domains in the top 20 Pfam domains ranked by association to diseases (Table 4.2). This census of repeat proteins highlights their prominent role in disease insurgence, as probable target of disease-related mutations.

ID	Name	#	Type	RepeatsDB	Class	Example	Figure 4.5
PF00069	Protein kinase domain	91	Domain	No			
PF00046	Homeobox domain	74	Domain	No			
PF07714	Protein tyrosine kinase	71	Domain	No			
PF00520	Ion transport protein	55	Family	No			
PF00041	Fibronectin type III domain	49	Domain	Yes	V	1fnhA	D
PF07679	Immunoglobulin I-set domain	46	Domain	Yes	V	2nziA	
PF00400	WD domain, G-beta repeat	45	Repeat	Yes	IV	2cnxA	C
PF00001	7 transmembrane receptor (rhodopsin family)	42	Family	No			
PF01391	Collagen triple helix repeat (20 copies)	40	Repeat	Yes	II	4gyxB	A
PF00169	PH domain	40	Domain	No			
PF12796	Ankyrin repeats (3 copies)	39	Repeat	Yes	III	4rlvA	
PF00089	Trypsin	37	Domain	No			
PF00038	Intermediate filament protein	35	Coiled-coil	No			
PF07645	Calcium-binding EGF domain	35	Domain	Yes	V	2vj3A	
PF00018	SH3 domain	34	Domain	No			
PF00271	Helicase conserved C-terminal domain	33	Family	No			
PF13855	Leucine rich repeat 8	33	Repeat	Yes	III	1m10B	B
PF00168	C2 domain	31	Domain	No			
PF00010	Helix-loop-helix DNA-binding domain	29	Domain	No			
PF00017	SH2 domain	29	Domain	No			

Table 4.2: Top 20 Pfam domains associated to diseases. Association between Pfam domains and diseases is computed by counting the number of proteins in Homo Sapiens harbouring a specific domain and being annotated by UniProt as involved in a disease. The Pfam ID, name are reported, together with the number of proteins mapped to the domain and to a disease (#), the domain type as from Pfam description, the overlap with RepeatsDB repeat regions (RepeatsDB). If the latter is true, the repeat class, an entry example and eventually the reference in Figure 4.5 are reported.

4. RESULTS & DISCUSSION

4.1.2.5 Centrality comes with a price

According to the observations presented here, TR proteins are highly connecting nodes sparsely distributed along the human interactome, supporting their predominant role as interactors, in cell signaling and transporting. The reason for the predominance of interaction role is accounted by their peculiar properties of a modular architecture, which have been exploited to develop binders across a plethora of pathways. This framework has evident consequences as regards TR proteins relationship with organism diseases. As a general rule, the highest is the number of a protein interactors, the most critical are the possible consequence of its disruption for the organism, a phenomenon known as the centrality-lethality rule (212). Since TR proteins show a higher number of interactors than UniProt background, they appear to fit this framework. In addition, we observed that the subset of disease-related TR protein show a degree higher both of TR proteins in general and non TR proteins still associated to diseases (Figure 4.4B), supporting the hypothesized association between TR protein role as binders and their involvement into diseases.

4.1.3 Structural in silico dissection of the collagen V interactome to identify genotype-phenotype correlations in classic Ehlers-Danlos Syndrome (EDS)

As a case-study of TR association with diseases we dissected Collagen V mutations, associated with Ehlers-Danlos syndrome (EDS) (213), a group of heritable collagenopathies. Collagen V structure is not available and the disease-causing mechanism is unclear. To address this issue, we manually curated missense mutations suspected to promote classic type EDS (cEDS) insurgence from the literature and performed a genotype-phenotype correlation study. Further, we generated a homology model of the collagen V triple helix to evaluate the pathogenic effects. The resulting structure was used to map known protein-protein interactions enriched with in silico predictions. An interaction network model for collagen V was created. We found that cEDS heterogeneous manifestations may be explained by the involvement in two different extracellular matrix pathways, related to cell adhesion and tissue repair or cell differentiation, growth and apoptosis.

4.1.3.1 Genotype/phenotype correlation

We aimed at characterizing different collagen V parts to analyze the pathologic effects of EDS correlated mutations. Based on literature data, pathogenic mutations were grouped according to their corresponding abnormal phenotype and used to perform our analysis. Collagen V chain $\alpha 1$ presents the highest number of pathogenic variants. These cause lack of collagen V secretion in the signal peptide (214). This specific condition is generally associated with haploinsufficiency of collagen V in the literature (215), thus they are not useful for the evaluation of the effects of collagen V structural impairments. The most frequent phenotypes associated with mutations affecting the N-terminus are characterized by hypermobility of joints and skin hyperextensibility (216), reduced deambulation, abnormal posture, neurological complications and reduced excretion and digestion (217, 218). Milder symptoms are associated with mutations located in the region not forming the triple helix (219, 220). The triple helix region appears to be related with the highest phenotypic variability, with a hypermobility Beighton score ranging from 0/9 to 9/9 (216). Of note, variants in the triple helix always present skin hyperelasticity. Bibliographic data for the C-terminal variants are not detailed enough to get more insights about related phenotypes (221). Less relevant information is present in the literature for chain $\alpha 2$ (221, 222). In addition, molecular details for compatibility with the triple helical structure are not currently available. However, the variability of symptoms reported for chain $\alpha 2$ associated variants seems to draw a scenario comparable to chain $\alpha 1$.

4.1.3.2 Model construction and evaluation of mutations

We modeled the collagen V triple helix region by homology, using the collagen I crystal structure (PDB code: 3HQV) as template. We found that variants affecting chain $\alpha 1$ can be grouped into two types. The first contains mutations interfering with triple helix packing. The main effect may be reduced chain compactness as the mutated amino acid side chains point to the geometric center of the molecule. Most pathogenic COL5A1 missense variants related to severe and complex cEDS phenotypes involve glycine residues, e.g. G1489D (c.4466G > A, 3) and G1564D (c.4691G > A, (223)). Glycine is known to play a key structural role in collagens as it allows the triple helix coil angle formation and the consequent right placement of helices. Less frequently, variants

4. RESULTS & DISCUSSION

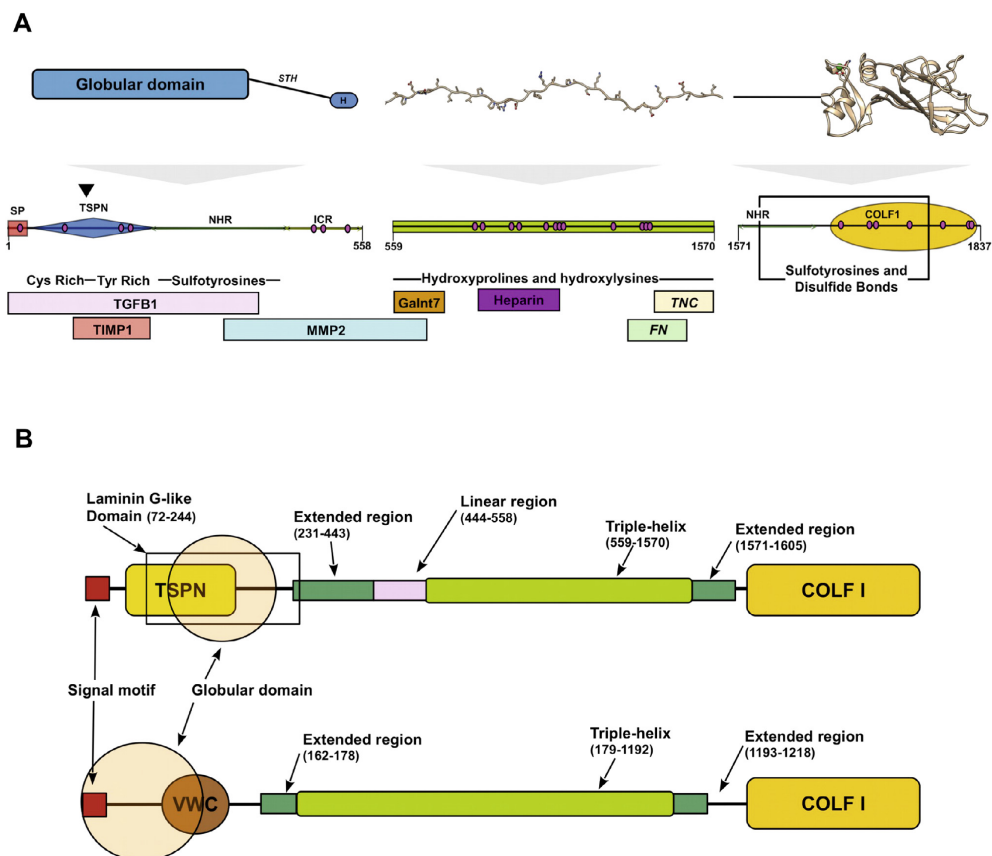


Figure 4.6: (A) Overview of collagen V chain structural organization. On top, collagen V is divided in three main regions depending on the functional specialization. A complex N-terminal region acting as regulative elements is opposed to a central region mainly structural and a C-terminal region though relevant for degradation. The globular domain (GD) stretching out from the triple helix axis to form a short rod (STH: short triple helices). Linker segment (H) connecting the globular domain to the extended region. 3D representation of chain alpha 1 forming the triple helix region (559-1570) followed by the C-terminal domain three-dimensional structure (PDB code: 4AE2). In the bottom part, functional domains composing the alpha 1 chain are presented. Variants discussed here are scattered on the entire molecule affecting the collagen V in total. Several protein-protein interaction domains were found in the globular region as well as on the triple helix. Signal peptide (SP), thrombospondin N-terminal-like (TSPN) domain, non-helix region (NHR), interrupted collagenous region (ICR), collagen C-propeptide (COLF I) domain (COLF1). Magenta dots show the relative mutations locations, while putative protein-protein interaction sites for which were identified a putative localization on collagen V structure are presented with colored bars, i.e. pleiotropic growth factors 1 (TGFB1), matrix metalloproteinase 1 (TIMP1), matrix metalloproteinase-2 (MMP2), N-acetylgalactosaminyltransferase 7 (Galnt-7), heparin binding site (heparin), tenascin C binding site (TNC), fibronectin binding site (FN). (B) Domain organization of the two chains forming collagen V. Schematic representation of collagen V alpha 1 (top) and alpha 2 (bottom) chains. The heparin binding domain overlaps the laminin G-like domain (TSPN), suggesting a regulative role for this region. Brown circle represents the Von Willebrand factor type C (VWC) oligomerization domain, while yellow box for the collagen C-propeptide (COLF I) domain. (175)

4.1 Protein tandem repeats characterization

of this type affect residues involved in the formation of stabilizing internal hydrogen bonds, such as E1292K (c.3874G > A, (224)). It is important to recall that collagen V contains two $\alpha 1$ chains. This peculiarity promotes a plethora of pathogenic phenotypes, as variants can independently localize on the first or second chain, either or both, depending on which of the two alleles (mutated or wild-type) is actively translated. Thus, the same collagen V variant may manifest different functionality depending on trimer composition. The second type of variant covers the ones that may disrupt the interaction between collagen V and other interacting proteins, e.g. E812D (c.2436A > T) and N951S (c.2852A > G) found in dbSNP (222). To validate this hypothesis, we used our model to map the interaction sites of different collagen V partners. Figure 4.6A summarizes the known position of the interactions detected along the triple helical region, while Figure 4.7A shows the network of all interactors known to bind this region. We found that E1292K (c.3874G>A) in one $\alpha 1$ chain forms an internal stabilizing bond. Instead, when the same variant affects two $\alpha 1$ chains, the long lysine side chain stretches from the groove between helices, facing the solvent. Further, E1292K localizes at the beginning of a larger area where frequency of hydrogen bonds stabilizing the triple helical fold decreases. This region is also known to form a fibronectin binding site (225). We found various sites for proline and lysine hydroxylation. In particular, collagen V hydroxyproline-rich regions are involved in protein self-assembly as well as interaction with PPII (polyproline-II), SH2 (Src homology 2) or SH3 (Src homology 3) domains (226). Mutation P1388S (c.4162C>T) is an example of proline substitution located in a poly-proline region. Similar features are reported in the literature for collagens and collagen V interactors (227). Analysis of chain $\alpha 2$ shows three structurally relevant glycine substitutions, i.e. G396R (c.1186G>C), G645R (c.1933G>A), G1146A (c.3437G>C). In addition, the P833L (c.2498C>T) mutation may interfere with two hydrogen bonds stabilizing the triple helical structure (Figure 4.6).

4.1.3.3 Detection and analysis of interactors

COL5A1 and COL5A2 genes harbor 90% of cEDS variant where the related phenotypes fulfill the three major diagnostic criteria (221). However, several cEDS patients harboring collagen V variants with milder manifestation are reported, i.e. not showing skin hyperelasticity or joint hypermobility. This observation suggested the hypothesis

4. RESULTS & DISCUSSION

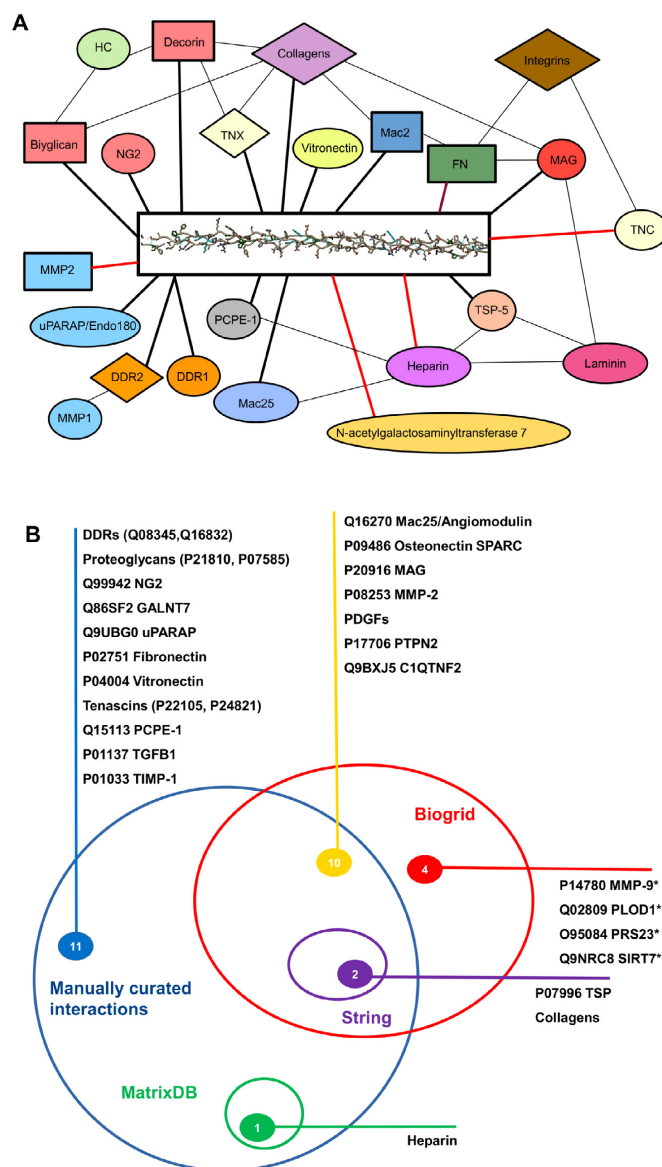


Figure 4.7: (A) Collagen V interaction network. The figure summarizes the collagen chain alpha 1 (V) interactors. Square boxes represent interactors found to correlate mainly with skin cEDS manifestations and diamond boxes the joint and skin” cEDS phenotype. Round boxes are interactors not connected with specific manifestations. Experimentally validated interactions are presented with red edges and black is used for the remaining interactions, i.e. predictions. Boxes colored with the same color belong to the same category. Proteoglycans, tenascin, discoidin domain containing receptors and matrix metalloproteinases are presented as pink, light yellow, orange and light blue boxes, respectively. (B) Venn diagram of collagen V interactors. The different interactors were grouped in respect to the considered source. UniProt (228) codes are also reported. (*) Interactors excluded due to lack of bibliographic data. (175)

4.1 Protein tandem repeats characterization

that simultaneous deregulation of different molecular pathways may promote cEDS insurgence. To confirm this hypothesis, we collected information about experimentally validated and high confidence predicted collagen V interactors from protein–protein interaction databases and manual search from the literature. A total of 24 different interactors were found (10, 2, 1 and 11 from BioGRID (229), STRING (230) and MatrixDB (231) and literature search, respectively). We found the most important interactions to be with collagens, in particular collagen I and collagen VI. Other relevant interactions are established during ECM organization and include proteoglycans, fibronectin and osteonectin for a total of 25 collagen V interactors (Table 4.1.3.4). The collected data were used to shed light on phenotypic cEDS variability. It was assumed that for two variants, one affecting collagen V and the other affecting one of its interactors, that correlate with the same specific phenotype, it can be speculated that the same biochemical pathway is compromised (i.e. the two molecules both play a key regulative role). Interestingly, we found that Fibronectin I and Tenascin X variants are related to other forms of EDS, where TGF-b1 variants are linked to Marfan Syndrome, a heritable disease sharing marked joint laxity. Collagen I variants are related to cEDS (232) and EDS VII (233) insurgence, as well as Osteogenesis Imperfecta (234), a defective connective tissue disease. Collagen II variants are related to severe arthritis and dysplasia development. Collagen VI is linked to UCDM (Ullrich congenital muscular dystrophy, (235), while ADAMTS mutations are causative of EDS VIIC. We then collected the OMIM annotation and category classifications of each symptom associated to interactors disruption, in order to group different interacting proteins by mutations/phenotype correlation. E.g. joint hypermobility was classified as a skeletal symptom and cigarette-paper scars as skin. Our statistical analysis resulted in two major clusters, one containing all the analyzed collagens, TGF-beta1, integrin alphaV-beta3, DDR2 and Tenascin-X. The common feature of non-collagenous proteins inside this cluster is that they intervene in cell migration, growth, differentiation and apoptosis. The diseases associated often affect joints structure and are related to dysplasia development. Interestingly, collagen V forms a separate sub-cluster with collagen XI. The two proteins are characterized by a high morphological affinity, in a murine model showing dosage compensation in tendons (215). Of note, we found integrin alphaV-beta3 in this cluster and not in the other, containing Fibronectin (FN) and the other integrins. Integrin alphaVbeta3 is indeed not typical for collagen V containing tissues,

4. RESULTS & DISCUSSION

but it is recruited in the EDS phenotype (236). Tenascin X was grouped with the collagen group, and this observation is confirmed by the fact that its variants are associated with some forms of EDS (237). The second cluster contains protein that are more strictly related to cell adhesion and tissue repair, as well as collagen assembly and binding. This proteins are MMP2, integrin alpha2beta1, Fibronectin, Mac25 and the proteoglycans biglycan and decorin. We identified this cluster as the skin one. The diseases associated often related to tissue repair defect and eye involvement.

4.1.3.4 Ehlers-Danlos association to Collagen V

Most variants affecting both COL5A1 and COL5A2 are associated with cEDS (238), a heterogeneous connective tissue disorder characterized by skin hyperextensibility, abnormal wound healing, and joint hypermobility (239). The disease is widely heterogeneous, with different manifestations even between relatives harboring the same mutation (240). Here, we provided a novel manually curated collection of collagen V missense variants causative of cEDS outcome annotated with the corresponding phenotypes. We performed an *in silico* investigation of collagen V associated cEDS manifestations building a homology model of the collagen V triple helix, which was used to map the pathogenic amino acid substitutions found in cEDS patients. Over the last two decades, different hypotheses were presented in the literature to explain disease variability (239). One explanation of this vast variability should be searched within collagen V interacting proteins. The ECM is characterized by many different proteins, which if mutated may in turn explain the contrasting cEDS phenotypes. In parallel, we created a manually curated dataset of phenotypes associated to the different variants. Our analysis confirmed the importance of the N-terminal domain as both initiator and organizer of collagen assembly. The central region shows variants affecting both the helical and non-helical sub-regions, the former have a stronger pathogenic effect. This finding confirms the well known structural role for this specific collagen V region. On the other hand, we found several putative binding sites of collagen V interactors to localize within the extended triple helix (Figure 4.7A) and not only within the N-terminal, suggesting a collagen V acting as a complex regulative scaffold protein. Indeed, the N-terminal domain seems to first initiate the assembling of ECM components, during this phase collagen may also serve as a scaffolding element. The resulting complex then relocates to the ECM, where it is buried within other collagen fibers to

4.1 Protein tandem repeats characterization

regulate their diameter. Hypermobility of joint and skin specific manifestations are frequently used as clinic criteria to separate different cEDS sub-types (238). Based on our results, cEDS phenotypes may be divided into two main classes, the first presenting joint hypermobility sometimes coupled with skin manifestation and a second class limited to skin manifestations. Our cluster analysis of interactors showed two well distinct molecular pathways, one relative to joint flexibility and the other to skin features. In other words, we suggest that skin-related phenotype is related to defects in ECM assembly and tissue repair, while joint flexibility is more generally associated to defects in cellular migration, differentiation, remodeling and proliferation. This observation also suggests that the interaction between collagen V impairments and variants of different proteins connected with collagen pathways may explain the extremely variable phenotypes observed in cEDS patients.

4. RESULTS & DISCUSSION

Table 4.3: Collagen V interactors. (*) Not shown in Figure 2 (175)

Interactor	Test	Pathway notes	Interaction data	Ref	Associated disease
DDR1	In tissue	Stimulates tyrosine autophosphorylation in SH2 or PTB docking sites. Promotes integrin recruitment	N or O glycosylated carbohydrates are required for binding	1	
DDR2	In tissue	Promotes MMP1 expression	N or O glycosylated carbohydrates are required for binding	1	Mutations are associated to spondylometaepiphyseal dysplasia
Biglycan	In vitro	Biglycan and decorin accelerate thrombin inhibition by heparin cofactor II after binding collagen V	Both core and GAG chains are required binding the collagen triple helix	2	The protein disruption leads to an osteoporosis-like phenotype
Decorin	In vitro	Biglycan and decorin accelerate thrombin inhibition by heparin cofactor II after binding collagen V	The interaction is located in the protein core. Binding site on collagen triple helix	2	Mutations are associated to congenital corneal dystrophy
NG2	In vitro	Proteoglycans are putative coreceptors, which act together with other matrix molecules	The central domain flanks with collagen	3	
Galnt7*	In vitro	Involved in Olinked oligosaccharide biosynthesis	The Nterminal propeptide interacts with the transferase	4	
TSP5	In vitro	Thrombospondin promotes platelets aggregation acting as link between collagen V and the basal lamina components	Collagen V specific thrombospondin binding site. heparin and fucoidin competing for the interaction	5	A functional polymorphism interferes with MMP binding, promoting the susceptibility to slipped disc
uPARAP	In vitro	Involved in ECM degradation, cellular adhesion and signal transduction	uPARAPs binds the fibronectin-type II domain (triple helix)	6	

Continued on next page

4.1 Protein tandem repeats characterization

Table 4.3 – continued from previous page

Interactor	Test	Pathway notes	Interaction data	Ref	Associated disease
Fibronectin	In tissue	Fibronectin regulates biological processes such as cellular adhesion, migration, proliferation. It also recruits the integrins	Differential splicing of FN is regulated by collagens, which stimulate the EDA + production	7	Mutations are associated with glomerulopathy, fibronectin deposition and FNdisfunction related EDS
MAG	In vitro	MAG plays a relevant role during cellular adhesion	MAG binds the collagens (alpha chain) and heparin	8	
Mac25/AGM	In vitro	Multifunctional protein expressed in secondary lymphoid tissues (blood vessels)	Mac25 binds the collagens (II and V) and heparin	9	Mutations lead to retinal arterial macroaneurysm with supra-valvular pulmonic stenosis
Heparin	In vitro	An essential molecule of ECM. Bind most collagen V interactors	Heparin binds the collagen V HepV motif (alpha chain)	10	
Vitronectin	In vitro	Glycoprotein promoting cellular adhesion and diffusion	Vitronectin binding competes with FN	11	Various phenotypes (e.g. adiposity, cataract, myeloids) not related to precise mutations
Osteonectin SPARC	In vitro	Promotes plasminogen conversion in plasmin (thrombolytic agent). It is also secreted by osteoblasts during bone formation	Osteonectin binds the first 17 residues of the collagen V Nterminus	12	
Tenascin X (TNX)	In vitro	Regulates collagen interfibrillar distance. It is relevant for the ECM organization	The complete collagen V trimer is required for the binding. The interaction is thought indirect and involving decorin	13	

Continued on next page

4. RESULTS & DISCUSSION

Table 4.3 – continued from previous page

Interactor	Test	Pathway notes	Interaction data	Ref	Associated disease
Tenascin C (TNC)	In vitro	Extracellular matrix protein driving the neuron and axon migration, synaptic plasticity and neuronal regeneration	The interaction involves the collagen V Nterminus, FN typeIII and TNC	4	Mutations are associated with asthma and allergies related traits
Collagen alpha1 (I)	In vitro	Collagen V regulates the collagen I deposition	The Nterminal propeptide interacts with the alpha1 (1) chain	4	Caffey disease, EDS (I and VIIA), OI (I, II, III, IV), bone mineral density variation and osteoporosis
Collagen alpha2 (I)	In vitro	Collagen V regulates the collagen I deposition	The Nterminal propeptide interacts with the alpha2 (1) chain	4	Several diseases such as arthritis, dysplasia and Stickler syndrome
Collagen alpha1 (VI)	In vitro	Collagen VI is abundant into the pericellular environment, where it acts as a scaffold elements	The interaction occurs between the collagen V Nterminus and the globular domain of collagen VI	4	Bethlem myopathy, Ullrich congenital muscular dystrophy and the posterior longitudinal spinal ligaments ossification
MMP2	In vitro	Degrades denatured collagens and TGF-beta1	MMP2 binds the collagen alpha 1 Nterminus (haemoplexinlike domain)	4	Mutations lead to TorgWinchester syndrome
TIMP1	In vitro	TIMP1 expression is regulated by TGF-beta1	The interaction involves the TIMP1 NTR domain and collagen V Nterminus	4	

Continued on next page

4.1 Protein tandem repeats characterization

Table 4.3 – continued from previous page

Interactor	Test	Pathway notes	Interaction data	Ref	Associated disease
PDGFs*	In vitro	Growth factors/cytokines modulated by interacting with ECM components	PDGF AA, BB and AB bind collagens, however they show low affinity for collagen V	14	Basal ganglia calcification, dermatofibrosarcoma, meningioma, gastrointestinal stromal tumor and somatic, hypereosinophilic syndrome
PTPN2*	In vitro	PTPs are known to be signaling molecules regulating cell growth, differentiation, mitotic cycle and oncogenic transformation	Data derived from high throughput affinitypurification mass spectrometry experiments	15	
C1QTNF2*	In vitro		Data derived from high throughput affinitypurification mass spectrometry experiments	15	

4.2 Protein tandem repeats identification and annotation

This chapter describes the curation and improvement of RepeatsDB database with respect to the first published version. A detailed structural characterization of repetitive elements was largely missing, as repeat unit annotation in RepeatsDB was manually curated and covered only 3% of bona fide TRPs at the time. This is the reason why we developed Repeat Protein Unit Predictor (ReUPred, section 4.2.1), a novel method for the fast automatic prediction of repeat units and repeat classification using an extensive repeat unit library derived from curated data in RepeatsDB 1.0. ReUPred (algorithm described in 3.1.1) uses an iterative structural search against the library to find repetitive units on target structures. We tested the method on solenoid proteins, i.e. the most canonical repeat structures constituted by the superhelical arrangement of simple and short repeating units. The accurate prediction of repeat units from ReUPred was exploited to increase the number of annotated repeat units in RepeatsDB by an order of magnitude comparable to the sequence-based Pfam classification. We presented the second release of RepeatsDB database (section 4.2.2). A substantial growth of repeat unit characterization that was possible by applying the ReUPred algorithm over the entire Protein Data Bank (PDB), indeed RepeatsDB now features information on start and end positions for the repeat regions and units for all entries. A new classification level has been introduced on top of the existing scheme, as an independent layer for sequence similarity relationships. The quality of the data is guaranteed by an extensive manual validation of ReUPred predictions for more than 60% of the entries. RepeatsDB is continuously updated, and therefore requires a continuous effort in the manual curation. To facilitate this process we designed RepeatsDB-lite (described in section 4.2.3), web server for the prediction and refinement of TR in protein structure. It takes advantage of ReUPred algorithm and an extended library that covers all different TR classes. The algorithm is described in section 3.1.2, and includes updates aimed at increasing the predictor speed and minimizing errors. The web interface allows to predict the position of repeat units and visualize similarity relationships between them at both the sequence and structure level, it also allows an intuitive revision of the prediction and submission of reviewed entries to RepeatsDB. The server represents a platform to harness community annotation efforts, which have been proven to be effective in RepeatsDB experience.

4.2.1 Identification of repetitive units in protein structures with ReUPred

In section 4.1.1 we provided a first look into the relationship between repeat structures (RepeatsDB subclasses) and their Pfam families. In the majority of cases a strict one-to-one relationship was found, with the expected tendency for structure to be more conserved than sequence in the remaining cases. The LRR example however shows that it is also possible for members of a large family to fall into different structural classes. In order to expand RepeatsDB dataset and get a better understanding of repeat protein evolution, we developed a method for Repeat Unit Prediction (ReUPred). ReUPred (algorithm described in 3.1.1) was developed to predict both unit position and classify repeat proteins to automate the time-consuming manual annotation process of detailed annotation in RepeatsDB 1.0. See Figure 4.8 for an example on plakophilin-1. Before benchmarking the main novel features, it is worthwhile to investigate whether ReUPred is able to correctly discriminate real repeats from non-repeat proteins. For this purpose, it has been compared with RAPHAEL (114) on the original datasets (see Table 4.4). ReUPred correctly classifies 324 out of 352 domains (92% accuracy). This is only somewhat lower than RAPHAEL on the same dataset (94.9% and 95.7%, for $S > 0$ and $S > 1$, respectively). A higher specificity could be obtained for ReUPred by setting a stronger filter on the last step of the algorithm, but that would affect coverage on the positive dataset. Even though ReUPred was designed to predict unit positions in tandem repeat proteins and not extensively optimized for repeat detection, this result demonstrates that the tool is also effective in discriminating repeat/non-repeat proteins.

Method	TP	FP	TN	FN	Solenoids	Non-solenoids
RAPHAEL ($S > 0$)	94	7	240	11	89.5	97.2
RAPHAEL ($S > 1$)	91	1	246	14	86.7	99.6
ReUPred	81	4	243	24	77.1	95.3

Table 4.4: The percentage of correctly classified solenoids and non-solenoids is shown together with the component true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). RAPHAEL is shown with the two SVM cutoff values as reported in the original paper. (176)

4. RESULTS & DISCUSSION

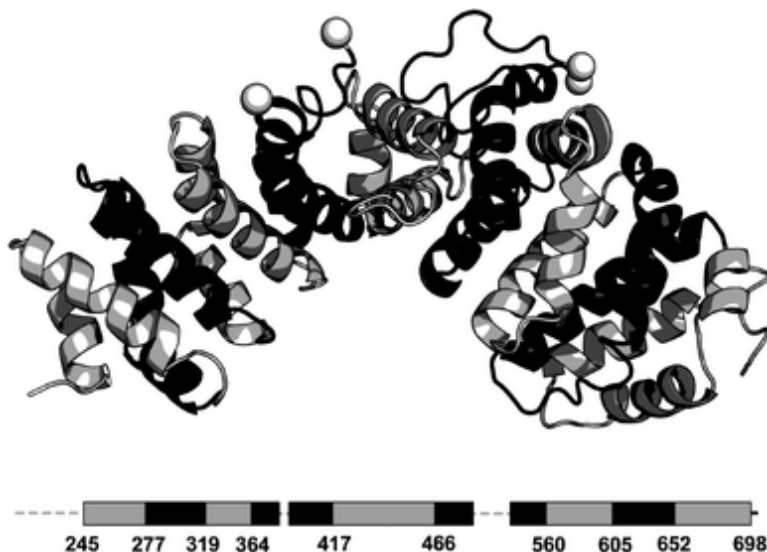


Figure 4.8: ReUPred unit prediction for Plakophilin-1 (PDB code 1XM9, chain A). The structure is shown in cartoon representation in the top part with the schematized sequence below. Predicted units are represented in black and gray. Dashed lines represent missing residues in the PDB file (residues 388396 and 481508). The N- and C-terminal residues flanking the missing residues are shown as spheres in the structure. (176)

4.2.1.1 Repeat classification

ReUPred predicts units and fine classification for 83% (893 proteins) of the RepeatsDB 1.0 classified set. The class assignment is obtained by simply transferring this information from the master unit found in SRUL. This approach has been proven to be effective as shown in Table 4.5. ReUPred works very well for the α class (III.3 in RepeatsDB). Instead, it is more difficult to correctly assign α/β and β examples. The low recall indicates that the cause of the problem is detecting units that do not have a good template in SRUL. This is an important result, as it indicates which RepeatsDB 1.0 entries are worth manually annotating at the detailed level to improve ReUPred sensitivity and SRUL representation of the repetitive structural element universe. Low precision for β and α/β classes is due to a high number of false positive assignments. Looking at the data in detail, we found some ambiguous class assignments, e.g., PDB code 3ZYI, chain A, is annotated as α/β solenoid in RepeatsDB, but there are no helix elements except for a small fragment (residues 309–318) which is not repeated in the units. Since ReUPred predicts the class by transferring annotation from SRUL, if a SRUL element is misclassified the error propagates. ReUPred could be very useful to

4.2 Protein tandem repeats identification and annotation

guide the manual refinement of RepeatsDB class annotations.

Class	Recall	Precision	F-Measure	Accuracy
All- β	0.81	0.74	0.78	0.63
Mixed α/β	0.55	0.65	0.60	0.43
All- α	1.00	0.99	1.00	0.99
Total	0.94	0.94	0.94	0.89

Table 4.5: ReUPred classification performance on the RepeatsDB 1.0 classified dataset. See section 3.1.1 "Performance evaluation" for details on the measures used. (176)

Class	Method	Recall	Precision	F-Measure	Accuracy
All- β	TAPO	0.47	0.59	0.53	0.47
	ConSole	0.39	0.69	0.50	0.46
	ReUPred	0.62	0.64	0.64	0.56
Mixed α/β	TAPO	0.66	0.70	0.68	0.59
	ConSole	0.62	0.69	0.66	0.57
	ReUPred	0.84	0.84	0.84	0.78
All- α	TAPO	0.64	0.78	0.70	0.57
	ConSole	0.50	0.74	0.59	0.46
	ReUPred	0.74	0.79	0.74	0.62
Total	TAPO	0.58	0.70	0.64	0.53
	ConSole	0.48	0.71	0.58	0.49
	ReUPred	0.71	0.75	0.73	0.62

Table 4.6: Performance evaluation is reported for each method on all RepeatsDB 1.0 solenoid structures (All) and for the three subclasses separately (β , α/β and α). The best value for each quality measure is shown in bold. See section 3.1.1 "Performance evaluation" for details on the measures used. (176)

4.2.1.2 Unit prediction accuracy

ReUPred has been evaluated for unit prediction using the metric described in section 3.1.1, i.e., penalizing predictions with a wrong phase or/and a wrong length. Table 4.6 shows a comparison with TAPO and ConSole in terms of predicted repeat residues on the detailed RepeatsDB 1.0 set. The results are reported for each of the three main solenoid classes and for all proteins together. ReUPred always outperforms the other methods for all evaluation measures. In particular, the greatest improvement is observed for the α/β subclass, with an increase of 19% accuracy compared with

4. RESULTS & DISCUSSION

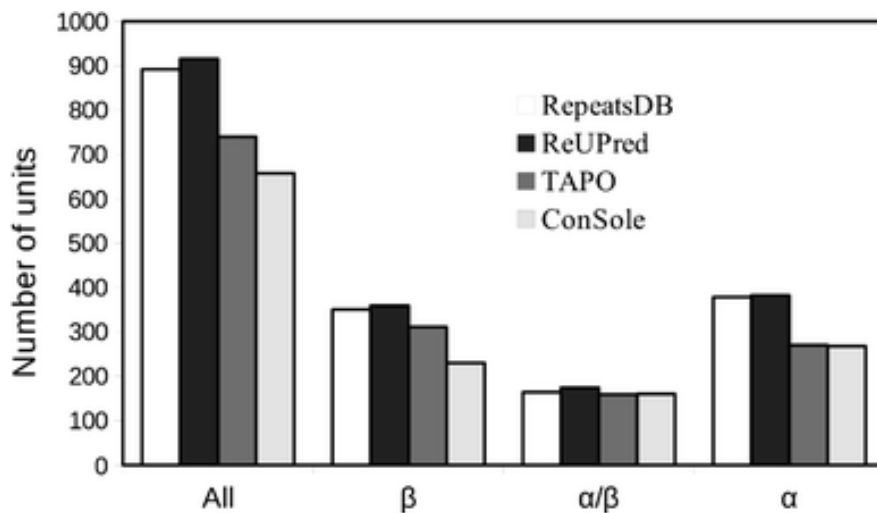


Figure 4.9: The number of predicted units on the RepeatsDB 1.0 detailed dataset. The manually curated reference (RepeatsDB) is shown next to the three prediction methods. ReUPred predicts more repeat units than the other two methods

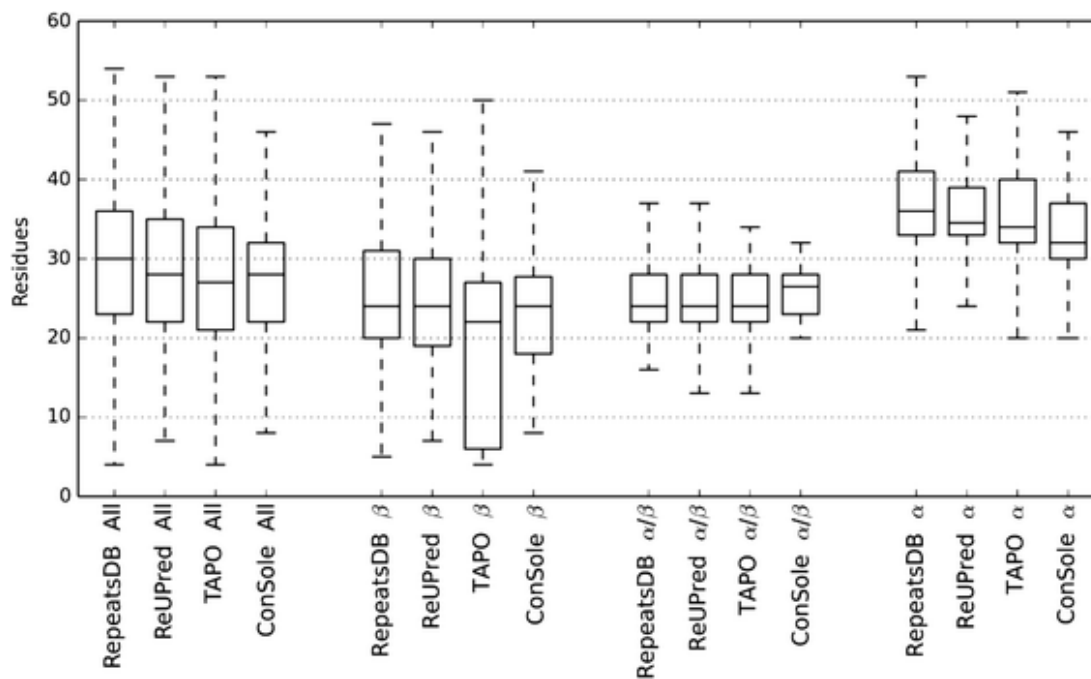


Figure 4.10: Repeat unit periodicity box plot distribution on the RepeatsDB 1.0 detailed dataset. The manually curated reference (RepeatsDB) is shown next to the three prediction methods. (176)

4.2 Protein tandem repeats identification and annotation

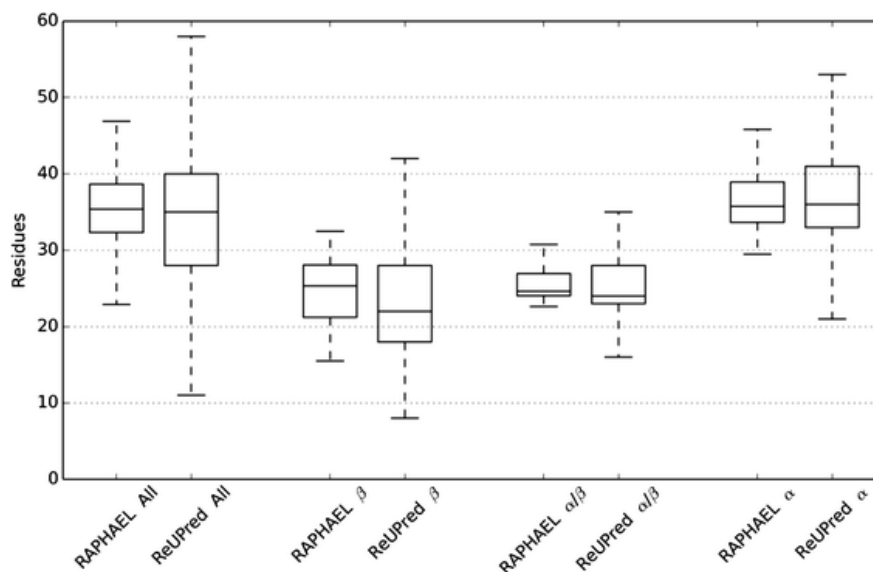


Figure 4.11: Large-scale periodicity predictions on the RepeatsDB 1.0 classified dataset. The original RAPHAEL periodicities are compared to ReUPred unit lengths as box plot. (176)

TAPO. The high accuracy for this class can be explained by the fact that mixed α/β units represent more structurally complex elements compared to all- α units. More information is coded in the structure unit, making it easier to discriminate wrong structural alignments. On the other hand, the most problematic subclass is all- β . Both recall and precision are lower for all methods compared with other subclasses. This may be explained by the fact that β solenoid units are more degenerated in the same protein than other solenoids and present a greater structural diversity with many insertions (data not shown). Moreover, they are shorter compared with all- α , generating worse structural alignments.

In addition to evaluating repeat annotations at the residue level, it is of interest to benchmark repeat units and their length distributions. Figure 4.9 shows the number of repeat units being identified by each method. Here again, ReUPred predicts more units than the other two methods. Both ConSole and TAPO generate units with the same size for a given structure and this may limit their ability to deal with insertions in solenoid proteins. ReUPred may therefore be better able to adapt to the irregular aspects of solenoid repeats. Figure 4.10 shows a box plot for the distribution of the predicted repeat periodicities against the RepeatsDB 1.0 classified set. The median

4. RESULTS & DISCUSSION

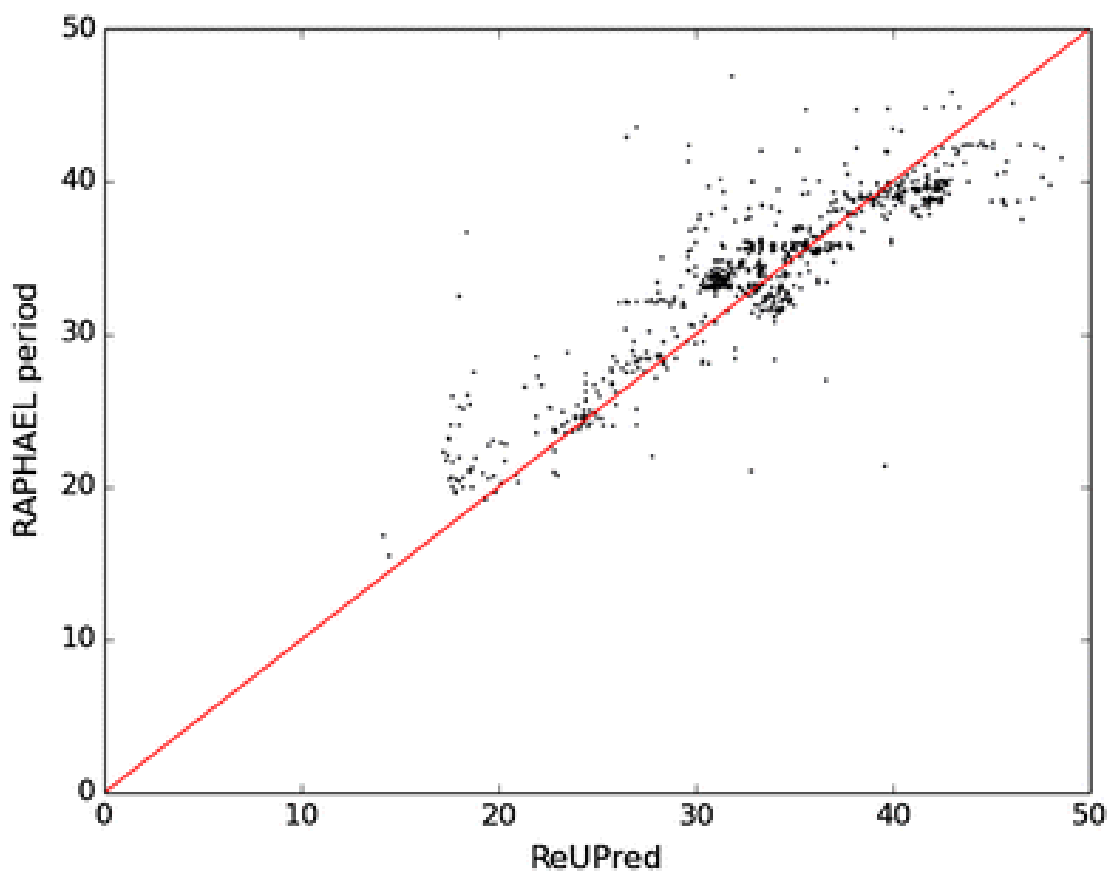


Figure 4.12: Scatter plot of RAPHAE and ReUPred periodicities on the RepeatsDB 1.0 classified dataset. RAPHAE produces a single periodicity per protein, whereas all predicted units were considered for ReUPred. (176)

repeat length and standard deviations of ReUPred are very similar to the reference definition and on average match better than TAPO and ConSole. TAPO appears to underpredict the repeat length in β structures, probably because it also uses sequence information. ConSole on the other hand appears to have more difficulties with α -helices.

4.2.1.3 Expanding the universe of known solenoids

Given the good performance of ReUPred for its intended purpose, i.e., classifying solenoid repeats and annotating their component units, it can be used to automatically expand the knowledge contained in RepeatsDB 1.0. The first step consists in establishing the baseline against the existing RAPHAE annotations on the classified dataset. This contains annotations for solenoid class and predicted average repeat

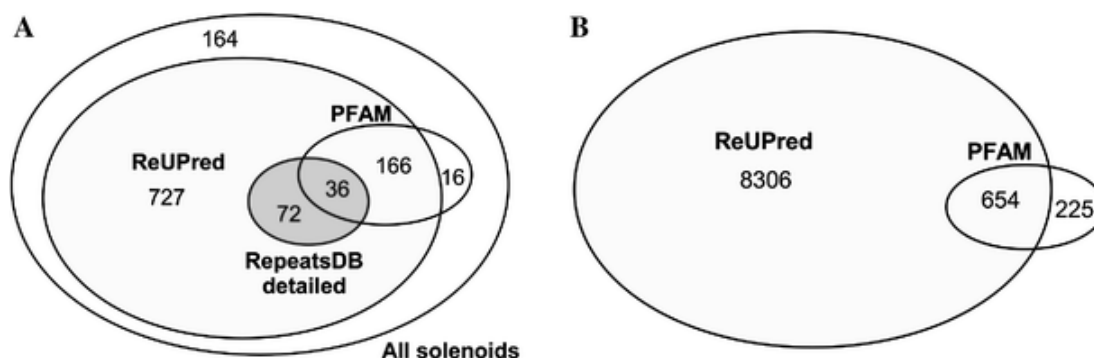


Figure 4.13: Venn diagram of available annotations for RepeatsDB 1.0 classified dataset. (a) Comparison of proteins with bona fide solenoid assignments. (b) The number of annotated repeat units in the dataset. The total number of repeat units in the dataset is unknown. ReUPred is able to increase the annotation by an order of magnitude in both cases. (176)

length. Since this dataset does not provide unit annotation, the simplest way to evaluate the performance is to compare the length of the predicted units with the repeat period predicted by RAPHAEL. This is the number of residues for which the symmetry signal is maximized, generating a single period for each protein. This is a big limitation, as it does not reflect the real situation where unit sizes vary inside a protein due to insertions which are frequent in solenoids. In particular, it is very relevant for the all- β class where almost all proteins have insertions. Figure 4.11 compares the distribution of ReUPred predicted unit length and RAPHAEL period for each solenoid class. Overall, both are very similar, with ReUPred having a wider range of periodicities as it is able to recognize irregularities in single repeat units. Only the distributions for all- β repeats differ more markedly. This class contains many structures with insertions which RAPHAEL struggles to summarize in a single fixed periodicity.

The scatter plot in 4.12 shows the correlation between the RAPHAEL period and ReUPred mean unit length calculated on each predicted protein. The two methods correlate strongly, with a Pearson correlation coefficient of 0.88 (P value = $4.59 \cdot 10^{290}$). On average, ReUPred predicts shorter units than the RAPHAEL period, 33.7 (SD 6.5) and 34.2 (SD 5.3) residues, respectively. When the RAPHAEL period is much larger (extreme points above the diagonal), ReUPred wrongly predicts two units instead of a single unit which would better represent the repetitive symmetry (e.g., PDB code 3L3F, chain X). For opposite cases, the contrary happens, i.e., ReUPred predicts a pair of units as a single element (e.g., PDB code 3PET, chain A).

4. RESULTS & DISCUSSION

To expand the annotation in RepeatsDB, ReUPred was used to predict all repeat units for classified RepeatsDB solenoids. Since there is no comparison and no structural validation is possible, we chose to compare the annotation to Pfam. Figure 4.13 shows the very substantial increase in annotations both in terms of bona fide solenoid proteins and especially in the number of identified repeat units. The latter yields an increase of an order of magnitude compared to state-of-the-art sequence-based annotation in Pfam.

4.2.1.4 Benchmarking

At the time of ReUPred design, detailed unit annotation was available in RepeatsDB for only 3% of the total putative repeat protein structures. ReUPred provides both the prediction of repetitive units and a finer classification in the RepeatsDB classification scheme. The algorithm works by exploiting a structure repeat unit library (SRUL) and an iterative decomposition of the input structure. While the performance was tested on the solenoid class, the method also works for other repeat types. ReUPred has been compared with other state-of-the-art methods, TAPO and ConSole, adopting an evaluation metric which takes into consideration both phase and size of the predicted units. Testing on a manually curated dataset obtained from the "detailed" RepeatsDB entries, ReUPred achieved the highest accuracy for all types of solenoids (β , α/β and α) with an overall increase of 9% over TAPO and 13% over ConSole. To provide an extended evaluation, a larger dataset with classified RepeatsDB entries without unit annotation was used. It was possible to test ReUPred ability of classifying solenoid structures and the correlation with periods predicted by RAPHAEL. ReUPred extended unit annotation and classification for almost all solenoids with high precision and accuracy. Moreover, the average unit length predicted by ReUPred strongly correlates with RAPHAEL, confirming the high quality of the predictions. Mixed α/β units are underrepresented in SRUL compared to the α and β classes, meaning that extending SRUL could yield a better recall and higher accuracy. ReUPred has also the ability to detect unit diversity inside a given target protein, recognizing fragment insertions that are not part of the repeat elements.

4.2.2 RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures.

4.2.2.1 Database description

RepeatsDB 2.0 data have been completely regenerated taking advantage of the new ReUPred predictor (176) for automatic detection of tandem repeat units. In the new database version all entries are annotated at the unit level, i.e. providing start and end position for each repeated segment, and classified at the subclass level. Compared to the old version, unit annotations have grown by more than an order of magnitude.

Data curation The initial dataset for RepeatsDB is the entire PDB (241). Repeat candidates are extracted with RAPHAEL (114) and processed with ReUPred (176) to confirm the presence of repeat regions and provide detailed unit information. ReUPred is able to identify the position of repeated fragments and to assign the class and subclass by transferring this information from the unit library. The final dataset available in RepeatsDB 2.0 is the result of an iterative process where the ReUPred library has been refined manually multiple times to resolve conflicts, improve its ability to generalize and include newly discovered subclasses. At the end of the process, an extensive validation and refinement of the predictions has been carried out by expert visual inspection. More than 60% of the entries have been reviewed and five new subclasses created, three for class IV (closed structures) and two for class V (beads on a string).

Innovations Apart from the new annotation pipeline, many improvements have been introduced since the last RepeatsDB release. All positional annotations are now based on SIFTS (242), making them consistent with both PDB (241) and UniProt (228) references. The search engine has been completely redesigned. An intuitive search interface allows to perform complex queries using logical operators and guides the user through all possible searching fields. A new classification level has been added to include evolutionary relationships among different repeat regions. An all-vs.-all alignment of the repeat regions allowed to group them according to sequence similarity and to identify different repeat families. The new classification has been implemented as an independent layer on top of the existing structural features, and is available at three different identity thresholds (40%, 60% and 90%). The web interface allows to

4. RESULTS & DISCUSSION

navigate entry clusters, providing an overview of the representative sequences inside each structural subclass.

4.2.2.2 Database usage

The user interface presents an intuitive summary table providing direct access to all entries by structural class directly from the home page. For a finer search, the user can visit either the Browse page providing subclass access or the Search page for generating complex queries (Figure 4.14 and 4.15, top). All entry points redirect to the same result page listing the retrieved proteins in a table (Figure 4.15, bottom). The table can be further filtered by providing additional matching strings in the column headers. The Browse page also provides direct access to sequence clusters, where entries are grouped by sequence similarity. The redesigned entry page (Figure 4.16) is much more informative compared to the previous RepeatsDB version, including several cross-links to third party resources. It also integrates several structural features useful for comparing CATH, SCOP, Pfam and DSSP annotations with RepeatsDB data. Regions, units and insertions are provided for all entries and correctly mapped both to UniProt and PDB reference (SEQRES field in the PDB file) sequences thanks to the SIFTS service. The correct mapping can strongly improve RepeatsDB impact since it is now very easy to link repeat data with other sequence features like mutations or post-translational modifications. Thanks to a RESTful architecture, all RepeatsDB data are accessible from external APIs and third party resources through HTTP URLs. Customized datasets can be downloaded in JSON or text format using the browse function or RESTful web services.

Statistics RepeatsDB provides high quality annotation for 5400 entries. Figure 4.17 compares the current RepeatsDB content to the previous version. The chart shows the total number of entries belonging to each class. However, the new version provides unit definition and subclass classification for all entries where the old version annotated only a tiny fraction (327 entries, cyan bar). Moreover, in RepeatsDB 2.0 more than 60% of the entries have been manually reviewed by expert curators (blue segment). Further details such as the number of regions, units and genes are available from the Stats page of the web site.

4.2 Protein tandem repeats identification and annotation

The screenshot shows the RepeatsDB website interface. At the top, there is a navigation bar with the RepeatsDB logo, a home icon, and links for 'Browse', 'Search', 'About', 'Help', and 'Stats'. A search box is also present. Below the navigation bar, there are two tabs: 'RepeatsDB classes' (selected) and 'RepeatsDB clusters'. The main content area is titled 'Currently available regions' and displays a table of protein tandem repeat classes and their units.



























Currently available regions			Units
II	Fibrous structures stabilized by interchain interactions		12
	II.1	collagen triple-helix	3 
	II.2	α helical coiled coil	9 
III	Elongated structures whose repeat units require one another to maintain structure		2388
	III.1	β-solenoid	499 
	III.2	α/β solenoid	497 
	III.3	α-solenoid	1329 
	III.4	trimer of β spirals	24 
	III.5	single layer anti-parallel β	39 
IV	Closed structures whose repeat units need one another to maintain structure		2883
	IV.1	TIM-barrel	1088 
	IV.2	β-barrel	116 
	IV.3	β-trefoil	74 
	IV.4	β-propeller	1188 
	IV.5	α/β prism	190 
	IV.6	α-barrel	22 

Figure 4.14: RepeatsDB data can be retrieved in three different ways. The Browse page provides the entry point for both the structural hierarchy and sequence clusters. (173)


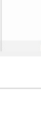

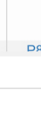


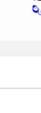
4. RESULTS & DISCUSSION

The top part of the image shows the RepeatsDB search interface. It includes a navigation bar with 'Browse', 'Search', 'About', 'Help', and 'Stats'. Below this is a search form with the following fields:

- Source database: RepeatsDB (dropdown)
- Field: No.units (dropdown)
- From: 6 (input)
- To: ∞ (input)
- Operator: - (dropdown)
- Boolean operator: AND (dropdown)
- Source database: PDB (dropdown)
- Field: Source organism (dropdown)
- Value: homo sapiens (input)
- Operator: - (dropdown)
- Boolean operator: AND (dropdown)
- Source database: Cross-reference (dropdown)
- Field: Pfam (dropdown)
- Value: PF13181 (input)
- Operator: - (dropdown)
- Boolean operator: + (dropdown)

Buttons for 'Reset' and 'Search' are located below the form.


The bottom part of the image shows the search results page. It includes a 'Download' button and tabs for 'Regions', 'PDB Chains', and 'UniProt entries'. The results are displayed in a table with the following columns:


Rev. \uparrow_1	RDB ID \downarrow	Structure Title \downarrow	Length \uparrow_3	Class \uparrow_2	UniProtID \downarrow	Image \downarrow
★	3sf4A	Crystal structure of the complex between the conserved c...	406	III.3	P81274 ↗	
★	4ay6A	Human O-GlcNAc transferase (OGT) in complex with UD...	723	III.3	O15294 ↗	
★	3ro2A	Structures of the LGN/NuMA complex	338	III.3	Q8VDU0 ↗	
★	4g11B	Crystal structure of interferon-stimulated gene 54	472	III.3	P09913 ↗	
★	1w3bA	THE SUPERHELICAL TPR DOMAIN OF O-LINKED GLC...	388	III.3	O15294 ↗	
★	4wndA	Crystal structure of the TPR domain of LGN in complex wi...	406	III.3	P81274 ↗	
★	1fchB	CRYSTAL STRUCTURE OF THE PTS1 COMPLEXED T...	368	III.3	P50542 ↗	
★	4wnnA			III.3	P81274 ↗	

At the bottom of the table, there is a pagination control showing '1 / 7' items per page and '1 - 10 of 70 items'.

Figure 4.15: RepeatsDB data can be retrieved in three different ways. (TOP) the Search page allows the user to perform advanced queries against a range of RepeatsDB-specific and third-party search fields. The input can be simple text or numeric (single value or range) according to the field type and multiple queries can be combined by boolean operators (AND, OR, NOT). Both the Browse and Search pages redirect to the results page. (BOTTOM) This page provides a list of retrieved entries and can be further filtered (and sorted) through column header fields. Results can be displayed by PDB chain (default), region or UniProt. (173)


4.2 Protein tandem repeats identification and annotation

 RepeatsDB Browse Search About Help Stats

A  **1ialA** IMPORTIN ALPHA

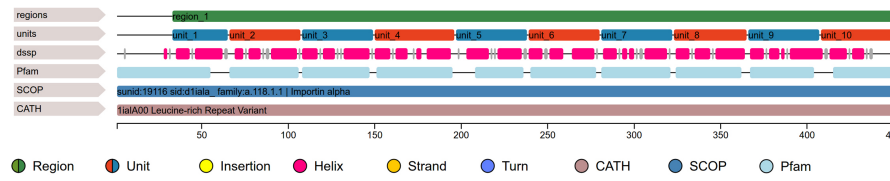
Title IMPORTIN ALPHA, MOUSE
Organism Mus musculus **Expression Host** Escherichia coli BL21(DE3) **Sequence length** 453
Cross-references PDB: 1ial; UniProt: P52293; MobiDB: P52293; SCOP: 19116; CATH: 1ialA00; Pfam: PF01749.16 PF00514.19 PF16186.1


B

Region	Classification	Start	End	Units	Period	Sequence clusters
1	III.3 α -solenoid 	76	496	10	41.10	RCL40_177 RCL60_189 RCL90_218

C Feature viewer

ZOOM POSITION



D  **453** Sequence viewer Structure viewer

```

1  DEQMLKRRNV SSFPDDATSP LQENRRNQGT VWWSVEDIVK
41  GINSNNLESQ LQATQAARKL LSREKPPID NITRAGLIPK
81  FVSLGKTDG SPIOFESAWA LTNIASGTSE QTKAVVDGGA
121 TPAFISLLAS PHAHISEQAV WALNGIAGDG SAFRDLVIKH
161 GAIDPLLALL AVPDLSLAC GYLRNLWTL SNLCRNKNPA
201 PPLDAVEQIL PTLVRLHHN DPEVLADSCW AISYLTGPN
241 ERLEMVVKKG VVPOLVLLG ATELPVTPA LRAIGNIVTG
281 TDEQTKVID AGALAVFPLG LTNPKNIOK EATWMSNIT
321 AGRDQIQVQ VNHGLVPLV GVLKADFVK QKEAAWAITN
361 YTSGGTVEQI VYLVHGLIE PLMNLSSAKD TKIQIVLDA
401 TSNIFQAEK LGETEKLSIH IEECGLDKI EALQRHENES
441 VYKASLNLTIE KYF
    
```

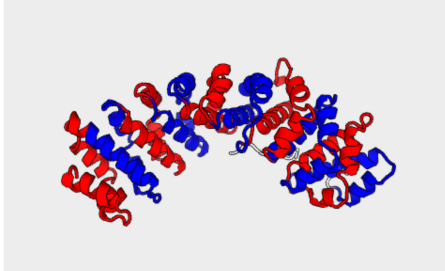


Figure 4.16: Screenshot of RepeatsDB sample entry page for PDB code 1ialA. The top part of the page (A) reports structure information from the PDB and cross-references to third-party databases including UniProt, MobiDB, SCOP, CATH and Pfam (when available). RepeatsDB annotations are available for download both in text and JSON formats on the top-right corner. (B) A table provides region details such as structural classification, start/end position, number of units, repeat period and cluster families. (C) The feature viewer summarizes available annotation for the PDB reference sequence, i.e. the SEQRES field in the PDB file. An overview of RepeatsDB information (regions, units and insertions) along with secondary structure (DSSP), Pfam, SCOP and CATH tracks (when available) are shown. (D) A detailed view of RepeatsDB annotations is highlighted in the sequence and PDB viewers. (173)

4. RESULTS & DISCUSSION

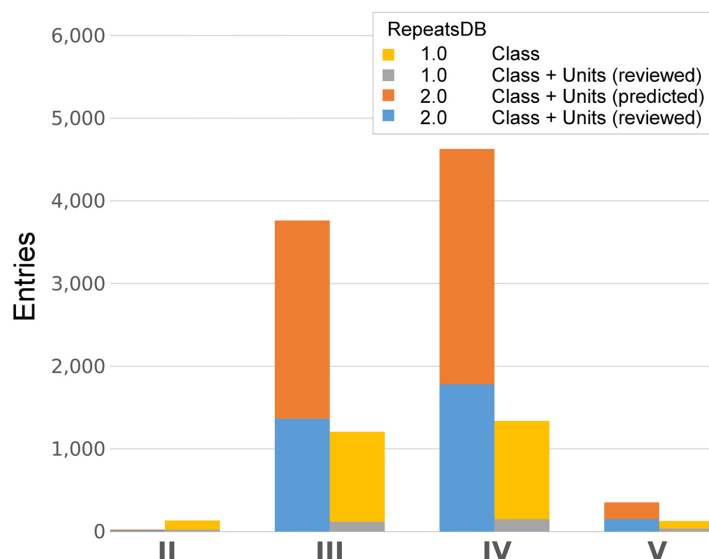


Figure 4.17: RepeatsDB 2.0 is compared to the previous release. Entries have unit and subclass annotation, with more than 60% manually reviewed (blue). For the old version, only a tiny fraction of entries have unit definition (cyan) and the rest is mostly annotated only at the class level (yellow). (173)

4.2.3 RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins

Since community annotation efforts have been proven to be effective in RepeatsDB experience, we developed RepeatsDB-lite, an interactive web server designed for the detection, classification and refinement of repeated structural modules from PDB files. RepeatsDB-lite (algorithm described in 3.1.2) extends the ReUPred algorithm to all TR types and strongly improves the performance both in terms of computational time and accuracy. RepeatsDB-lite takes a PDB structure in input and predicts TR units and the repeat classification along the RepeatsDB schema (class, subclass, type and fold). The server accepts either a PDB identifier (ID) or file. By default, the predictor considers only the first PDB chain. Alternatively, the user can specify the chain ID or an all chains mode. Submitted jobs can be retrieved using the search box or bookmarking the result page URL. The RepeatsDB-lite output page features an intuitive visualization of predicted TR regions and units. In addition, another page allows the user to modify the prediction and visualize the effect on the unit alignments on the fly. The reviewed prediction can optionally be submitted for review and inclusion in RepeatsDB.

4.2 Protein tandem repeats identification and annotation

Input PDB: 3vbn Input PDB Reupred log

Chains: E

Session Name: a0444a6e-104e-409b-b262-3a0719ba8f40

Click the tabs below to navigate between chains

E

Chain E Input PDB file Output Mapping Edit Annotation

COLOR LEGEND
■ Units
■ Insertions

186 Sequence viewer Search in sequence. (Regex)

```

1 MNSFYSQEEL KKIEFLSVQK NVLISKQSI YNQVISIGN NVRIDDFCIL
51 SRVTIGSYS HIAAVTALVG QVGIEMYDF ANISSRTIVV AAIADFSGNA
101 LMGPTIPNQY KQNTIKGYL KRHVIIGAHS IIFMVVIGE GVAVGAMSMY
151 KESLDDWVIY VGVPVRKIPA RRRRIVELEN EFLKSM
  
```

LEU120

Use your mouse to rotate (left-mouse) and zoom (scroll-wheel) the structure. Mouse-over to identify atoms.

Region 1 Aligned units PDB file Aligned units FASTA file Aligned units DSSP file Units PDBs Structural similarity matrix summary

Classification: III.1 Beta-solenoid

Master unit: PDB code , residues

Structure viewer (aligned units)
 3D view of aligned units, each unit is colored differently

LEU120

Use your mouse to rotate (left-mouse) and zoom (scroll-wheel) the structure. Mouse-over to identify atoms.

Sequence alignment score based on the structural alignment

14-34	1.0	0.33	0.39	0.29	0.43	0.32	0.04
35-52		1.0	0.41	0.33	0.39	0.33	0.0
53-74			1.0	0.5	0.41	0.28	0.08
75-90				1.0	0.5	0.4	0.1
118-133					1.0	0.55	0.1
134-153						1.0	0.09
154-166							1.0

Sequence viewer (aligned units)

```

ID Label      2  4  6  8  10 12 14 16 18 20 22 24 26
1 3vbnE 14-34  G E  L S V G K N V L I S K K A S I Y N P G - - -
2 3vbnE 35-52  - - V I S I G N V R I D D F C I L S G - - -
3 3vbnE 53-74  - - K V I I G S S S H I A A V T A L Y G - G E W G
4 3vbnE 75-90  - - I E N Y D F A N I S S R T I V Y - - -
5 3vbnE 118-133 - - V I L K K H V I I G A H S I I F - - -
  
```

Secondary structure viewer (DSSP, aligned units)

```

ID Label      2  4  6  8  10 12 14 16 18 20 22 24 26 28 30 32 34 36 40 42 44 46 48
1 3vbnE 118-133 - - - - - S S - - - - T T T T - - - - - T T T T - - - - - S S - - - - -
2 3vbnE 14-34  - - - - - S S - - - - - S S S S - - - - - T T T T S S - - - - - S S - - - - -
3 3vbnE 75-90  - - - - - T T T T - - - - - T T T T - - - - - T T T T - - - - - S S - - - - -
4 3vbnE 35-52  - - - - - S S S S - - - - - T T T T - - - - - T T T T - - - - - T T T T - - - - -
5 3vbnE 53-74  - - - - - S S S S - - - - - T T T T - - - - - T T T T - - - - - T T T T T - - - -
  
```

If you think that the annotation of this entry is correct, please provide a feedback to RepeatsDB. Otherwise, you can edit it by clicking the "Edit Annotation" button

Name: Email: Submit to RepeatsDB

Figure 4.18: Result page. The header provides summarizing information about the job (PDB code: 3vbn). The tabs below allow the navigation between chain predictions. Each chain tab shows some general information about the chain and a specific card for each region. Download buttons allow the retrieval of text file results, while sequence, structure and alignment viewer guide data visualization. The unit sequence similarity matrix shows the relationship between units in the region. The orange button redirects to the form for annotation editing. (177)

4. RESULTS & DISCUSSION

4.2.3.1 Web server description

Output page The different visualizations contained in the RepeatsDB-lite output page (Figure 4.18) are designed to guide the analysis of the repeat structure. The page header (Figure 4.18, top) provides general information such as the name of the input PDB (or file), processed chains and session identifier. Multiple chains are visualized in different tabs (Figure 4.18, middle). When multiple regions (groups of units) are identified in the same chain, they are visualized in the same page separated in different blocks. For each chain, regions and units are visualized in a structure and sequence viewer. For each region (Figure 4.18, bottom) the multiple structure alignment of the units and resulting sequence and secondary structure alignments are visualized along with the similarity matrix (see unit similarity paragraph) representing sequence similarity based on an all-against-all structure alignment. The PDB input, the predictor output, log and mapping files (between position along the SEQRES and PDB indices of each residue) are available for download. The manual refinement of the unit in the single chain can be accessed through the Edit annotation button in the corresponding chain tab.

Edit page RepeatsDB-lite includes a page for the manual refinement of unit annotations (Figure 4.19). The form fields (Figure 4.19, left) allow the curator to add/delete regions, change classification and modify unit annotation. On the right side of the page (Figure 4.19, right), a sequence and structure viewer react to the user edits, allowing a preliminary evaluation of the changes. Upon clicking the Submit button, the user is redirected to the results page whose content is updated according to the provided new annotation. Finally, the Submit to RepeatsDB button, available both in the edit page and in the output page, allows to submit the curated annotation to RepeatsDB for review and inclusion.

4.2.3.2 Usage example

The AntD N-acyltransferase from *Bacillus cereus* (PDB code: 3vbn) forms a ternary complex of three solenoid chains (243). Each chain folds into a left-handed β -helix of seven turns, interrupted by a loop and ending with an α -helix. The loop extends toward the flanking subunits and provides a binding platform for the ligands (Coenzyme

4.2 Protein tandem repeats identification and annotation

The screenshot shows the RepeatsDB annotation editing interface for PDB 3vbn, Chain E. The top header includes 'Input PDB: 3vbn', 'Chains: E', and 'Session Name: a0444a6e-104e-409b-b262-3a0719ba8f40'. Below this is a 'Chain E' section with a 'Back to chain visualization' button. The main area is divided into a 'Curator' section with 'ReUPred' and an 'Email' field, and a 'Sequence viewer' section showing a protein sequence with highlighted units and insertions. A 'Structure viewer' on the right displays a 3D ribbon model of the protein. At the bottom, there are 'Add Region', 'Save', and 'Submit to RepeatsDB' buttons.

Figure 4.19: Annotation editing page of the example job (PDB code: 3vbn). The user can modify region classification and the unit start end position. Changes are reflected in the viewers on the right. Reviewed annotation can be submitted directly to RepeatsDB maintainers to be included in the database. (177)

A and dTDP). RepeatsDB-lite correctly identifies five β -solenoid elements and the 27 residue long insertion between the fourth and fifth unit (Figure 4.18). It is possible to appreciate the good phasing from the multiple structure alignment. In addition, the similarity matrix shows some darker cells close to the diagonal indicating how adjacent units are more similar to each other compared than distant ones. Even if shorter, two other units are missing in the RepeatsDB-lite output. The user can add them from the edit page and immediately see the results in the sequence and structure viewer. By clicking the Save button the user is redirected to the result page and the similarity matrix is recalculated as well as the sequence and structural alignment (Figure 4.20). The latter in particular shows how the added unit diverge slightly from the perfect superimposition pattern of the previous multiple alignment including only

4. RESULTS & DISCUSSION

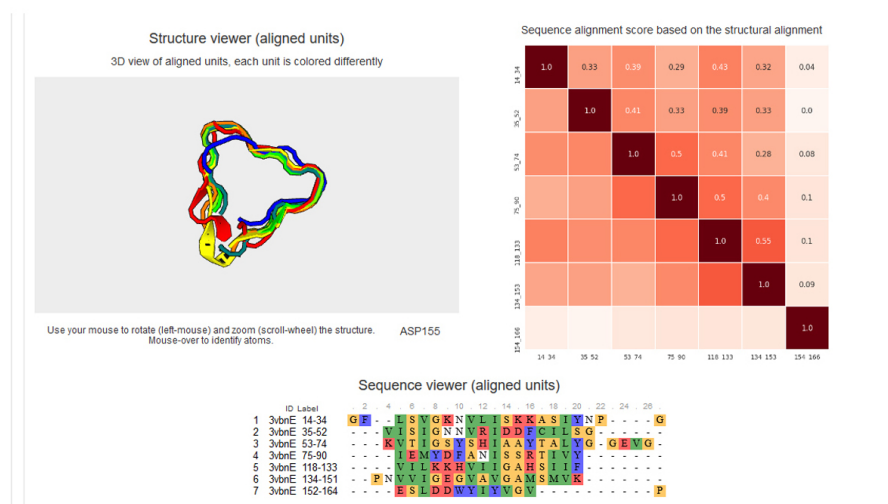


Figure 4.20: Results after resubmission of the example job (PDB code: 3vbn). By saving the repeat annotation edits, the user is redirected to RepeatsDB output page where he is provided a detailed visualization of new results to evaluate the annotation quality. (177)

the first five units (Figure 4.20, top left). The similarity matrix, where two additional elements are added (Figure 4.20, top right), shows how the last unit in particular diverges significantly from the others. The different visualizations are designed to guide the user in the annotation refinement process. Users are encouraged to send the reviewed annotation to the RepeatsDB maintainers by clicking the corresponding button.

4.2.3.3 RepeatsDB-lite performance

RepeatsDB-lite is able to predict all types of TR proteins. In Table 4.7, a comparison with other methods is provided. The benchmark is the same used previously for ReUPred (176) but with updated unit annotations, i.e. considering reviewed information from the last RepeatsDB release. The dataset includes 87 solenoid regions from 84 proteins with 679 units for a total of 19 646 repeat and 5560 non-repeat residues. The region column corresponds to the evaluated regions. RepeatsDB-lite consistently reaches a precision above 95% and outperforms the other methods both considering balanced accuracy and F-measure (Table 4.7). ConSole has a better precision for α -solenoids at the cost of missing about half of the truly repeated residues (low sensitivity). Filtered

4.2 Protein tandem repeats identification and annotation

	Method	Regions	Sn	Sp	Pr	Acc	F
III.1, β -solenoid	TAPO	31	0.546	0.802	0.851	0.674	0.665
	ConSole	31	0.510	0.811	0.848	0.661	0.637
	RepeatsDB-lite 40	31	0.398	0.959	0.952	0.678	0.561
	RepeatsDB-lite 60	31	0.543	0.962	0.967	0.752	0.695
	RepeatsDB-lite 80	31	0.560	0.912	0.929	0.736	0.699
	RepeatsDB-lite	31	0.598	0.953	0.963	0.776	0.738
III.2, α/β -solenoid	TAPO	18	0.692	0.699	0.925	0.696	0.792
	ConSole	18	0.644	0.834	0.954	0.739	0.769
	RepeatsDB-lite 40	18	0.558	0.912	0.967	0.735	0.707
	RepeatsDB-lite 60	18	0.790	0.851	0.961	0.820	0.867
	RepeatsDB-lite 80	18	0.788	0.847	0.960	0.818	0.866
	RepeatsDB-lite	18	0.838	0.864	0.971	0.851	0.900
III.2, α -solenoid	TAPO	38	0.665	0.577	0.916	0.621	0.771
	ConSole	38	0.552	0.820	0.955	0.686	0.700
	RepeatsDB-lite 40	38	0.747	0.561	0.914	0.654	0.822
	RepeatsDB-lite 60	38	0.859	0.595	0.930	0.727	0.893
	RepeatsDB-lite 80	38	0.839	0.668	0.946	0.754	0.889
	RepeatsDB-lite	38	0.885	0.684	0.951	0.784	0.917
III	TAPO	87	0.630	0.747	0.898	0.688	0.740
	ConSole	87	0.556	0.823	0.917	0.690	0.693
	RepeatsDB-lite 40	87	0.589	0.849	0.932	0.719	0.722
	RepeatsDB-lite 60	87	0.737	0.850	0.945	0.793	0.828
	RepeatsDB-lite 80	87	0.733	0.844	0.944	0.788	0.825
	RepeatsDB-lite	87	0.778	0.855	0.950	0.816	0.855

Table 4.7: Comparison with other methods. The regions column corresponds to the number of evaluated TR regions, i.e. for which a predictor provides an output, including fully negative predictions (zero units). Sensitivity (Sn), specificity (Sp), precision (Pr), balanced accuracy (Acc) and F-measure (F) values are in the range [0, 1]. Best values are in bold. (177)

versions of the RepeatsDB-lite unit library at 80, 60 and 40% sequence identity are benchmarked to assess the effects of redundancy with the test dataset. RepeatsDB-lite still shows a good accuracy even at 40% identity. In order to evaluate unit detection accuracy with a higher significance, RepeatsDB-lite was evaluated against all reviewed entries of RepeatsDB, for a total of 3666 proteins with 3835 TR regions and 29 113 units. The dataset contains 1 051 562 repeated and 193 338 non repeated residues (i.e. outside TR units). Insertion residues (29 403) are masked, i.e. not considered in the evaluation. Considering them as negatives does not affect the performance (data not shown). Results for the entire dataset and each subclass are reported in Table 4.8. RepeatsDB-lite provides prediction for 3628 proteins, 136 of which contain multiple

4. RESULTS & DISCUSSION

Classification	Description	Regions	Sn	Sp	Pr	Acc	F
II.1	Collagen triple-helix	3	0.000	0.000	0.000	0.000	0.000
II.2	α helical coiled coil	9	0.594	0.865	0.875	0.730	0.708
II	Fibrous repeats	12	0.545	0.865	0.875	0.705	0.672
III.1	β -Solenoid	325	0.561	0.926	0.911	0.743	0.694
III.2	α/β solenoid	350	0.797	0.907	0.984	0.852	0.881
III.3	α -Solenoid	888	0.784	0.628	0.943	0.706	0.856
III.4	β trefoil / β hairpins	76	0.583	0.960	0.968	0.772	0.728
III.5	Anti-parallel β layer / β hairpins	63	0.642	0.723	0.869	0.683	0.739
III	Elongated repeats	1702	0.750	0.819	0.948	0.785	0.838
IV.1	TIM-barrel	538	0.669	0.731	0.932	0.700	0.778
IV.2	β -Barrel / β hairpins	77	0.682	0.863	0.970	0.772	0.801
IV.3	β -Trefoil	24	0.449	0.754	0.731	0.602	0.556
IV.4	β -propeller	849	0.677	0.845	0.968	0.761	0.797
IV.5	α/β prism	185	0.782	0.956	0.997	0.869	0.876
IV.6	α -Barrel	18	0.419	0.931	0.917	0.675	0.576
IV.7	α/β barrel	5	0.986	0.000	0.995	0.493	0.991
IV.8	α/β propeller	117	0.591	0.836	0.933	0.713	0.723
IV.9	α/β trefoil	70	0.836	0.929	0.973	0.883	0.899
IV.10	Aligned prism	45	0.856	0.978	0.998	0.917	0.921
IV	Closed repeats	1928	0.685	0.826	0.961	0.755	0.800
V.1	α -Beads	13	0.758	0.652	0.980	0.705	0.855
V.2	β -Beads	42	0.813	0.779	0.975	0.796	0.887
V.3	α/β -beads	14	0.296	0.864	0.990	0.580	0.456
V.4	β sandwich beads	37	0.429	0.850	0.984	0.639	0.597
V.5	α/β sandwich beads	48	0.452	0.698	0.969	0.575	0.616
V	Beads on a string	154	0.537	0.759	0.975	0.648	0.692
All		3796	0.706	0.821	0.956	0.764	0.812

Table 4.8: RepeatsDB-lite performance against RepeatsDB reviewed entries. Columns headers have the same meaning of Table 4.7. (177)

regions. α/β solenoid (III.2), α/β prism (IV.5), α/β trefoil (IV.9) and aligned prism (IV.10) are the best predicted subclasses, with a balanced accuracy over 0.8. In general, the majority of the examples come from class III and IV with the former having better sensitivity and the latter better specificity. RepeatsDB-lite fails when the unit length and structure diverge too much. Class IV has a larger unit structural variability that is remarkable also inside the same region. Another source of errors are those cases for which a single unit in the reference corresponds to multiple units in the prediction (or vice versa). Even when a unit perfectly matches multiple units in the counterpart, these cases are strongly penalized because the evaluation algorithm selects at most one match for each reference unit and counts non-overlapping residues as false negatives.

4.2 Protein tandem repeats identification and annotation

Class V contains globular bead domains and the size of the dataset is much smaller than the other two. In this case RepeatsDB-lite accuracy is lower because it identifies repetitions inside domains that are generally annotated as single units by curators. Class II includes single helix fibrous structures stabilized by inter-chain interactions and lack structural repetitions. As unit annotation is completely arbitrary and unrelated to structural properties, any evaluation can be considered meaningless. RepeatsDB-lite is also able to detect the structural classification of the TR region. In particular, it correctly detects the subclass for 77% of class III proteins and 80% of class IV (data not shown).

4.2.4 Improving Repeat Definitions in Pfam

RepeatsDB project represents a powerful resource for the annotation of repeat proteins, both in terms of structure and sequence. To address TRPs issues in the sequence-based detection, we extended our collaboration with Pfam (63) curators in order to improve existing Pfam domains and create accurate models of repeats based on structural information. Different strategies are applied when building TR Pfam models. The seed alignment may include multiple consecutive repeat units instead of a single one as longer HMMs are better at discriminating true positives. The tendency of repeated sequences to diverge is especially true for flanking units, so this solution partially escapes the problem. The TR framework has been applied in cases like Leucine Rich Repeats or Ankyrins, and in newly defined families. These include the bacteriophage spike domain used to penetrate the host cell membrane, represented by an entry containing three copies of the repeat. However, this may lead to partial overlaps between detected repeats and/or missing some units, since they are present in variable number in TR regions. A different strategy is the definition of several, specific sequence models representing the same structural unit, grouped in the same clan. This has been extensively applied in the case of HEAT repeats, characterized by high sequence diversity. These include the Importin HEAT-like repeat, with six different and specific entries, and the LRR Ribonuclease Inhibitor capping repeat (located at the N- or C-terminus of the repeat region). This way, each model is found a few times along the repeat region, but the overall coverage is high. Repeat units tend to diverge also for other reasons. Variations in the typical unit sequence may be related to a specialization of function (e.g. a specific binding site) or structure (e.g. change in structural curvature). Functional unit

4. RESULTS & DISCUSSION

variations are typically conserved through evolution and their recognition requires the definition of more specific models. E.g. F-box protein Transport inhibitor response 1 (TIR1) shows a unit in the LRR region which includes the insertion of one short α -helix in the loop between the β -strand and the following helix. The unit shares sequence similarity with other units including a similar insertion in other LRR-containing proteins, supporting the hypothesis that repeat units with different structural and functional features are combined by evolution as building blocks of repeat regions similar to protein domain architectures. Indeed, while these models may have just one hit per TR region, they show a basic repeat structure perfectly compatible with the rest of the region. In these cases the model description is especially important, as it will guide the user in understanding the reasons for its specificity. Strategies for the detection of family-specific repeats are being explored along with the revision of repeat entries descriptions.

4.3 Intrinsically disordered proteins

In addition to the methods and databases for the annotation of repeat proteins, I contributed to the improvement of another database developed in BioComputingUP Lab that represents a central resource for the scientific community working in the field of NGPs: MobiDB. We published the new release (244). Several curated datasets for intrinsic disorder and folding upon binding have been integrated to MobiDB from specialized databases. MobiDB 3.0 contains information for the complete UniProt protein set and it is also linked from the UniProt entry page. A large amount of information and cross-links to more specialized databases are intended to make MobiDB the central resource for the scientific community working on protein intrinsic disorder and mobility.

4.3.1 MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.

4.3.1.1 Database description

MobiDB 3.0 is intended to be a central resource for large-scale intrinsic disorder sequence annotation. This new version is organized by both type of disorder annotation and quality of disorder evidence (Figure 4.21). Disorder information is grouped in three different sections: disorder, linear interacting peptides (LIPs) and secondary structure populations. The latter represents the conformational heterogeneity of IDPs and IDRs as the ability to populate different secondary structure populations in solution. LIPs are structure fragments that interact with other molecules preserving an elongated structure or folding upon binding. The data in MobiDB is organized hierarchically. The top tier is formed by manually curated data from external databases and represents the highest quality annotations. Annotations derived from experimental data such as X-ray and NMR chemical shifts are indirect but far more abundant. At the bottom, predictions provide disorder annotation at lower confidence than experimental evidence. The main disorder definition in MobiDB is provided by a consensus combining all available sources prioritizing curated and indirect evidences over predictions in analogy to the previous version (245). In the following, we will describe the main recent improvements since the previous release. The database schema, web interface and server have been completely redesigned and the underlying technology updated.

4. RESULTS & DISCUSSION

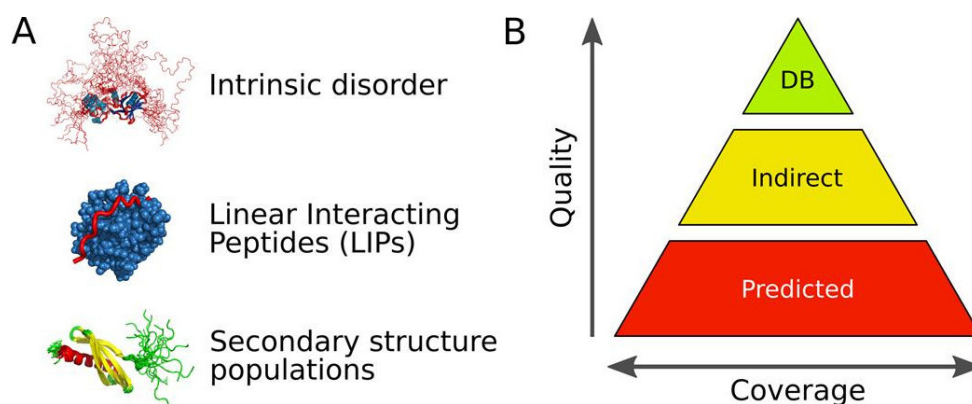


Figure 4.21: Overview of different annotation data types (A) and levels of accuracy (B) in MobiDB 3.0. (174)

The feature viewer showing sequence annotations is now fully dynamic and allows the generation of high quality images for publications with a click. Where available, MobiDB annotation is projected directly onto the structure and shown in a new 3D viewer. The look and feel and organization of the page and loading latency were also improved.

New curated data MobiDB 3.0 includes different sources of manually curated disorder annotations (Table 4.9). These annotations fall into two categories: disorder and LIPs. LIPs are binding regions presumed or demonstrated to be intrinsically disordered that fold upon binding. These come under different names such as SLiMs (short linear motifs) or MoREs (molecular recognition elements) in the literature. The IDEAL database calls them protean segments (ProS) (246). MobiDB includes both verified and possible ProS from IDEAL, where verified means disorder has been experimentally observed in the isolated molecule. The Database of Disordered Binding Sites (DIBS, (247)) collects cases where a disordered region folds upon binding with a globular domain and the Mutual Folding Induced by Binding (MFIB, (248)) database includes disordered regions that fold upon binding with another disordered region. ELM (249) provides SLiM annotations involved in binding and post-translational modifications. General disorder annotation, i.e. without any knowledge about transition driven by interactions, is collected from UniProtKB (250), DisProt (251) and FuzDB (252). UniProtKB provides manually curated disorder annotations under the region field in the features section. FuzDB collects cases of fuzzy complexes, where conformational

4.3 Intrinsically disordered proteins

diversity has a functional role in the regulation and formation of protein complexes or higher-order assemblies. DisProt has been recently revamped and MobiDB now propagates DisProt disordered regions by homology transfer. Regions homologous to experimentally characterized IDRs are mapped across homologs obtained from Gene-Tree alignments (253). Regions with identity and similarity >80% and an alignment of at least 10 residues are retained as homologous IDRs. Gene3D (254) contributes complementary order annotation to the MobiDB consensus calculation, while Pfam (102) is used to highlight protein domains. Lastly MobiDB also maps CoDNaS information to highlight conformation diversity in globular regions. CoDNaS measures structural differences among conformers of the same protein (255).

Database	Type	Comment	URL
UniProt	Curated	Disorder	http://www.uniprot.org/
DisProt	Curated	Disorder	http://www.disprot.org/
FuzDB	Curated	Disorder	http://protdyn-database.org/
ELM	Curated	LIPs	http://elm.eu.org/
MFIB	Curated	LIPs	http://mfib.enzim.ttk.mta.hu/
DIBS	Curated	LIPs	http://dibs.enzim.ttk.mta.hu/
IDEAL	Curated	LIPs	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/
Gene3D	Curated/Prediction	Structure	http://gene3d.biochem.ucl.ac.uk/
Pfam	Curated/Prediction	Domains/Families	http://pfam.xfam.org/
CoDNaS	Indirect	Conformational diversity	http://ufq.unq.edu.ar/codnas/

Table 4.9: Overview of databases integrated into MobiDB 3.0. (174)

New indirect annotations Previous releases of MobiDB provided indirect annotations from the PDB through missing residues in X-ray structures and mobile regions from NMR ensembles as calculated with the Mobi software (256). In the current release, this annotation has been complemented with additional indirect information from experimental data in the PDB and chemical shifts from the Biological Magnetic Resonance Data Bank (BMRB) (257). The new Mobi 2.0 software (244) is used to extract LIPs and disorder information from PDB files. Disorder is encoded by three different parameters: high-temperature, missing and mobile residues. High-temperature residues are detected from B-factor regions for X-ray and cryo-EM structures using a threshold proportional to the resolution of the structure. Missing residues are available for all experimental types and obtained comparing the experimental sequence (i.e. PDB SEQRES entries) with the observed residues in the structure (i.e. PDB ATOM entries).

4. RESULTS & DISCUSSION

A mobility estimate is provided for NMR structures by comparing C displacement and local conformations in different aligned models (256). LIPs are identified by comparing intra- versus inter-chain contacts calculated using RING (258). The closest atoms between two residues are used to establish a contact which is then distinguished by chemical type (e.g. hydrogen bond, salt bridge, stack). LIPs are identified as any region where the number of inter-chain contacts is at least two times the number of intra-chain contacts (244). MobiDB 3.0 better exploits the power of NMR spectroscopy to probe the structural properties of proteins in solution, as well as their dynamics on a wide range of timescales (259). Chemical shifts quantify structural fluctuations of proteins up to the millisecond timescale and are relatively easy to measure. Using chemical shifts to obtain information about the statistical populations of different structural motifs allows for a more comprehensive structural description of proteins in solution than static structures or binary definitions such as ordered and disordered (259). MobiDB 3.0 uses chemical shift data from BMRB directly as reported without applying chemical shift re-referencing methods. The software packages δ 2D (260) and Random Coil Index (RCI) (261) are used to calculate two-dimensional ensembles in terms of secondary structure populations (259) and backbone flexibility. Secondary structure populations are calculated only for residues with at least three atom types with measured chemical shifts, as using fewer chemical shifts results in less accurate mappings of the populations (260). MobiDB 3.0 reports the experimental conditions at which the chemical shifts were measured as the structural properties of some proteins can change drastically between different conditions (e.g. binding partners, lipids, pH) and these can help elucidate protein function (259). When an entry in MobiDB is associated to multiple chemical shifts, an overview of the predominant secondary structure conformation is provided in a consensus track. This can be expanded in the feature viewer to show experimental conditions such as pH, temperature, binding partners, molecular state, sample information and the title of the corresponding BMRB entries.

New predictors MobiDB 3.0 includes the same set of disorder predictors used in the previous release: ESpritz (183), IUpred (262), DisEMBL (130) and VSL2b (263). Consensus generation is handled by MobiDB-lite (51), which uses a stronger majority threshold and enforces at least 20 consecutive disordered residues to provide highly specific predictions. This is completed by a continuous representation of the fraction of

4.3 Intrinsically disordered proteins

methods predicting disorder for each residue. DynaMine (264), Anchor (265) and FeSS (185) are now also part of the annotation pipeline. DynaMine (264) predicts backbone flexibility where 1.0 means complete order (stable conformation, i.e. rigid) and 0 means fully random bond vector movement (highly dynamic, i.e. flexible). Anchor predicts binding regions located in disordered proteins, providing LIP annotations for all proteins in the database. FeSS is a component of the FIELDS method (185) providing three-state (helix, sheet, coil) secondary structure propensity. FeSS prediction confidence can be interpreted similarly to the dynamic behavior measured by 2D in chemical shifts, i.e. a propensity to remain in a given state of secondary structure. The complete list of tools is available in Table 4.10.

Tool	Type	Description
Mobi 2.0	Indirect	Missing, high-temperature and mobile residues from PDB structures
RING 2.0	Indirect	Residue interactions from PDB structures, used to define LIPs
RCI	Indirect	Random coil index from BMRB chemical shifts
2D	Indirect	Secondary structure populations from BMRB chemical shifts
DynaMine	Prediction	Random coil index
FeSS	Prediction	Secondary structure prediction component of FIELDS
MobiDB-lite	Prediction	Long disorder based on consensus
DisEMBL	Prediction	Disorder. Versions: 465, Hot-loops
ESpritz	Prediction	Disorder. Versions: DisProt, NMR, X-ray
IUPred	Prediction	Disorder. Versions: Short, Long
VSL2b	Prediction	Disorder
GlobPlot	Prediction	Globular regions, used as opposite of disorder
SEG	Prediction	Low complexity
Pfilt	Prediction	Low complexity

Table 4.10: Overview of tools used into MobiDB 3.0. (174)

4.3.1.2 Usage and annotated data

MobiDB now contains all sequences from UniParc, the most comprehensive non-redundant set of protein sequences. Entries are identified also by UniProtKB (250) accession numbers and can be retrieved by organism, taxonomy and other identifiers provided by UniProtKB. Prediction results are combined with indirect disorder evidences derived from PDB data (using Mobi 2) and data extracted from manually curated third party databases. MobiDB annotations are used by DisProt (251) curators to guide the annotation of disorder regions. MobiDB data is made available to the public via a web

4. RESULTS & DISCUSSION

interface allowing extensive search functionalities and RESTful services for programmatic access. MobiDB 3.0 includes a pre-calculated consensus for all entries allowing real-time statistics and download of entire datasets in different formats directly from the web interface. The new database schema makes it possible to perform complex search queries and to generate custom datasets, for example retrieving all entries with manually curated annotations. The MobiDB update has been automatized and is scheduled every three months due to the high computational cost of generating predictions for new sequences.

4.4 Low complexity sequences

A common feature of TRPs, IDPs and other NGPs is that they are characterized not only by a non-canonical structure, but also by a non-canonical sequence which hampers their detection and analysis. In particular, several NGPs are characterized by low complexity (LC) sequences. A low complexity (LC) sequence often shows malignant aggregation propensity, therefore flanking identical sequences are prone to diverge, as already discussed in section 4.1.1. This originated a continuum between repeated structures, disordered proteins and aggregation-prone domains which is not easy to explore from the structural point of view and even more complicated as regards the relationship between the sequences of these proteins. This chapter presents a critical review focusing on the definition of sequence features of LC regions and their connection with structure. We presented statistics and methodological approaches that measure low complexity and related sequence properties. Composition bias is often associated with low complexity and disorder, but repeats, while compositionally biased, might also induce ordered structures. We illustrated this dichotomy, and more generally the overlaps between different properties related to LC regions, using examples. We argued that statistical measures alone cannot capture all structural aspects of LC regions and recommend the combined usage of a variety of predictive tools and measurements. While the methodologies available to study LC regions are already very advanced, we foresee that a more comprehensive annotation of sequences in the databases will enable the improvement of predictions and a better understanding of the evolution and the connection between structure and function of LC regions.

4.4.1 Disentangling the complexity of low complexity proteins

This section presents statistics and methodological approaches that measure low complexity and related sequence properties. Composition bias is often associated to low complexity and disorder, but repeats, while compositionally biased, might induce ordered structures. This study illustrates this dichotomy, and more generally the overlaps between different properties related to LCRs, using examples.

4. RESULTS & DISCUSSION

AC	ID	Description	Length (aa)	Organism
Q38PT6	Q38PT6.9HEXA	6.5 kDa glycine-rich antifreeze protein	103	Hypogastrura harveyi
P35226	BMI1_HUMAN	Polycomb complex protein BMI-1	326	Homo sapiens
P20226	TBP_HUMAN	TATA-box-binding protein	339	Homo sapiens
P04637	P53_HUMAN	Cellular tumor antigen p53	393	Homo sapiens
P32583	SRP40_YEAST	Suppressor protein SRP40	406	Saccharomyces cerevisiae
P34945	SYS.THET2	Serine-tRNA ligase	421	Thermus thermophilus
P0C2W0	YADA2_YEREN	Adhesin YadA	422	Yersinia enterocolitica
P02930	TOLC_ECOLI	Outer membrane protein TolC	493	Escherichia coli (s. K12)
P35637	FUS_HUMAN	RNA-binding protein	526	Homo sapiens
P49711	CTCF_HUMAN	Transcriptional repressor CTCF	727	Homo sapiens
P15502	ELN_HUMAN	Elastin	786	Homo sapiens
P42566	EPS15_HUMAN	Epidermal growth factor receptor substrate 15	896	Homo sapiens
Q9BVN2	RUSC1_HUMAN	RUN and SH3 domain-containing protein 1	902	Homo sapiens
P10275	ANDR_HUMAN	Androgen receptor	920	Homo sapiens
Q8WVM7	STAG1_HUMAN	Cohesin subunit SA-1	1258	Homo sapiens
Q9NZW4	DSPP_HUMAN	Dentin sialophosphoprotein	1301	Homo sapiens
Q8ZL64	SADA_SALTY	Autotransporter adhesin Sada	1461	Salmonella typhimurium
P02452	CO1A1_HUMAN	Collagen alpha-1(I) chain	1464	Homo sapiens
A3M3H0	ATA.ACIBT	Adhesin Ata autotransporter	1873	Acinetobacter baumannii
P24928	RPB1_HUMAN	DNA-directed RNA polymerase II subunit RPB1	1970	Homo sapiens
P42858	HD_HUMAN	Huntingtin	3142	Homo sapiens

Table 4.11: Illustrative set of proteins with LCRs, ordered by the length of the protein. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

4.4.1.1 The many shades of complexity

To illustrate the overlap between amino acid composition, periodicity and structure we use a 2D diagram where we can compare proteins (or regions) of various degrees of complexity from intermediate to unbiased ("normal") sequences according to their compositional bias and repetitiveness (Figure 4.22). This diagram applies ideally to sequence regions with lengths in the range of 10 to 50 residues, for the sake of simplicity (considering that long structural repeats have a length of about 50 residues (79) and fragments of less than 10 residues would suffer from low-count statistical effects). Suppose that we compute for one such region two simplified measurements of complexity: one reflecting variability of amino acid usage (compositional bias) and the other indicating periodicity. For example, AEEAEAAEEA and a perfect direpeat like AEAEAEAEAE have the same amino acid composition (50% A, 50% E) but different periodicities. As a simplified measurement of amino acid variation, we can take the percentage of the most frequent amino acid in the region (see (120) for another measure of repeat perfection). For example, given the ten-amino acid sequence ACDEFEGEIE, the most abundant amino acid is E, at 40%. To measure repetitiveness, we could calculate how distant this sequence is from a sequence with perfect repeats. A

simple measure for that distance is how many residues we need to mutate to convert the query sequence to a perfect repeat. The simplest instance of a repeat is the homorepeat; any sequence with $n\%$ for the most frequent amino acid can be converted to a homorepeat by changing the other residues to the most frequent residue, i.e. $100\% - n\%$. For our example sequence, ACDEFEGEF, we would have to change six residues to E, 60% , to have 10 E residues. This sets the upper limit to this value. But if a less trivial repeat can be found using fewer mutations, this second value will be necessarily lower. In this case, we can change ACDEFEGEF to FEFEFEF with only 40% of changes. Using these metrics, we can conceptually position in the diagram (Figure 4.22) examples of regions of variable degrees of complexity (y-axis) and repetition (x-axis). All perfect repeats are placed at $x = 0$, and homorepeats have $y = 100\%$. Direpeats have $y = 50\%$, AABAAB repeats have $y = 66\%$, ABCABC repeats have $y = 33\%$, and so forth. Proteins without repeats are placed in the trivial diagonal, with a y value for the most frequent amino acid and $x = 100\% - y$. A protein composition with all 20 amino acids equally abundant sets 5% as the lower limit for y . Rather, most proteins will have unbiased compositions where the most abundant amino acid forms around 10% of the sequence (e.g. aspartate 10.7% or glutamate 9.9% in (266)). Then, unbiased proteins, far from repeats and with the expected amino acid variation, will populate the bottom-right corner of the diagram. We can imagine intermediate situations, which can be constructed by adding mutations from regions with perfect repeats. In this manuscript, we will discuss the hypothesis that there is a border between LCRs influenced by periodicity (i.e., repetitiveness), so that given two LCRs with the same amino acid composition, the one with more repetitiveness might be prone to form a structure, whereas the other one would have a stronger tendency to be disordered. This would give a slant to the low complexity border (line separating the "Low complexity" area, Figure 4.22). Not all repeats are LCRs, but LCRs tend to be close to short repeat sequences, since groups of short repeats have necessarily a limited number of amino acids, and thus can be considered a low complexity unit. In other words, low complexity can only be compositionally biased, while compositional bias can be of low or high complexity. In order to explore how the different measurements of complexity and repetition relate to this graphical representation in reality, we will take a few proteins with LCRs, repeats of various types, and a range of structures, measure their complexity using available methods and locate these regions in the model graph. Note

4. RESULTS & DISCUSSION

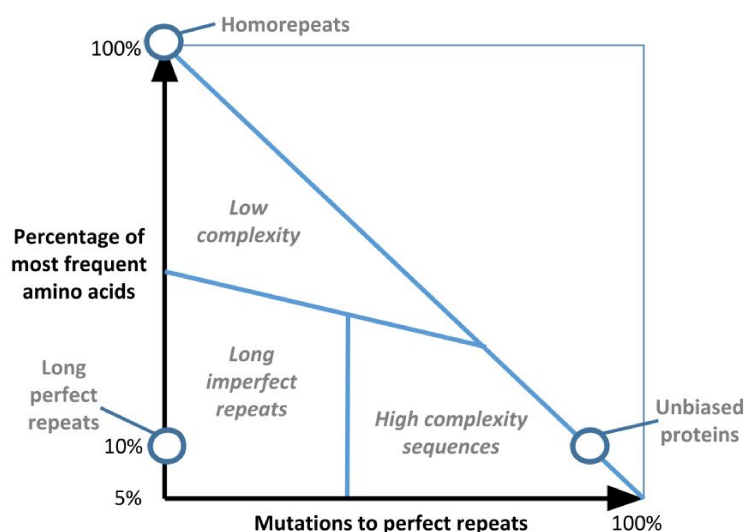


Figure 4.22: The diagram illustrates where several types of sequences would be placed in relation to two measures related to sequence complexity. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

that parts of these proteins will have an expected composition, which will populate the point of unbiased proteins, where most globular proteins reside. This will constitute a contrast to which we can compare their LCRs.

4.4.1.2 Detection of low complexity sequences

We collected a set of 21 protein sequences to illustrate the phenomena involved in LCRs (Table 4.11). This dataset is a collection of examples of what is commonly defined as a compositionally biased protein. It includes enzymes (Serine tRNA ligase, P34945 UniProt accession number), transcription factors (Transcriptional repressor CTCF, P49711), membrane channels (Outer membrane protein TolC, P02930), transporters (Autotransporter adhesin SadA, Q8ZL64), structural proteins (Collagen alpha-1 chain, P02452), proteins that respond to changes of physical states (Glycine-rich antifreeze protein, Q38PT6), typical disordered proteins (Cellular tumor antigen p53, P04637) and proteins related to diseases (Huntingtin, P42858). With this selection, we aim at relating the concept of compositional bias in proteins to a variety of cellular processes, compartments, and structural states. We note that associating function to LCRs is not our goal here; rather, the functional variety in the set of proteins chosen to highlight

the diversity of biological situations where low complexity plays a relevant role. In the following sections, a series of methods that are widely used to detect low complexity in protein sequences are introduced and applied to the dataset of the selected 21 proteins. The methods are presented in chronological order, to facilitate the understanding of the historical context within which each method was developed. In each section, we discuss the features and possible functions of detected LCRs, to illustrate the current knowledge on those regions and directions to obtain further insights about them. Related structural aspects and methods that take them into account are discussed after this part.

SEG (1993): Detection of LCRs SEG was the first algorithm developed to specifically detect LCRs within protein sequences (267), as masking of LCRs has been found to improve the detection of homology (268). This method is based on the concept of local complexity of a subsequence defined for a window of length L . Such subsequences can be represented in the form of a state complexity vector, where each position represents the number of amino acid occurrences in that window. For any state complexity vector, its compositional complexity and probability of occurrence of the particular complexity state can be computed. Based on these values any subsequence can be classified as a low or high complexity subsequence. Here, we applied SEG to the collected set of proteins (Table 4.11) to characterize their LCRs and putative function based on their sequence homology with other non-related proteins. As proposed in (269), we used the SEG algorithm with intermediary parameters (these are window length $W=15$, trigger complexity $k_1=1.9$, extension complexity $k_2=2.5$). We found that twelve proteins from the dataset contain a total of 46 LCRs, with the longest having 760 residues (dentin sialophosphoprotein). Moreover, both elastin and Collagen alpha-1(I) chain have eleven LCRs each. On average, the twelve LCR-containing proteins have 3.8 LCRs with an average length of 67 residues. Similarity between LCRs in different proteins can be used to propose hypotheses about the function of the similar proteins. However, many caveats apply i.e. in the case of low complexity sequences, matching hits do not guarantee evolutionary relationship even with statistically significant scores. We illustrate this with one of our example proteins: dentin sialophosphoprotein, which contains the longest LCR of all the examples. We used the NCBI BLAST search engine with default options to find other proteins with similar LCRs. Dentin sialophosphoprotein (DSPP; UniProt:Q9NZW4) is cleaved into two chains: dentin phosphophoryn

4. RESULTS & DISCUSSION

(DPP; amino acids 16-462) and dentin sialoprotein (DSP; amino acids 463-1301). A very long LCR was detected in DSP covering most of the sequence (amino acids 511-1270). DSP is an extracellular matrix protein synthesized by odontoblasts. It is highly acidic, and the phosphorylated protein possesses a strong affinity for calcium ions. Therefore, DSP in the extracellular matrix can promote hydroxyapatite nucleation and can regulate the size of the growing crystal (270, 271, 272). Apart from its calcium binding property, DSP can initiate signaling functions from the extracellular matrix (273, 274, 275, 276). We found a high degree of similarity of the DSP fragment of DSPP to two hypothetical proteins, BCR41DRAFT_427036 (NCBI Reference Sequence AC: XP_021875136.1) from *Lobosporangium transversale* (a fungus) and JF76_17750 (GenBank AC: KJY54264) from *Lactobacillus kullabergensis* (a bacterium). Both are highly acidic sequences, rich in serine and aspartic acid. The bacterial protein possesses three MucBP domains, which are characteristic for peptidoglycan binding proteins; the presence of these domains suggests a function outside of the cell, probably in adhesion.

CBR type	Nr.CBRs	Nr.CBRP	%CBRPs	Nr.CBRPs in UniProt	%CBRPs in UniProt
A	4	4	19	19465	19.5
D	1	1	4.8	5293	5.3
E	8	7	33.3	25438	25.5
G	7	5	23.8	8771	8.8
K	2	1	4.8	14936	15
N	2	2	9.5	5428	5.4
P	9	8	38.1	12000	12
Q	5	5	23.8	9149	9.2
S	14	13	61.9	25081	25.1
T	2	2	9.5	4216	4.2
R	0	0	0	3768	3.8
C	0	0	0	1083	1.1
H	0	0	0	2584	2.6
I	0	0	0	2178	2.2
L	0	0	0	2422	2.4
M	0	0	0	766	0.8
F	0	0	0	756	0.8
W	0	0	0	274	0.3
Y	0	0	0	562	0.6
V	0	0	0	1487	1.5

Table 4.12: Compositionally biased regions (CBR) and CBR containing Proteins (CBRPs) detected by CAST. A single protein sequence may contain one or more CBRs of the same or even different residue types. The last two columns refer to UniProt/Swiss-Prot entries (release 2014_05) as retrieved from LCR-eXXXplorer. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

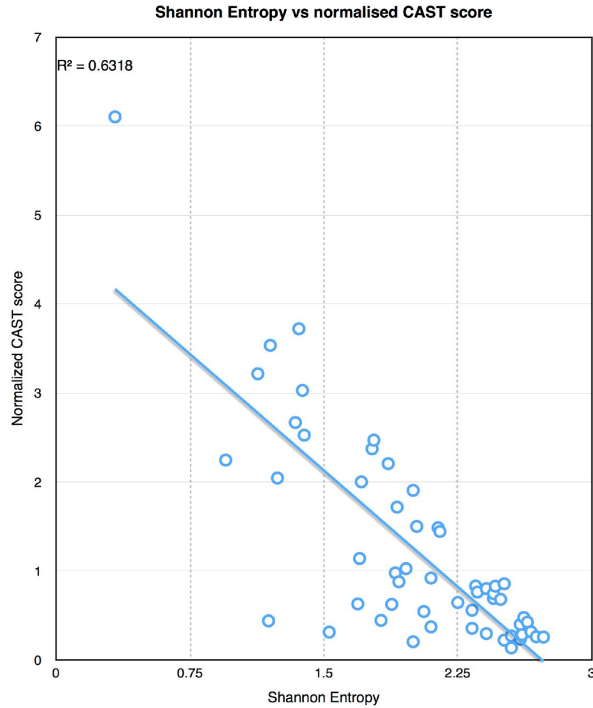


Figure 4.23: Shannon entropy value for each detected CBR against the CAST score normalized by the sequence length. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

CAST (2000): Detection of compositionally biased regions A next logical step following the detection of LCRs with SEG is to focus on compositionally biased regions (CBRs). While the usage of the terms LCR and CBR has been interchangeable in many contexts (Table 1.1), as they overlap significantly, the use of one term or the other depends on the focus of the method used for their detection, i.e., sequence variability or amino acid composition, respectively. Indeed, the terms LCR and CBR are somehow imprinted by the fields of computer science and biology, respectively. CAST was developed based on the idea that CBRs are enriched in at least one amino acid type (277). In brief, CAST detects (and scores) CBRs using comparisons of a query sequence against a database of 20 degenerate homopolymeric sequences based on each of the 20 amino acid types. Overlapping CBRs of different type (residue) may be detected in the same sequence tract. Here, we applied the CAST algorithm to our dataset with default parameters (BLOSUM62 substitution matrix and a detection threshold value

4. RESULTS & DISCUSSION

of 40). All 21 proteins from the dataset were detected to contain at least one CBR, with 54 CBRs in total (mean: 2.6 CBRs/sequence, median: 2, standard deviation: 1.5) (Table 4.12). The number of CBRs per protein vary between 1 (n=7 proteins) and 5 (n=3 proteins). CBRs vary considerably in length, with the shortest one being just 10 residues long (a P-rich region in the androgen receptor) and the longest being a S-rich region extending over 1436 residues covering almost the entirety of the autotransporter adhesin SadA. It is worth mentioning that in our dataset CAST did not detect half of the possible CBR types, namely CBRs enriched in R, C, H, I, L, M, F, W, Y and V residues. Some of these CBR types are indeed rare in the overall sequence database (Table 4.12). Our analysis stresses the fact that composition bias is related to low complexity (as discussed in the complexity diagram) but is more widely spread and commonly found in many proteins. Along these lines, of the 54 CBRs detected in this dataset using CAST, only 12 instances correspond to sequences with high sequence complexity values ($k_2 > 2.5$), illustrating that the majority of CBRs in this dataset are also LCRs. Interestingly, these 12 CBRs with high complexity values correspond to relatively long regions (often spanning along hundreds of residues) and, nevertheless, dominated by serine-rich tracts (9 out of 12). Importantly, CAST offers the possibility to explore another dimension of LCRs, which is the residue type characterizing each region. In addition, when plotting the CAST score normalized by the sequence length for each detected CBR against the Shannon entropy (Figure 4.23), we observe a correlation sorted in a triangle with many points crowding the bottom-right corner (high entropy and low normalized CAST score), which is reminiscent of the low complexity diagram (Figure 4.22).

SIMPLE (2002): Detection of tandem and cryptic repeats The tool SIMPLE was first developed in 1986 to quantify the amount of simple sequences in DNA (278). A version for proteins was developed in 2002 (279). The original aim of SIMPLE was to identify genomic sequences with a propensity to undergo replication slippage and to quantify the concept of cryptic simplicity, which corresponds to one or more short sequence motifs within a sequence region, above a baseline, random concentration. The 2002 implementation extends this original concept to detect comparably cryptic sequences at the amino acid sequence level. To provide a rich overview of the repeat landscape of the 21 proteins in our dataset, we analyzed them using an updated

4.4 Low complexity sequences

ID	Nr.repeats identified	Characteristic repeat(s) frequency
Q38PT6.9HEXA	23	G (19)
TBP_HUMAN	336	Q (41)
P53_HUMAN	11	AP (6)
SRP40_YEAST	794	S (168)
FUS_HUMAN	175	G (60)
CTCF_HUMAN	1	EP (1)
ELN_HUMAN	350	A (30), GV (28)
EPS15_HUMAN	11	DPF (6)
RUSC1_HUMAN	6	PP (3)
ANDR_HUMAN	351	Q (25), G (23)
DSPP_HUMAN	3082	S (459)
SADA_SALTY	3	NTT (2)
CO1A1_HUMAN	113	GP (17)
ATA_ACIBT	21	NTK, TKTEL (3)
RPB1_HUMAN	948	SP (96)
HD_HUMAN	211	P (27)

Table 4.13: Numbers and major classes of repeats identified by SIMPLE analysis. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

version of the SIMPLE tool (280). Significant repeat motifs of length 1 to 10 were identified at a per-analysis probability cutoff of 0.99 (aggregate cutoff probability 0.9) by awarding a score of 1 for the selected length and 0 for all other lengths. Analyses were carried out using an 11-residue moving window. Sixteen of the sequences analyzed using the SIMPLE method contained significant repeat motifs to some degree (Table 4.13). SIMPLE analysis provides two types of motif information: motif identity and motif hit frequency information defined as the frequency with which a given motif is detected as being significantly repeated within a given sequence. As examples, three of the proteins in the test set (huntingtin, TATA-binding protein and androgen receptor) contained significantly repeated motifs of all possible Q_n motifs (from n=1 to n=10), characteristic of a simple polyQ repeat. However, the most prominently repetitive protein in the set was Dentin, which, as described before, contained numerous highly repeated motifs with serine as the primary repeated amino acid. Examining the list of motifs detected in the most repetitive proteins in the dataset reveals many similar or closely related motifs. To portray these relationships, the motifs can be represented graphically. As an example, Figure 4.24 shows a motif graph for Collagen alpha-1(I) chain. The representation links different motifs identified in the sequence with their sequence overlap. The example in Figure 4.24 shows a closely-knit set of motifs linked to the submotifs PGP and GPP alongside others linked to PGA. Some motifs in this

4. RESULTS & DISCUSSION

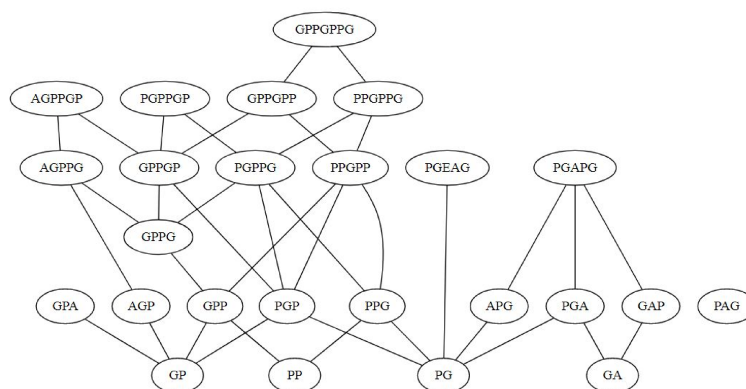


Figure 4.24: Motif graph based on SIMPLE analysis of CO1A1_HUMAN. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

example (and in other sets) are less connected than others - the extreme example in P02452 being PAG which, although related to others by circular permutation, does not overlap with them.

A		% residues predicted by				
		IUPred	SEG	CAST	SIMPLE	
% residues predicted by		Total	44.89	15.04	50.16	18.51
		IUPred	100	27.07	78.66	32.03
		SEG	80.78	100	98.41	90.32
		CAST	70.4	29.51	100	35.27
		SIMPLE	77.69	73.42	95.89	100
B		Enrichment of overlap				
		IUPred	SEG	CAST	SIMPLE	
Enrichment of overlap		IUPred	1	1.8	1.57	1.73
		SEG	1.8	1	1.96	4.88
		CAST	1.57	1.96	1	1.91
		SIMPLE	1.73	4.88	1.91	1

Table 4.14: (A) Fraction of residues predicted by one method (columns) that are predicted by another method (rows). (B) Enrichment ratio of overlapping residues between two methods compared to random overlap. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

4.4 Low complexity sequences

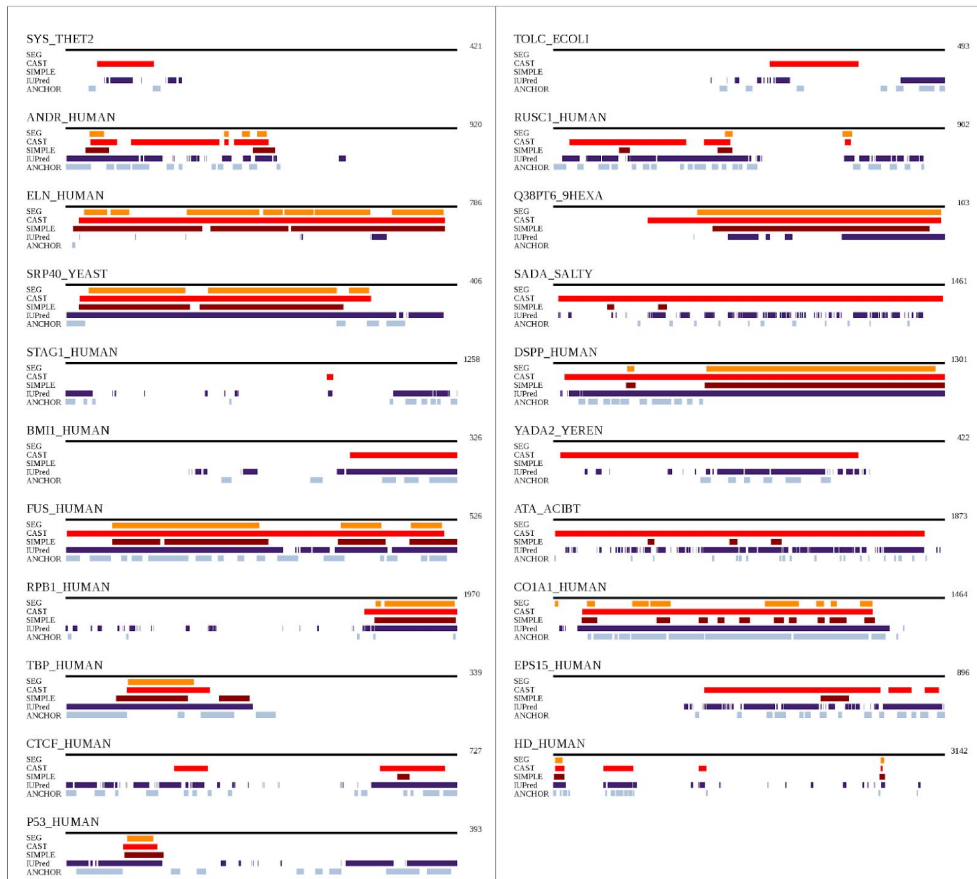


Figure 4.25: Comparison of positions detected to be of low complexity in the 21 proteins of our dataset. Methods SEG (in orange), CAST (in red), SIMPLE (in brown) and IUPred (in purple) were used. ANCHOR (in light blue), which includes structural aspects, is also compared. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

4.4.1.3 Correlation between low complexity and disorder

Low complexity and compositionally biased sequences often overlap with protein disorder (281). However, their precise relation largely depends on the applied methods used for their quantification. Here, the IUPred method was used to characterize protein disorder and to calculate the overlap with the various features determined with the methods SEG, CAST and SIMPLE described earlier. IUPred captures the basic biophysical properties of ordered and disordered sequences by relying on an energy estimation scheme. According to this, sequences composed of amino acids that cannot form enough favorable intrachain interactions would be disordered and can be recog-

4. RESULTS & DISCUSSION

nized from the amino acid sequence by their less favorable estimated energies (282). All the 21 sequences in our dataset contained at least one disordered segment, and nearly 45% of residues were predicted as disordered. This was lower compared to the average residues predicted by CAST, but higher than those predicted by SEG (15%). Table 4.14 and Figure 4.25 describe the overlap between the various methods. The matrix of overlaps is non-symmetrical (Table 4.14A), as the overlap is computed on the percentage of residues with a given feature. For example, 81% of SEG low complexity residues are predicted to be in disordered regions by IUPred. However, only 27% of residues predicted to be disordered by IUPred are found in a SEG detected region. Overall, there is a fairly good agreement between the methods that detect low complexity and the disordered regions detected by IUPred. Between the methods that detect low complexity, the largest agreement (relative to random overlap) was observed in the case of SEG and SIMPLE, likely because both produce relatively conservative predictions (Table 4.14B). Interestingly, by this metrics, the overlap between IUPred and the low complexity methods was not much lower as the overlap between CAST and the other methods.

The low complexity diagram: a proof of principle The low complexity (LC) diagram described before (Figure 4.22) allows us to situate and compare protein sequences in a framework that reflects two simple properties that are intimately associated to low complexity: compositional bias and repeats. These two features are measured by computing the abundance of the most frequent amino acid in the tract, and by the fraction of residues that needs to be mutated to have a perfectly repeated tract, respectively. We calculated the properties that define the two axes of the LC-diagram for a dataset of globular monomeric proteins (globular), a dataset of disordered proteins (IUP) (283), and for fragments of our own protein dataset (Table 4.11) determined to be of low complexity by the SEG, CAST, and SIMPLE methods (with a minimum length of 10 residues) (Figure 4.26). To place them in the LC-diagram, the percentage of the most common amino acid in each sequence was determined as a function of the percentage of the mutations to form perfect repeats. The latter quantity was calculated in a brute force way by considering all potential fragments of the sequence of lengths between 1 and 30. From these fragments, an artificial sequence of perfect repeats was generated by iterating these elements to be long enough to cover the original sequence region. At least three repeats were required, therefore only fragments no longer than a third of

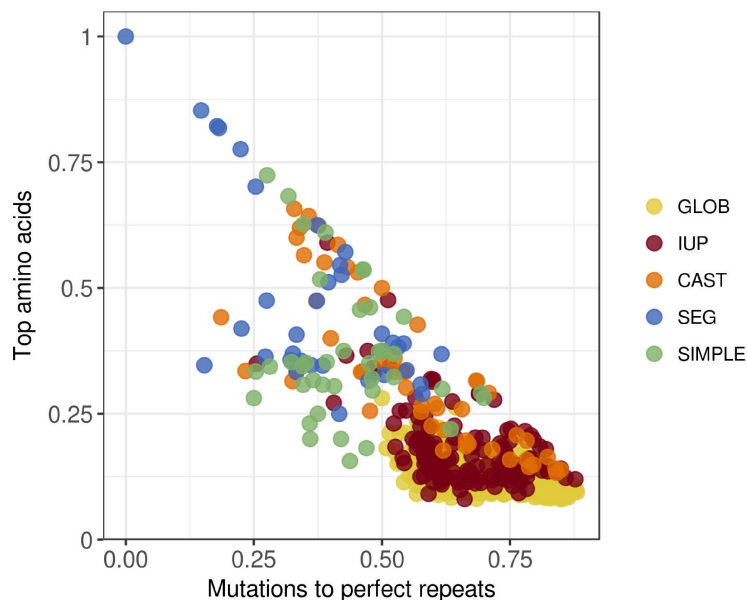


Figure 4.26: The percentage of the top amino acid as a function of the percentage of mutations to perfect repeats calculated for a dataset of globular (GLOB), disordered (IUP) sequences as well as fragments of our protein dataset with low complexity character according to the SEG, CAST and SIMPLE methods. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

the sequence were considered. The minimum number of mutations between the original and these artificial sequences was calculated and normalized by the sequence length. This approach cannot consider insertions and deletions. Thus, the x-values calculated represent an estimate, and the real values (if different) can only be closer to zero. The regions from globular proteins are distributed as a compact cloud (yellow points) that edges on the point described as globular in Figure 4.22 (bottom-right corner; Figure 4.26). An inferior limit around 10% of top amino acid agrees with the estimation published in 1966 (266). The globular cloud overlaps with the disorder cloud (red points) outside the immediate vicinity of regular proteins and extends into the realm of low complexity (orange, blue and green points). The separation between the globular cloud and the low complexity cloud described by SEG is very strong: the clouds touch each other but they do not overlap. Disordered regions overlap with both globular proteins and low complexity regions, as expected. The disorder cloud overlaps with the globular cloud but does not touch the extreme, indicating that a globular sequence can transition to disorder both by gaining a biased sequence but also via slight repetitions.

4. RESULTS & DISCUSSION

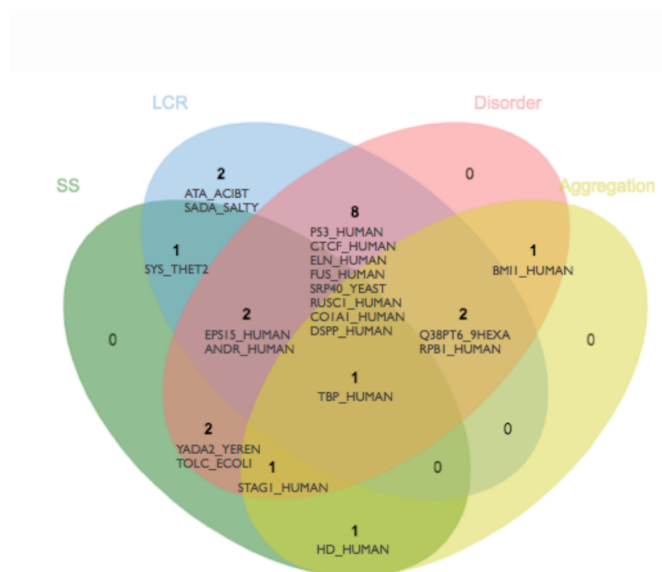


Figure 4.27: Venn diagram representing the FELLs prediction of dataset proteins, in four categories: Secondary Structure (SS), Low Complexity Regions (LCR), Disorder and Aggregation. Each protein is assigned to a category if more than 30% of the residues in its sequence are predicted in that state. License: Attribution-NonCommercial-NoDerivatives 4.0 International. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

In this respect, however, it is interesting to note that the disorder cloud overlaps very little with the repeat cloud, confirming that long perfect repeats are predicted to confer order. This is a structural aspect that we address in the next section.

4.4.1.4 Structural properties of LCRs

The experimental determination of protein structure is much more challenging for LCRs than for globular and fibrous proteins (284), and only few cases have been studied experimentally. This is due to various reasons that we will explain in this section. To guide our tour from the sequence to the structural aspects of LCRs, we will continue our strategy to illustrate low complexity with the set of 21 examples, taking into consideration the previously obtained information for these sequences. There are prediction tools specialized for the study of the structural properties of proteins, which we will apply to the selected proteins with LCR. It should be noted that for many of them there is experimentally known 3D structure covering parts of the sequence, but these generally

do not overlap with LCRs. For example, the recently solved structure of Huntingtin (285) does not resolve the N-terminal 90 amino acids, which contains a CBR including the polyQ whose expansion causes Huntingtons disease, and the 2622-2660 fragment, both of which practically overlap to the regions identified as LCRs in our SEG analysis.

Analysis of the structural properties of low complexity sequences The structural properties of LCRs can be predicted with several bioinformatics methods. To classify the incidence of different phenomena in the dataset, we used FIELDS, a predictor that aggregates sequence and structural propensity predictions in a single view (286); this includes secondary structure, LCR, disorder and aggregation predictions displayed along sequence positions. We focused on four predictions: LCRs (SEG), disorder (ESpritz-NMR), aggregation propensity (Pasta 2.0 (287)) and secondary structure (FESS). We classified each protein in the dataset as belonging to one category (low complexity, disordered, aggregating, structured) if more than the 30% of its sequence is predicted to be in that state. The results are shown in a Venn diagram (Figure 4.27). In our dataset, focused on LCRs, only one protein falls outside the LCR and/or disorder categories. This is Huntingtin, the longest of the 21 proteins (3142 amino acids) known to harbor homorepeats, alpha-solenoid repeats and globular domains (285, 288). In agreement with the sequence analyses presented before, we observe a large overlap between LCR and disorder (13 of 21 proteins), including proteins such as the Glycine-rich antifreeze protein (Q38PT6_9HEXA), Dentin (DSPP_HUMAN), and human RNA binding protein FUS (FUS_HUMAN). Regarding aggregation, while three of the six proteins classified as aggregating are also in the LCR category (TBP_HUMAN, RPB1_HUMAN and Q38PT6_9HEXA), we need to look at the sequence level (Figure 4.28). For example, for both TBP_HUMAN and RPB1_HUMAN the regions with aggregation propensity (minima in the aggregation score plot) do not overlap with the LCRs (Figure 4.28). Even in FUS, a largely disordered protein with generally low sequence complexity, its few regions presenting aggregation propensity are localized in the small ordered part of the protein (Figure 4.28). A possible explanation of this is that LCRs and aggregation prone regions have different amino acid frequencies. Hydrophobic residues inducing aggregation are probably less abundant in LCRs. This was the case in our dataset (see Table 4.12 for CBRs). Therefore, our small dataset supports the previous association between LCR and disorder but not to aggregation propensity. However, TBP leads

4. RESULTS & DISCUSSION

to another turn in our story, by bringing another player relating LCR, structure and aggregation: homorepeats. TBPs LCR is a large stretch of consecutive glutamines (positions 55-95), which is interestingly predicted both in helical conformation and as a disordered region. These contradictory predictions are most probably due to the lack of detailed understanding of the conformational preferences adopted by homorepeats. In the next section, we discuss the challenges posed by homorepeat structure prediction and determination, and the strategies that have been proposed for their study.

Deciphering the structural basis of homorepeat function Homorepeats are an extreme case of low complexity and in this respect, they can help us to illustrate the origin of the difficulties in relating sequence and structure in LCRs. In homorepeats, the presence of multiple copies of a single amino acid in a protein region confers very specific physicochemical properties to the hosting protein and enables it to perform specialized biological tasks (289). Despite their relevance, the connection between amino acid sequence, 3D structure and biological function in homorepeats remains poorly understood due to the challenges they pose to structural biology. Homorepeats and short repeats are found in disordered regions, a property that typically precludes their crystallization. In the case of polyQ, there are, however, examples that have been crystallized in the presence of fusion proteins (290, 291) or specific antibodies (292, 293). These studies yield contradictory results regarding the secondary structural preferences of polyQ tracts. This observed structural variability most likely originates from the inherent conformational plasticity of the homorepeat regions, which cannot be captured in crystallographic studies. Nuclear Magnetic Resonance (NMR), a high resolution structural technique in solution, seems more adapted to study homorepeats. However, the similarity of the nuclear resonance frequencies within homorepeats have hampered these studies. Some pioneering NMR studies of polyQ homorepeats in Huntingtin (294, 295), and the androgen receptor (296) have shown these studies are possible. These examples show that the N-terminal flanking region of the polyQ adopts an α -helical conformation that extends towards the homorepeat. In the absence of this structured flanking region, polyQ adopts a random coil conformation (296, 297). Homorepeats are frequent in our LCR-focused set of 21 proteins. Using a relatively lax cutoff of 4 residues of the same type in a window of 6 (which was identified as already inducing structural effects for polyQ (298)), only TOLC_ECOLI has no homorepeat region (as detected with dAPE

(299)), hinting at the large overlap of LCRs with homorepeats. While there is a variety of homorepeat types, we can observe preferences in particular sequences, like polyS in SRP40_YEAST, DSPP_HUMAN and RPB1_HUMAN, polyP in CO1A1_HUMAN, or polyG in FUS_HUMAN. Elastin has many polyA and polyG tracts, since these residues participate in motifs discussed above that surround and support functional lysines and prolines. PolyQ is present once in TBP_HUMAN (followed by polyA), EPS15_HUMAN, HD_HUMAN, and three times in ANDR_HUMAN. All overlap the predicted regions by CAST (which identifies the Q-rich region) and IUPred (indicating disorder). While there was no overlap with FIELDS (PASTA 2.0) indicating aggregation, the aggregation propensity regions predicted by ArchCandy (300) do overlap with the three regions (in TBP, HD and ANDR) that are involved in polyQ repeat expansions causing disease (301). This result suggests that ArchCandy detects aggregation of the type involved in CAG/CAA triplet expansions. The ArchCandy analysis of our dataset identifies aggregation regions in a subset of the proteins identified by PASTA 2.0, suggesting that distinct methods for detection of aggregation have different sensitivity depending on the sequence.

Analysis of repeating patterns of charged regions/residues As discussed above, repetition within LCRs can result in structure and function. Another type of repetition that can occur within LCRs, beyond homorepeats, are those with alternating blocks of oppositely charged residues. To our knowledge, the only such motif that has been characterized in detail is the Charged Single Alpha-Helix (CSAH), also often referred to simply as Single Alpha-Helix (SAH). In these regions, generally 3 to 4 negatively charged residues are followed by 3 to 4 positively charged ones, although only few of such repeats are perfect. The structure of these segments is an alpha-helix that is stable in water as a monomer. CSAH segments can act as rigid linkers, rulers or lever arms in various proteins (302, 303, 304), and may also behave as constant force springs [PMID 25122759]. CSAHs are very rare in protein sequences, and in a number of cases are adjacent to coiled coil segments. One of the most well-characterized segments is found in myosin 6, where it forms the extended lever arm (302). There are currently three methods for detecting CSAHs in protein sequences, Waggawagga (305), FT_CHARGE and SCAN4CSAH, which are generally used together for consensus predictions (306). Of these, FT_CHARGE identifies repeating charge patterns of any frequency, not just

4. RESULTS & DISCUSSION



Figure 4.28: Secondary structure prediction (helix in yellow, sheet in purple and coil in grey) along the sequence of three proteins in the dataset. License: Attribution-NonCommercial-NoDerivatives 4.0 International. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

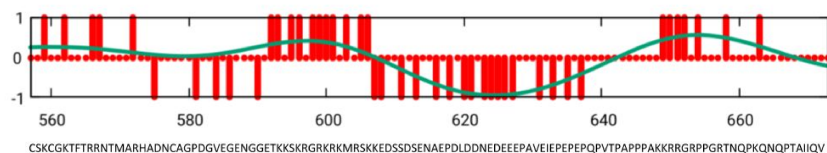


Figure 4.29: The C-terminal segment of CTCF_HUMAN with alternating, oppositely charged blocks. The region between residues 557-673 is shown. Charges of individual residues are plotted (red sticks) along with a smoothed bezier curve (green). License: Attribution-NonCommercial-NoDerivatives 4.0 International.

those characteristic of CSAHs. We applied the FT_CHARGE method (306) allowing all repeat frequencies to our dataset of 21 proteins. In agreement with their known low frequency, we only found CSAHs in two of the 21 proteins: a short region in Huntingtin (HD_HUMAN, residues 2633-2664), and a 120 amino acid segment in the human transcriptional repressor CTCF (CTCF_HUMAN, residues 557-673, Figure 4.29). The first 20 residues of the CTCF region largely match the 11th, atypical Zinc-finger motif of the protein as annotated in UniProt (positions 555-577). The structural information available for this protein suggests that its C-terminal part is intrinsically unstructured (307). However, this is typically found for CSAHs because, due to their highly charged nature, they are almost always predicted to be intrinsically disordered for most of their length (308). However, CSAHs can adopt a stable conformation as monomers (e.g. (302)). The notion that several structural motifs formed by LCRs are predicted to be intrinsically disordered is often found in the literature (309, 310, 311, 312). Most notably, there are many segments that are predicted to form alpha-helical coiled coils and also to be intrinsically disordered. In the case of coiled coils this can be justified on the basis that coiled coil forming regions are generally viewed as disordered in their monomeric state and they adopt helical conformation upon dimerization/multimerization (313). Collagen triple-helical motifs are another example of similar behavior, providing a case of folding upon binding/multimerization (314). In the next section, we study the overlaps of these structural predictions to LCRs.

ID	dis	only dis	dis + cc	dis + coll	cc	only cc	coll	only coll
Q38PT6_9HEXA	0	0	0	0	0	0	48	48
SYS_THET2	0	0	0	0	63	63	0	0
EPS15_HUMAN	287	228	59	0	161	102	0	0
STAG1_HUMAN	202	202	0	0	31	31	0	0
CO1A1_HUMAN	1168	390	0	778	0	0	778	0
ATA_ACIBT	546	450	96	0	96	0	0	0

Table 4.15: Number of residues predicted to be in different structural states. (dis) disordered, (cc) coiled coils, (coll) collagen. License: Attribution-NonCommercial-NoDerivatives 4.0 International.

Overlap of structural predictions and LCRs Our previous analyses suggest that LCRs tend to lie in regions without much structure. However, there are LCRs with repetitions that seem to provide structure, even multiple structures influenced by interactions with protein partners. To illustrate the overlaps of different structural pre-

4. RESULTS & DISCUSSION

dictions and LCRs, we use again our protein dataset. Overlaps of predictions were computed in three steps. First, we applied IUPred (282), VSL2B (315), ncoils (316), Paircoil2 (317) and hmmsearch (318) using Collagen.hmm (Pfam family PF01391), all with default parameters. Then, using in-house scripts, we computed (i) the consensus of the two disorder predicting methods, IUPred and VSL2B (only regions with a minimum of 30 residues predicted by both methods were considered), and (ii) the consensus of the coiled coil predicting methods, ncoils and Paircoil2 (only regions with a minimum of 21 residues predicted by both methods were considered). Finally, we computed the number of residues predicted to be disordered, located in coiled coil regions or in collagen helices (according to their similarity to collagen evaluated with hmmsearch). No residue was predicted to be both in a collagen helix and in a coiled coil: such overlap is unrealistic because of the incompatible structural preferences of amino acids (both Gly or Pro, abundant in collagen helices, are very rare in alpha-helical regions). Collagen and coiled coils were predicted for two and four proteins, respectively (Table 4.15). Full overlap to disorder was found for the collagen predicted for Q38PT6_9HEXA (Glycine-rich antifreeze protein) and partially for the coiled coils in EPS15_HUMAN (epidermal growth factor receptor substrate 15) and ATA_ACIBT (adhesin autotransporter). While these overlaps might reflect reality in terms of dynamic rearrangements of the segments, the general wisdom could be that the more specific prediction should usually be considered, meaning that coiled coil and collagen predictions have prevalence over disorder predictions. In this respect, disorder detection is regarded as a method to recognize non-globular sequences that might either form fibrillar structures or be disordered in their functional form, depending, among others, on their repetitiveness.

Multimerization: a final variable adding complexity to the study of LCRs

As discussed above, structural variability and folding upon binding are properties that can characterize some LCRs. Thus, the structural behavior of LCRs is context dependent. The interactions of LCRs with either additional copies of the same molecule (homomultimers) or other proteins/(macro)molecules (heteromeric complexes) is a key factor and largely influences the ability of the sequence to adopt a specific structure or interchange between conformations. Current methods are typically either able to predict the structure of the isolated molecule or the propensity to form specific structures, which typically stem from the underlying repeated sequence. The limitation of

such methods is that they usually predict homomultimeric structures, because it is impractical to consider the sequence information of all possible interaction partners. However, there are efforts to identify interaction motifs that might fold upon partner interaction (e.g. ANCHOR (319)). Indeed, application of this method to our protein dataset indicates some cases where this property applies (Figure 4), and, while there is a general overlap of folding propensity overlapping LCRs, there are also examples of striking complementarity (e.g. DSPP_HUMAN).

4.5 Protein aggregation and protein solubility

Finally, we exploited the knowledge acquired in the studies described in sections 4.1, 4.2, 4.4 and 4.3 to design a novel method, SODA, to predict changes in protein solubility. SODA uses the propensity of the protein sequence to aggregate as well as intrinsic disorder, plus hydrophobicity and secondary structure preferences derived from sequence features and complexity to predict a sequence-based solubility profile. SODA is able to evaluate solubility changes introduced by a mutation by comparing the profiles of the wild type (WT) and mutated sequences, and it is compatible with different types of variation including point mutations, deletions and insertions. The comparison to other recently published methods shows that SODA has state-of-the-art performance and is particularly well suited to predict mutations decreasing solubility. The method is fast, returning results for single mutations in seconds. A usage example estimating the full repertoire of mutations for a human germline antibody highlights several solubility hotspots on the surface.

4.5.1 SODA: Prediction of protein solubility from disorder and aggregation propensity

SODA is a novel method to predict the effects of variations on protein solubility. It is based on the disorder and aggregation propensities of a protein plus secondary structure and hydrophobicity in comparison to the same values of its mutated form. The difference between the two determines the effect on solubility of the variation. SODA is entirely based on sequence features and allows to quickly scan a large number of mutations. The web server was designed to allow large-scale annotation through its RESTful web service, while the user interface provides an intuitive form to guide detailed selection of mutations based on sequence solubility plot and, if the protein structure is given, residues accessibility to solvent. SODA can be useful for several applications. Its main envisaged application is in protein engineering, where predicting the variation in protein solubility upon mutation can help design proteins with more favorable surface properties. This can be of interest to pharmaceutical companies designing novel antibodies, as demonstrated by the usage example, as lack of solubility is a bottleneck in the development of biologicals. In addition, SODA may be of use in the context of

4.5 Protein aggregation and protein solubility

studying the impact of natural protein variants and their potential effect on disease insurgence.

4.5.1.1 Benchmarking

In Table 4.16, SODA performance using only sequence information is compared with the published solubility predictors CamSol (320), SOLpro (321) and Proso II (322). SODA correctly predicts all variations and its accuracy is higher than the other tested methods, even though the dataset is biased towards positive examples increasing solubility.

	Trevino	Miklos	Tan	Dudgeon	Total	Accuracy
SolPro	15 / 22	3 / 3	1 / 1	21 / 30	40 / 56	71.4
PROSO II	16 / 22	3 / 3	1 / 1	12 / 30	32 / 56	57.1
CamSol	22 / 22	3 / 3	1 / 1	28 / 30	54 / 56	96.4
SODA	22 / 22	3 / 3	1 / 1	30 / 30	56 / 56	100.0

Table 4.16: SODA is compared to three published methods. The dataset is the same used in the recent CamSol paper (187) and includes 19 proteins and 56 variants from four publications: Trevino (188), Miklos (189), Tan (190) and Dudgeon (191). Accuracy is calculated as the percentage of correct predictions over the dataset size.

4.5.1.2 Server description

SODA provides two types of analysis, namely mutation mode and full-protein mode. The first provides the solubility change on sequence mutation. The second generates a profile describing the contribution to solubility of each sequence position deduced from the effect of all possible mutations. The mutation mode requires the sequence and a list of mutations as input. The full-protein mode requires just the sequence since SODA automatically generates all possible single point variations (19 amino acid alternatives x sequence length) and then calculates the fraction of mutations increasing (and decreasing) the solubility for each position. In both cases, a PDB structure can be provided to label buried/exposed residues. The input page is the same for both modes but after input the route splits. While the mutation mode requires only seconds, full-protein analysis is more time consuming, with linear complexity proportional to sequence length. For example, evaluating a protein of 350 residues takes about 3 h. The SODA interface is straightforward to use. The home page features an input form, which accepts either a sequence or PDB structure. When the structure is provided (file

4. RESULTS & DISCUSSION

or ID) the server parses the PDB file, extracts the sequence and masks buried residues. Even though SODA is sequence based, this can help the user avoid introducing mutations in the core of a globular protein, which can potentially break the fold, altering its function and leading to meaningless results.

Mutation mode When the user chooses the mutation mode, the web server redirects to a new submission page (see Figure 4.30). The user introduces mutations by clicking on the stretch of residues to be modified directly from the input wild type sequence. A new edit box pops up when residues are selected, allowing to introduce/modify/delete residues until the save command is issued. Multiple mutation instances can be created and submitted as a single job. The solubility profile of the WT is plotted on the top of the page to help the user in the editing process. When a PDB input is provided, buried residues are shaded but still editable. The results page provides a table summarizing the comparison between WT and mutation (Figure 4.31). It provides WT/mutation differences for SODA and its components (aggregation, disorder, secondary structure helix and strand). Detailed SODA output is reported on the bottom, including the wild type and mutated stretches. When a PDB file is provided, the results page also shows the corresponding structure, highlighting the mutated region (Figure 4.31).

Full-protein mode The full-protein mode only requires the sequence or PDB file as input. Like the mutation mode, the results page (Figure 4.32) provides the solubility profile for the input sequence. When the structure is available, buried residues are missing from the plot and excluded from the calculation. For each position all possible amino acid substitutions are evaluated. The number of mutations increasing (and decreasing) solubility is plotted. Below the plot, a table reports for each position the list of substitutions sorted according to their impact on solubility.

4.5.1.3 Usage examples

The crystal structure of human germline antibody IGHV1-69/IGKV1-39 (PDB code 5i15) was recently determined (323). The light and heavy chains are composed of 214 and 228 amino acids respectively. SODA was used to calculate the potential effect of mutations on each residue of the molecule (full-protein mode). It predicts the effect of each possible point substitution on each position of the light and heavy chains,

4.5 Protein aggregation and protein solubility

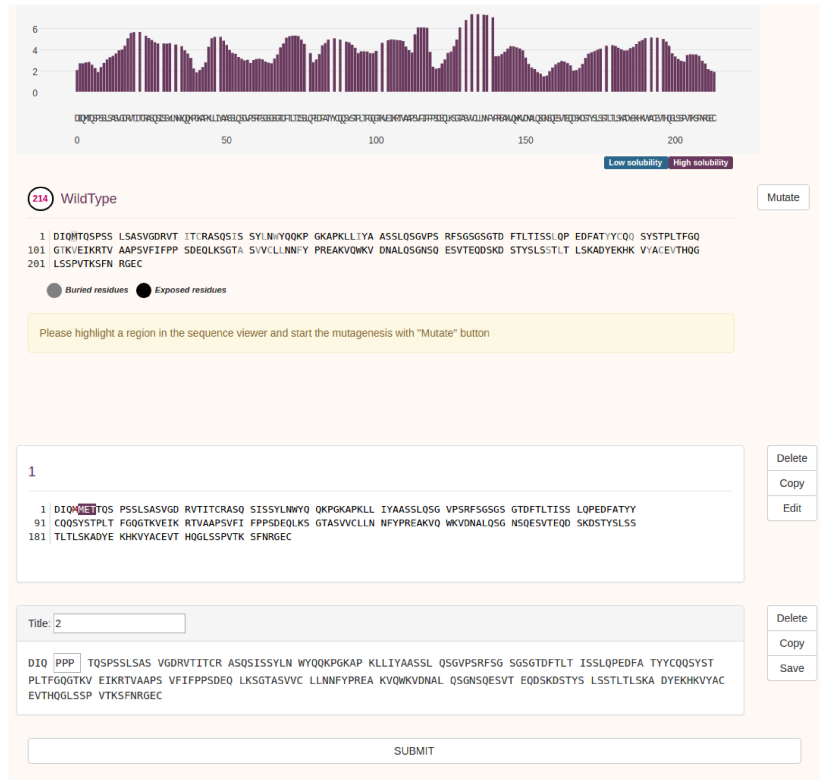


Figure 4.30: The mutation input page is returned when the user chooses mutation mode. It allows to create multiple instances of mutations/deletions/insertions. (178)

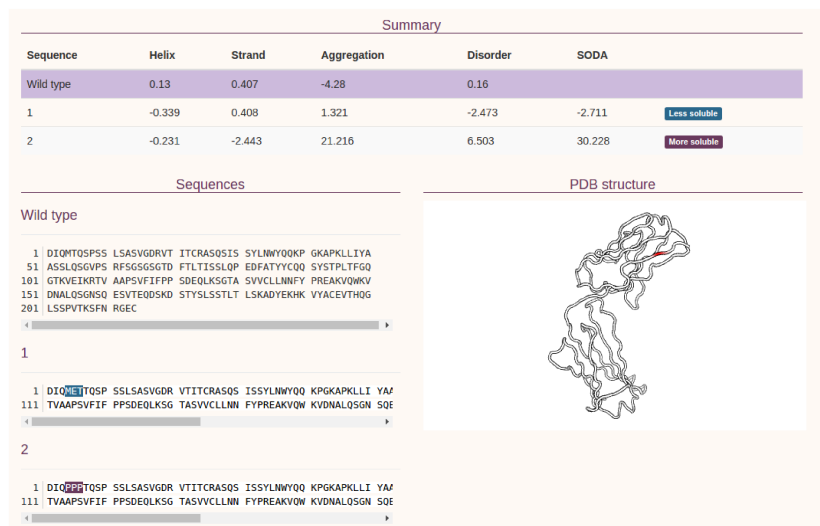


Figure 4.31: The mutation mode result reports changes of the protein solubility upon mutation. When a PDB is provided as input the mutated region is highlighted in the structure. (178)

4. RESULTS & DISCUSSION



Figure 4.32: The full protein mode provides the solubility profile (first plot) and the propensity of increasing or decreasing solubility for all sequence positions (second plot and table). (178)

for a total of 8398 mutations (19 amino acid substitutions on 214 + 228 positions). Figure 4.33 shows the SODA output for the antibody light and heavy chains in the 3D structure of the protein complex. The light (L) and heavy (H) chains are shown with different representation in order to show the connecting surfaces. Red residues have high probability of increasing protein solubility when mutated. On the contrary, blue positions indicate an aggregation propensity upon mutation. The wild type residues in this position show a selective pressure to be the most soluble among all possibilities, thus the simulated mutations are likely to impair this property. Notably, blue positions are mostly localized in the surface indicating them as hot spots for solvent interactions. The "Full-protein mode" enables the analysis of the whole protein surface such as the case presented, while the "Mutation mode" may be used to evaluate effects on solubility of a mutation. The sickle cell anemia is a disease usually caused by a missense mutation in human Hemoglobin subunit beta (UniProt AC: P68871), substituting a Glutamate with a Valine at position 7 along the protein sequence (324). It leads to a deformation of red cells, resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs in homozygous patients (325). Sickle cell hemoglobin

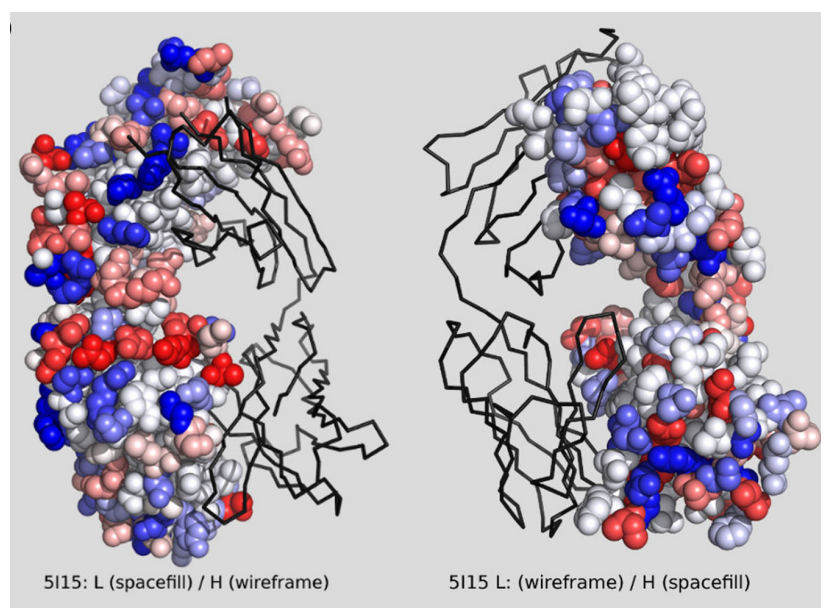


Figure 4.33: The human germline antibody IGHV1-69/IGKV1-39 (PDB code 5i15) is shown alternatively as wireframe and space fill between light (L) and heavy (H) chain. For each position, the probability of increasing (red) or decreasing (blue) solubility upon mutation is mapped on the structure. On the left, the light (L) and heavy (H) chains are shown as wireframe and space fill respectively, on the right the same protein with opposite chain visualization mode is provided. (178)

detection methods, from the very beginning, rely on the protein solubility and have proved to be practical, cheap and to have a high degree of sensitivity (326). Indeed, sickle cell hemoglobin was demonstrated to be less soluble than the wild type (327). There is a clear relationship between the disease and the protein solubility. SODA correctly predicts the mutation as decreasing protein solubility.

4. RESULTS & DISCUSSION

Conclusions

The proteomes of organisms, even the more studied such as human, are far from being completely annotated. This is true both at the sequence level, e.g. the less than 50% of human proteome residues is covered by a Pfam domain (77), and at the structure level, e.g. 44% of the structures of eukaryotic proteins have not been determined and cannot be derived from homologous proteins (78). The main reason for this lack of annotation is that most of the methods for protein characterization were designed for the first (1, 2) type of proteins that were identified, namely globular proteins. Whatever diverges from it, in terms of stability, solubility or architecture, since the very beginning of protein science was considered non-functional and thus few efforts have been done for the identification of these exceptions. Only in the last few decades, a number of phenomena has been discovered, that not only are widespread in evolution and in different cellular compartments and tissues, but also perform fundamental functions. During my PhD, I focused on this type of phenomena contributing to their detection, classification and annotation.

The main focus of this thesis is the building of resources and analysis protocols for the characterization of protein tandem repeats, results are described in sections 4.1 and 4.2. Despite their wide distribution and functional heterogeneity, repeat proteins still belong to the dark matter of structural biology due to the inherent difficulties of

5. CONCLUSIONS

sequence and structural identification and classification (328). Several efforts have been made during the years to improve the tools for their detection (328). These methods are either sequence- or structure-based and they are used for the construction of sequence- and structure-based databases and classification. However, the protein universe cannot be explored without taking into account the relationship between amino acid sequences and their 3D structures, analyzed in light of their evolutionary and functional background.

In section 4.1.1 I provided a first look into the relationship between repeat structures and their Pfam families. In the majority of cases a strict one-to-one relationship was found, with the expected tendency for structure to be more conserved than sequence in the remaining cases. The Leucine Rich Repeat example however shows that it is also possible for members of a large family to fall into different structural classes. The comparison between structures and protein families proved beneficial to a better understanding of repeat protein evolution and design, and guided the next steps of our analysis. In section 4.1.2 I presented a census of TRPs in the human proteome. A central observation derived from this analysis is that TRPs localize in several different cell compartments and human tissues, and they are involved in a plethora of different pathways. The common denominator is that this heterogeneous group of proteins serve as binders. The reason for preferential election as connector nodes in the interaction network lies in their very own nature as modular structures. They are extended structures, with high surface to volume ratio, characterized by very few long-range internal interactions. The high availability of exposed residues prone to evolve binding specialization allows high sequence plasticity, able to evolve and specialize fast. The evolutionary flexibility does not only concern the amino acids in the sequence of the TPRs, but also in the length and number of repeats (79). In many cases, TPRs can easily tolerate the insertion of a new structural unit, thus adding a new potential binding interface free to specialize for a new interactor without compromising the binding to the ones already present. Given the fact that the presence of a repetition at the DNA level promotes additional duplications (204), the process of new unit insertion is probably widespread in the TPR universe, which commonly evolved specificity for more than one partner at the same time (81). Their structure makes them the best eligible candidates to be highly interacting nodes in the protein-protein interaction network. However, once the protein is specialized for interaction, it immediately needs the

residues involved to be conserved and plasticity decreases dramatically, increasing the conservation level (329). This, and their centrality in the network (212), provides an explanation for their enrichment in disease-associated proteins. All findings therefore suggest that the better understanding of TRPs behaviour may considerably contribute to a better understanding of the cell system and of the pathways compromised in disease insurgency. In addition, their binding role suggests a myriad of potential uses in the bioengineering of target recognition, which has already been demonstrated to be a promising field of application of TRPs science (204, 208). As a case-study of TRPs association to disease in the light of their role as platforms and hubs, in section 4.1.3 I presented the *in silico* dissection of collagen V genotype-phenotype correlations in relation to its interactome. Collagen V mutations are indeed associated with EhlersDanlos syndrome (EDS) (213), a group of heritable collagenopathies with heterogeneous phenotype. Collagen V structure is not available and the disease-causing mechanism is unclear. To address this issue, we manually curated missense mutations suspected to promote classic type EDS (cEDS) insurgence from the literature. Further, we generated a homology model of the collagen V triple helix to evaluate the pathogenic effects. The resulting structure was used to map known proteinprotein interactions enriched with *in silico* predictions. An interaction network model for collagen V was created. We found that cEDS heterogeneous manifestations may be explained by the involvement in two different extracellular matrix pathways, related to cell adhesion and tissue repair or cell differentiation, growth and apoptosis. We believe that the data presented here can give a useful insight on collagen V specific properties and will be useful to drive future experimental validation as well as helping in patient classification.

TRPs importance in biological systems and their potential pathogenicity motivates our effort in their identification and annotation. In section 3.1.1 I presented ReUPred, a predictor that we used to identify repeat structures in the Protein Data Bank and populate the second version of RepeatsDB database. RepeatsDB was originally presented in 2014 with the goal to provide the community with a central resource for high-quality tandem repeat protein structure annotation. It has been cited in a number of different studies regarding repeat proteins, and has been used to extract datasets for repeat proteins analysis and to test algorithms for repeat proteins annotation. The detailed annotation of entries performed in the first version by RepeatsDB curators has allowed us to build a high quality Structure Repeat Unit Library (SRUL). This

5. CONCLUSIONS

library was exploited by the ReUPred algorithm (results described in section 4.2.1) as a gold standard to define unit position in new entries in an iterative process. The tool takes as input a target PDB structure and aligns it against the SRUL. Through structural alignment, it is able to annotate on the target repeat units and insertions. The comparison of ReUPred to other tools for repeat detection shows that the predictor has state-of-the-art performances in repeat classification and is one of the few and best performing resources for unit position annotation. Indeed, the prediction of tandem repeat units is a challenging problems currently mainly addressed by expert manual curation. This work has demonstrated that repeat protein annotation can be made by repetitive template-based structural searches. Moreover, it shows that the approach can be applied reliably on a large scale, i.e., over all uncharacterized RepeatsDB entries, unveiling new scenarios for the analysis of the entire repeat protein universe. The second release of RepeatsDB (in section 4.2.2) includes a new annotation pipeline, combining the RAPHAEL algorithm for repeat detection and ReUPred for annotation, producing extensive annotation for all entries. The pipeline is fully automated and allows the easy regular update of the database. The iterative execution of the pipeline already demonstrated its efficacy both because it identified a large number of new entries, and because new subclasses were detected and added to the structural classification scheme. RepeatsDB will benefit from regular updates, which will steadily increase the number of available annotations. Future perspectives for the development of the database included exploiting repeat unit definitions to create profiles for use in detecting repeats from sequence for genome-scale analysis and the facilitation of RepeatsDB revision process in order to achieve a complete coverage of RepeatsDB in terms of manually curated data. In line with the former, we started a collaboration with Pfam database to curate repeat families based on structural data. In line with the latter, in section 4.2.3 I presented RepeatsDB-lite. It is a web server that allows to identify units and classify the protein exploiting a structural similarity search and the information available in RepeatsDB. The prediction outperforms existing methods and can be applied to all types of TR proteins. The web interface allows to visualize similarity relationships between TR units at both the sequence and structure level. The prediction can be manually refined by the user, visualizing the effects of the edits in real time. Annotations can be submitted to RepeatsDB for reviewed with the aim to increase the amount of community-curated entries in the database. RepeatsDB-Lite

can be seen as an example of gamification principles to engage a wider community towards database curation.

Another non-globular phenomena that, for different reasons, shares the high connectivity character of TRPs is protein disorder. Intrinsically disordered proteins (IDPs) or regions (IDRs) are devoid of order in their native unbound state (126, 330). Intrinsic disorder is prevalent in the human proteome (125), appears to play important signaling and regulatory roles (126) and is frequently involved in disease (127). The discovery of intrinsic disorder and its prevalence and functional importance is transforming the field of molecular biology. As intrinsic disorder is emerging as a general phenomenon, databases are collecting and presenting disorder related data in a systematic manner. MobiDB has been a major contributor by providing consensus predictions and functional annotations for all UniProt proteins, driving the field ahead (157, 245). The MobiDB upgrade I presented in section 4.3.1 is essential for several reasons. MobiDB 3.0 improves on previous releases by adding descriptions of conformational diversity and disorder-related functions, both in terms of experimental data and predictions. A particular field where it may have a significant impact is the establishment of a long-awaited disorder sequence-function relationship schema. The most reliable proxy to this goal is to assess the function of a protein by homology transfer, i.e. transferring functional annotation based on sequence similarity. A large-scale analysis of IDP functional annotations will be necessary to find adequate boundaries for transferring IDP functions by homology. As sufficient data is now available in MobiDB 3.0, we expect a rapid advance in the field of sequence-function correlations of IDPs. In addition, for proteins with sufficient NMR data, MobiDB now features quantitative annotations incorporating structure and equilibrium dynamics in a unified framework. These large-scale quantitative annotations will help understand the biological role of order and disorder, and serve as a basis to construct predictive models. MobiDB is widely used by scientific community and by third party services, it is becoming a thematic hub for IDPs and future work will focus on including IDP annotations into core data resources such as UniProt.

From the sequence point of view, non-globular proteins are characterized by non-typical sequences as well, as they are largely characterized by low complexity (LC). In section 4.4.1 I presented a critical review where we focused on the description of several features of LCRs by using computational methods. We chose a dataset of proteins with

5. CONCLUSIONS

a variety of functions and types of LCRs to test these methods and their overlapping predictions. At the strict level of sequence, low complexity is related to composition bias and repeats. At the level of structure, there is a direct, yet not fully understood, relation to disorder, aggregation and flexibility. While some connections have been established previously, we demonstrate the difficulty of defining general rules connecting sequence features and structural properties. We hypothesize that the problem lies in the strong non-linearities of the connections between the sequence/structure relationships in low complexity sequences. On one side, the non-linearity depends from the fact that variables used to characterize sequences cannot capture all the effects of amino acid combinations at the structural level. The second reason for this non-linearity lays on the unusual pattern of conservation of disordered regions, which complicates any standalone predictions. We have tried a pragmatic approach with two sides. On the one hand, a diagram of sequence properties that allows to explore the overlaps in three variables (repeat perfection, composition bias and low complexity). Along this exemplary path, we exploited the dataset to submit it to a variety of analyses and illustrate their potential overlaps. The structural aspects were discussed separately. The main conclusion from this latter section is that low complexity manifests itself in apparently opposite effects: while disorder and flexibility seem to be common features of LCRs, repetition/periodicity in sequence at multiple levels can induce structure. In evolutionary terms, this might imply that a disordered (low complexity) sequence can escape disorder by either gaining a richer (higher complexity) composition maintaining aperiodicity, or by attaining a highly periodic structure. We have demonstrated the intricacies of analyzing low complexity in protein sequences: even methods that are supposed to study the same properties (low complexity and sequence bias) might not share similar assumptions. Our recommendation for researchers investigating a particular protein is to use several of these methods together. The additional advantage in having these multiple outputs is that the sequence context might be influencing the structure adopted by a low complexity region. In this respect, joint bioinformatics research and development efforts to make the outputs of these methods compatible and consistent are highly desirable. We expect that ongoing efforts will lead to a more specific classification of LCRs, aiming at the prediction of their function.

A concept tightly connected to protein folding, and consequently to the features that determine their stability or flexibility, is protein solubility. In section [4.5.1](#) I presented

SODA, a novel method to predict the changes of protein solubility based on several physico-chemical properties of the protein. It is based on the disorder (183) and aggregation (182) propensities of a protein plus secondary structure (185) and hydrophobicity (184) of the wild type protein in comparison to the same values of its mutated form. The difference between the two determines the effect on solubility of the variation. SODA is entirely based on sequence features and allows to quickly scan a large number of mutations. The web server was designed to allow large-scale annotation through its RESTful web service, while the user interface provides an intuitive form to guide detailed selection of mutations based on sequence solubility plot and, if the protein structure is given, residues accessibility to solvent. SODA can be useful for several applications. Its main envisaged application is in protein engineering, where predicting the variation in protein solubility upon mutation can help design proteins with more favorable surface properties (40, 41, 42). This can be of interest to pharmaceutical companies designing novel antibodies (331), as demonstrated by the usage example on a human germline antibody. Lack of solubility is indeed a bottleneck in the development of biologicals. In addition, SODA may be of use in the context of studying the impact of natural protein variants and their potential effect on disease insurgence (37, 38, 39), as shown in the example of sickle cell hemoglobin.

NGPs recognition and classification is essential to shed a light on the so called "dark proteome", i.e. the large fraction that we know almost nothing about. An entire community of scientists accepted the challenge, and during my PhD I had the opportunity to work together with several of them within the framework of the NGPnet COST Action. I contributed to this goal through the development of new resources dedicated to NGPs, with the aim to identify and classify them. Defining not only non-globular phenomena themselves but also their relationships, e.g. the overlap between tandem repeat and intrinsical disorder, is of paramount importance for understanding a wide variety of functional arrangements, molecular processes and mechanism of evolution currently unknown. This knowledge is essential also to dissect pathogenic mechanisms yet to be discovered, e.g. linked to protein aggregation. Finally, this opens up new possibilities for the redesign of enzyme activities and the building of proteins tailored to have specific properties, as demonstrated by the case of repeat proteins that are turned into new biomaterials.

5. CONCLUSIONS

References

- [1] Kendrew, J., Dickerson, R., Strandberg, B., Hart, R., Davies, D., Phillips, D., and Shore, V. Structure of myoglobin. *Nature* **185**(422), 427–1960 (1960). [1](#), [135](#)
- [2] Muirhead, H. and Perutz, M. F. Structure of reduced human hemoglobin. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 28, 451–459. Cold Spring Harbor Laboratory Press, (1963). [1](#), [135](#)
- [3] Crippen, G. M. The tree structural organization of proteins. *Journal of molecular biology* **126**(3), 315–332 (1978). [1](#)
- [4] Ramachandran. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**(1), 9599 (1963). [3](#), [4](#), [5](#)
- [5] Kabsch, Wolfgang, S. How good are predictions of protein secondary structure? *FEBS Letters* **155**(2), 179182 (1983). [4](#)
- [6] Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**(31), 7133–7155 (1990). [5](#)
- [7] Chou, P. Y. and Fasman, G. D. Empirical predictions of protein conformation. *Annual review of biochemistry* **47**(1), 251–276 (1978). [5](#)
- [8] Blaber, M., Zhang, X. J., and Matthews, B. W. Structural basis of amino acid alpha helix propensity. *Science* **260**(5114), 1637–1640 (1993). [5](#)
- [9] Avbelj, F. and Baldwin, R. L. Role of backbone solvation in determining thermodynamic β propensities of the amino acids. *Proceedings of the National Academy of Sciences* **99**(3), 1309–1313 (2002). [5](#)
- [10] Zimmerman, S. B. and Trach, S. O. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of escherichia coli. *Journal of molecular biology* **222**(3), 599–620 (1991). [5](#)
- [11] Miller, S., Janin, J., Lesk, A. M., and Chothia, C. Interior and surface of monomeric proteins. *Journal of molecular biology* **196**(3), 641–656 (1987). [6](#)
- [12] Lins, L., Thomas, A., and Brasseur, R. Analysis of accessible surface of residues in proteins. *Protein science* **12**(7), 1406–1417 (2003). [6](#)
- [13] Devos, D. and Russell, R. B. A more complete, complexed and structured interactome. *Current opinion in structural biology* **17**(3), 370–377 (2007). [6](#)
- [14] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. From molecular to modular cell biology. *Nature* **402**(6761supp), C47 (1999). [6](#)
- [15] Smyth, M. and Martin, J. x ray crystallography. *Molecular Pathology* **53**(1), 8 (2000). [7](#)
- [16] Wüthrich, K. The way to nmr structures of proteins. *Nature Structural & Molecular Biology* **8**(11), 923 (2001). [7](#)
- [17] Abola, E. E., Bernstein, F. C., and Koetzle, T. F. The protein data bank. In *Neutrons in Biology*, 441–441. Springer (1984). [7](#)
- [18] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. Scop2 prototype: a new approach to protein structure mining. *Nucleic acids research* **42**(D1), D310–D314 (2013). [8](#)
- [19] Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research* **45**(D1), D289–D295 (2016). [8](#)
- [20] Alva, V., Remmert, M., Biegert, A., Lupas, A. N., and Söding, J. A galaxy of folds. *Protein Science* **19**(1), 124–130 (2010). [8](#)
- [21] Silverstein, T. P. Principles of physical biochemistry (van holde, kersal e.; johnson, w. curtis; ho, p. shing). *Journal of Chemical Education* **76**(4), 474 (1999). [8](#)
- [22] Perrot, P. *A to Z of Thermodynamics*. Oxford University Press on Demand, (1998). [9](#)
- [23] Pullman, P. *His dark materials*. Bluefire, (2007). [9](#)
- [24] Wu, H. Studies on denaturation of proteins xiii. a theory of denaturation. In *Advances in protein chemistry*, volume 46, 6–26. Elsevier (1995). [9](#)
- [25] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**(4096), 223–230 (1973). [9](#)
- [26] Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. Forces contributing to the conformational stability of proteins. *The FASEB journal* **10**(1), 75–83 (1996). [10](#)
- [27] Levinthal, C. How to fold graciously. *Mossbauer spectroscopy in biological systems* **67**, 22–24 (1969). [10](#)
- [28] Rooman, M., Dehouck, Y., Kwasigroch, J. M., Biot, C., and Gilis, D. What is paradoxical about levinthal paradox? *Journal of Biomolecular Structure and Dynamics* **20**(3), 327–329 (2002). [11](#)

REFERENCES

- [29] Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics* **21**(3), 167–195 (1995). [11](#)
- [30] Ohgushi, M. and Wada, A. 'molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS letters* **164**(1), 21–24 (1983). [12](#)
- [31] Loladze, V. V. and Makhatadze, G. I. Removal of surface charge-charge interactions from ubiquitin leaves the protein folded and very stable. *Protein Science* **11**(1), 174–177 (2002). [12](#), [13](#)
- [32] Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N., and Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophysical journal* **102**(8), 1907–1915 (2012). [13](#)
- [33] Manning, M. C., Chou, D. K., Murphy, B. M., Payne, R. W., and Katayama, D. S. Stability of protein pharmaceuticals: an update. *Pharmaceutical research* **27**(4), 544–575 (2010). [13](#)
- [34] Garnier, J., Osguthorpe, D. J., and Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology* **120**(1), 97–120 (1978). [13](#)
- [35] Trainor, K., Broom, A., and Meiering, E. M. Exploring the relationships between protein sequence, structure and solubility. *Current opinion in structural biology* **42**, 136–146 (2017). [13](#)
- [36] Balch, W. E., Morimoto, R. I., Dillin, A., and Kelly, J. W. Adapting proteostasis for disease intervention. *science* **319**(5865), 916–919 (2008). [13](#)
- [37] Ciryam, P., Tartaglia, G. G., Morimoto, R. I., Dobson, C. M., and Vendruscolo, M. Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell reports* **5**(3), 781–790 (2013). [13](#), [141](#)
- [38] Thal, D. R., Walter, J., Saido, T. C., and Fändrich, M. Neuropathology and biochemistry of $\alpha\beta$ and its aggregates in alzheimers disease. *Acta neuropathologica* **129**(2), 167–182 (2015). [13](#), [141](#)
- [39] Knowles, T. P., Vendruscolo, M., and Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nature reviews Molecular cell biology* **15**(6), 384 (2014). [13](#), [141](#)
- [40] Esposito, D. and Chatterjee, D. K. Enhancement of soluble protein expression through the use of fusion tags. *Current opinion in biotechnology* **17**(4), 353–358 (2006). [13](#), [141](#)
- [41] Williams, H. D., Trevaskis, N. L., Charman, S. A., Shanker, R. M., Charman, W. N., Pouton, C. W., and Porter, C. J. Strategies to address low drug solubility in discovery and development. *Pharmacological reviews* **65**(1), 315–499 (2013). [13](#), [141](#)
- [42] Savjani, K. T., Gajjar, A. K., and Savjani, J. K. Drug solubility: importance and enhancement techniques. *ISRN pharmaceuticals* **2012** (2012). [13](#), [141](#)
- [43] Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D., and Davidson, A. R. A simple in vivo assay for increased protein solubility. *Protein Science* **8**(9), 1908–1911 (1999). [13](#)
- [44] Meulemans, A., Seneca, S., Pribyl, T., Smet, J., Alderweirdt, V., Waeytens, A., Lissens, W., Van Coster, R., De Meirleir, L., di Rago, J.-P., et al. Defining the pathogenesis of the human atp12p w94r mutation using a *saccharomyces cerevisiae* yeast model. *Journal of Biological Chemistry* **285**(6), 4099–4109 (2010). [13](#)
- [45] Andley, U. P. and Reilly, M. A. In vivo lens deficiency of the r49c α -crystallin mutant. *Experimental eye research* **90**(6), 699–702 (2010). [13](#)
- [46] Chothia, C. and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**(4), 823–826 (1986). [14](#)
- [47] Chothia, C. and Lesk, A. The evolution of protein structures. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, 399–405. Cold Spring Harbor Laboratory Press, (1987). [14](#)
- [48] Rost, B. Twilight zone of protein sequence alignments. *Protein engineering* **12**(2), 85–94 (1999). [15](#)
- [49] Bork, P., Sander, C., and Valencia, A. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science* **2**(1), 31–40 (1993). [15](#)
- [50] Devos, D. and Valencia, A. Practical limits of function prediction. *Proteins: Structure, Function, and Bioinformatics* **41**(1), 98–107 (2000). [15](#)
- [51] Lee, D., Redfern, O., and Orengo, C. Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* **8**(12), 995 (2007). [15](#)
- [52] Walusinski, G. The game of possibilities-essay on the diversity of the living-french-jacob, f, (1982). [15](#)
- [53] Jensen, L. J., Ussery, D. W., and Brunak, S. Functionality of system components: conservation of protein function in protein feature space. *Genome research* **13**(11), 2444–2449 (2003). [15](#)
- [54] Lupas, A. N., Ponting, C. P., and Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of structural biology* **134**(2-3), 191–203 (2001). [15](#)
- [55] Söding, J. and Lupas, A. N. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**(9), 837–846 (2003). [15](#)
- [56] Alva, V., Söding, J., and Lupas, A. N. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, e09410 (2015). [15](#)
- [57] Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. Structure, function and evolution of multidomain proteins. *Current opinion in structural biology* **14**(2), 208–216 (2004). [15](#), [16](#)

REFERENCES

- [58] Phillips, D. C. The three-dimensional structure of an enzyme molecule. *Scientific American* **215**(5), 78–93 (1966). [15](#)
- [59] Richardson, J. S. The anatomy and taxonomy of protein structure. In *Advances in protein chemistry*, volume 34, 167–339. Elsevier (1981). [15](#)
- [60] Wetlaufer, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences* **70**(3), 697–701 (1973). [15](#)
- [61] Bork, P. Shuffled domains in extracellular proteins. *FEBS letters* **286**(1-2), 47–54 (1991). [15](#), [16](#)
- [62] Davidson, J. N., Chen, K. C., Jamison, R. S., Musmanno, L. A., and Kern, C. B. The evolutionary history of the first three enzymes in pyrimidine biosynthesis. *Bioessays* **15**(3), 157–164 (1993). [15](#)
- [63] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–285 (2016). [15](#), [99](#)
- [64] Finn, R. D., Mistry, J., Schuster-Bekler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. Pfam: clans, web tools and services. *Nucleic acids research* **34**, D247–D251 (2006). [15](#), [16](#)
- [65] Eddy, S. R. Accelerated profile HMM searches. *PLoS Computational Biology* **7**(10) (2011). [16](#)
- [66] Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., and Tramontano, A. Protein function annotation by homology-based inference. *Genome biology* **10**(2), 207 (2009). [16](#)
- [67] Koshland Jr, D. E. The key-lock theory and the induced fit theory. *Angewandte Chemie International Edition in English* **33**(23-24), 2375–2378 (1995). [17](#), [18](#)
- [68] Bu, Z. and Callaway, D. J. Proteins move! protein dynamics and long-range allostery in cell signaling. In *Advances in protein chemistry and structural biology*, volume 83, 163–221. Elsevier (2011). [18](#)
- [69] Li, D. and Roberts, R. Human genome and diseases: Wd-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cellular and Molecular Life Sciences CMLS* **58**(14), 2085–2097 (2001). [19](#)
- [70] Midic, U., Oldfield, C. J., Dunker, A. K., Obradovic, Z., and Uversky, V. N. Protein disorder in the human diseaseome: unfoldomics of human genetic diseases. *Bmc Genomics* **10**(1), S12 (2009). [19](#)
- [71] Mosavi, L. K., Cammett, T. J., Desrosiers, D. C., and Peng, Z.-y. The ankyrin repeat as molecular architecture for protein recognition. *Protein Science* **13**(6), 1435–1448 (2004). [19](#)
- [72] D’Andrea, L. D. and Regan, L. Tpr proteins: the versatile helix. *Trends in biochemical sciences* **28**(12), 655–662 (2003). [19](#)
- [73] Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. A census of protein repeats. *Journal of molecular biology* **293**(1), 151–160 (1999). [19](#), [22](#), [38](#)
- [74] Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., and Rost, B. Protein disorder breakthrough invention of evolution? *Current opinion in structural biology* **21**(3), 412–418 (2011). [19](#)
- [75] Wright, P. E. and Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of molecular biology* **293**(2), 321–331 (1999). [19](#)
- [76] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradović, Z. Intrinsic disorder and protein function. *Biochemistry* **41**(21), 6573–6582 (2002). [19](#)
- [77] Mistry, J., Coggill, P., Eberhardt, R. Y., Deiana, A., Giansanti, A., Finn, R. D., Bateman, A., and Punta, M. The challenge of increasing pfam coverage of the human proteome. *Database* **2013** (2013). [19](#), [23](#), [135](#)
- [78] Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences* **112**(52), 15898–15903 (2015). [19](#), [135](#)
- [79] Kajava, A. V. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology* **179**(3), 279–288 (2012). [19](#), [22](#), [23](#), [24](#), [53](#), [54](#), [60](#), [108](#), [136](#)
- [80] Di Domenico, T., Potenza, E., Walsh, I., Gonzalo Parra, R., Giollo, M., Minervini, G., Piovesan, D., Ihsan, A., Ferrari, C., Kajava, A. V., and Tosatto, S. C. E. RepeatsDB: A database of tandem repeat protein structures. *Nucleic Acids Research* **42**, 1–6 (2014). [19](#), [24](#), [38](#), [54](#)
- [81] Kobe, B. and Kajava, A. V. When protein folding is simplified to protein coiling: The continuum of solenoid protein structures. *Trends in Biochemical Sciences* **25**(10), 509–515 (2000). [20](#), [60](#), [136](#)
- [82] Kajava, A. V. Review: proteins with repeated sequence structural prediction and modeling. *J Struct Biol* **134**, 132 – 144 (2001). [20](#)
- [83] Kobe, B. and Kajava, A. V. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* **11**, 725 – 732 (2001). [21](#), [56](#), [57](#)
- [84] Gamblin, S. J. and Smerdon, S. J. Nuclear transport: what a kary-on! *Structure* **7**(9), R199–R204 (1999). [21](#)
- [85] Ekman, D., Light, S., Björklund, Å. K., and Elofsson, A. What properties characterize the hub proteins of the protein-protein interaction network of *saccharomyces cerevisiae*? *Genome biology* **7**(6), 1 (2006). [21](#), [63](#)
- [86] Jones, D. A. and Jones, J. D. The role of leucine-rich repeat proteins in plant defences. In *Advances in botanical research*, volume 24, 89–167. Elsevier (1997). [21](#)

REFERENCES

- [87] Paladin, L. and Tosatto, S. Comparison of protein repeat classifications based on structure and sequence families. *Biochemical Society Transactions* **43**(5), 832–837 (2015). [21](#), [34](#), [55](#)
- [88] Zweifel, M. E., Leahy, D. J., Hughson, F. M., and Barrick, D. Structure and stability of the ankyrin domain of the drosophila notch receptor. *Protein Science* **12**(11), 2622–2632 (2003). [21](#)
- [89] Lowe, A. R. and Itzhaki, L. S. Biophysical characterization of the small ankyrin repeat protein myotrophin. *Journal of molecular biology* **365**(4), 1245–1255 (2007). [21](#)
- [90] Aksel, T. and Barrick, D. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods in enzymology* **455**, 95–125 (2009). [21](#)
- [91] Truhlar, S. M., Mathes, E., Cervantes, C. F., Ghosh, G., and Komives, E. A. Pre-folding *ikba* alters control of *nf- κ b* signaling. *Journal of molecular biology* **380**(1), 67–82 (2008). [22](#), [29](#)
- [92] Bjrkklund, . K., Ekman, D., and Elofsson, A. Expansion of protein domain repeats. *PLoS Comput Biol* **2**(8), e114 (2006). [22](#), [24](#), [38](#)
- [93] Moore, A. D., Bjrkklund, . K., Ekman, D., Bornberg-Bauer, E., and Elofsson, A. Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences* **33**(9), 444–451 (2008). [22](#)
- [94] Rose-Martel, M., Smiley, S., and Hincke, M. T. Novel identification of matrix proteins involved in calcitic biomineralization. *Journal of Proteomics* **116**(0), 81–96 (2015). [22](#)
- [95] Soler-Llavina, G., Arstikaitis, P., Morishita, W., Ahmad, M., Schof, T., and Malenka, R. Leucine-rich repeat transmembrane proteins are essential for maintenance of long-term potentiation. *Neuron* **79**(3), 439–446 (2013). [22](#)
- [96] Bukowska, M. A. and Grtter, M. G. New concepts and aids to facilitate crystallization. *Current Opinion in Structural Biology* **23**(3), 409–416 (2013). [22](#)
- [97] Nikitovic, D., Aggelidakis, J., Young, M. F., Iozzo, R. V., Karamanos, N. K., and Tzanakakis, G. N. The biology of small leucine-rich proteoglycans in bone pathophysiology. *Journal of Biological Chemistry* **287**(41), 33926–33933 (2012). [22](#)
- [98] Esteves, A. R., Swerdlow, R. H., and Cardoso, S. M. LRRK2, a puzzling protein: Insights into parkinson’s disease pathogenesis. *Experimental Neurology* **261**(0), 206–216 (2014). [22](#)
- [99] Ferreira, D. U., Walczak, A. M., Komives, E. a., and Wolynes, P. G. The energy landscapes of repeat-containing proteins: Topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Computational Biology* **4**(5) (2008). [22](#)
- [100] Pellegrini, M., Renda, M. E., and Vecchio, A. Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. *BMC Bioinformatics* **13**, S8 (2012). [23](#)
- [101] Andrade, M. A., Ponting, C. P., Gibson, T. J., and Bork, P. Homology-based method for identification of protein repeats using statistical significance estimates. *Journal of molecular biology* **298**(3), 521–537 (2000). [23](#)
- [102] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. Pfam: The protein families database. *Nucleic Acids Research* **42**, 222–230 (2014). [23](#), [103](#)
- [103] Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research* **43**(D1), D213–D221 (2014). [23](#)
- [104] Heger, A. and Holm, L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**(2), 224 – 237 (2000). [23](#)
- [105] Szklarczyk, R. and Heringa, J. Tracking repeats using significance and transitivity. *Bioinformatics* **20**(suppl.1), i311–i317 (2004). [23](#)
- [106] Newman, A. M. and Cooper, J. B. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinforma* **8**, 382 (2007). [23](#)
- [107] Jorda, J. and Kajava, A. V. T-REKS: identification of tandem REpeats in sequences with a k-meanS based algorithm. *Bioinformatics* **25**(20), 2632–2638 (2009). [23](#)
- [108] Soding, J., Remmert, M., and Biegert, A. HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res* **34**, W137 – W142 (2006). [23](#)
- [109] Biegert, A. and Sding, J. De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* **24**(6), 807–814 (2008). [23](#)
- [110] Gruber, M., Soding, J., and Lupas, A. N. REPPER-repeats and their periodicities in fibrous proteins. *Nucleic Acids Res* **33**, W239 – W243 (2005). [23](#)
- [111] Schaper, E. and Anisimova, M. The evolution and function of protein tandem repeats in plants. *New Phytol.* **206**(1), 397–410 (2015-04). [23](#)
- [112] Abraham, A.-L., Rocha, E. P. C., and Pothier, J. Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics* **24**(13), 1536–1537 (2008). [23](#)
- [113] Sabarinathan, R., Basu, R., and Sekar, K. ProSTRIP: A method to find similar structural repeats in three-dimensional protein structures. *Computational Biology and Chemistry* **34**(2), 126–130 (2010). [23](#)
- [114] Walsh, I., Sirocco, F. G., Minervini, G., Di Domenico, T., Ferrari, C., and Tosatto, S. C. E. RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics* **28**(24), 3257–3264 (2012). [23](#), [24](#), [41](#), [79](#), [87](#)
- [115] Hrabe, T. and Godzik, A. ConSOLE: using modularity of contact maps to locate solenoid domains in protein structures. *BMC Bioinformatics* **15**(1), 119 (2014). [23](#), [42](#), [45](#)

REFERENCES

- [116] Do Viet, P., Roche, D. B., and Kajava, A. V. TAPO: A combined method for the identification of tandem repeats in protein structures. *FEBS Letters* **589**(19), 2611–2619 (2015). [23](#), [42](#), [45](#)
- [117] Brunette, T., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A., and Baker, D. Exploring the repeat protein universe through computational protein design. *Nature* **528**(7583), 580 (2015). [24](#)
- [118] Heidenfelder, B. L. and Topal, M. D. Effects of sequence on repeat expansion during DNA replication. *Nucleic Acids Res* **31**(24), 7159–7164 (2003). [24](#)
- [119] Gemayel, R., Vincens, M. D., Legendre, M., and Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**(1), 445–477 (2010). [24](#)
- [120] Jorda, J., Xue, B., Uversky, V. N., and Kajava, A. V. Protein tandem repeats - the more perfect, the less structured. *FEBS Journal* **277**(12), 2673–2682 (2010). [24](#), [29](#), [30](#), [57](#), [60](#), [108](#)
- [121] Ellegren, H. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in genetics* **16**(12), 551–558 (2000). [24](#)
- [122] Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Vermeesch, J. R., Verstrepen, K. J., and Froyen, G. Large-scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res* **42**(9), 5728–5741 (2014). [24](#)
- [123] Wright, P. E. and Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* **293**(2), 321–331 (1999). [25](#)
- [124] Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **25**(9), 847–855 (2003). [25](#), [29](#)
- [125] Pancsa, R. and Tompa, P. Structural disorder in eukaryotes. *PLoS one* **7**(4), e34687 (2012). [25](#), [139](#)
- [126] Wright, P. E. and Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology* **16**(1), 18 (2015). [25](#), [29](#), [139](#)
- [127] Uversky, V. N., Oldfield, C. J., and Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008). [25](#), [29](#), [139](#)
- [128] Oldfield, C. J. and Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual review of biochemistry* **83**, 553–584 (2014). [26](#)
- [129] 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., et al. Intrinsically disordered protein. *Journal of molecular graphics and modelling* **19**(1), 26–59 (2001). [26](#), [27](#)
- [130] Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. Protein disorder prediction: implications for structural proteomics. *Structure* **11**(11), 1453–1459 (2003). [26](#), [104](#)
- [131] Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A. K., Daughdrill, G. W., and Uversky, V. N. The alphabet of intrinsic disorder: I. act like a pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins* **1**(1), e24360 (2013). [26](#)
- [132] Uversky, V. N. The alphabet of intrinsic disorder: II. various roles of glutamic acid in ordered and intrinsically disordered proteins. *Intrinsically disordered proteins* **1**(1), e24684 (2013). [26](#)
- [133] Uversky, V. N. The intrinsic disorder alphabet. iii. dual personality of serine. *Intrinsically disordered proteins* **3**(1), e1027032 (2015). [26](#)
- [134] Hernández, M. A., Avila, J., and Andreu, J. M. Physicochemical characterization of the heat-stable microtubule-associated protein map2. *European journal of biochemistry* **154**(1), 41–48 (1986). [26](#)
- [135] Schweers, O., Schönbrunn-Hanebeck, E., Marx, A., and Mandelkow, E. Structural studies of tau protein and alzheimer paired helical filaments show no evidence for beta-structure. *Journal of Biological Chemistry* **269**(39), 24290–24297 (1994). [26](#)
- [136] Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. Nacp, a protein implicated in alzheimer’s disease and learning, is natively unfolded. *Biochemistry* **35**(43), 13709–13715 (1996). [26](#)
- [137] DeForte, S. and Uversky, V. N. Order, disorder, and everything in between. *Molecules* **21**(8), 1090 (2016). [26](#), [27](#)
- [138] Collins, M. O., Yu, L., Campuzano, I., Grant, S. G., and Choudhary, J. S. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Molecular & Cellular Proteomics* **7**(7), 1331–1348 (2008). [27](#)
- [139] Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., and Uversky, V. N. Analysis of molecular recognition features (morfs). *Journal of molecular biology* **362**(5), 1043–1059 (2006). [27](#)
- [140] Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T. J. Attributes of short linear motifs. *Molecular BioSystems* **8**(1), 268–281 (2012). [27](#)
- [141] Pawson, T. and Scott, J. D. Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278**(5346), 2075–2080 (1997). [27](#)
- [142] Niklas, K. J., Bondos, S. E., Dunker, A. K., and Newman, S. A. Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. *Frontiers in cell and developmental biology* **3**, 8 (2015). [27](#)

REFERENCES

- [143] Bhowmick, P., Pancsa, R., Guharoy, M., and Tompa, P. Functional diversity and structural disorder in the human ubiquitination pathway. *PLoS one* **8**(5), e65443 (2013). [27](#)
- [144] Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., et al. Classification of intrinsically disordered regions and proteins. *Chemical reviews* **114**(13), 6589–6631 (2014). [27](#)
- [145] Hyman, A. A., Weber, C. A., and Jülicher, F. Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology* **30**, 39–58 (2014). [27](#)
- [146] Shorter, J. Membraneless organelles: phasing in and out. *Nature chemistry* **8**(6), 528 (2016). [27](#)
- [147] Toretzky, J. A. and Wright, P. E. Assemblages: functional units formed by cellular phase separation. *J Cell Biol* **206**(5), 579–588 (2014). [27](#)
- [148] Brangwynne, C. P., Tompa, P., and Pappu, R. V. Polymer physics of intracellular phase transitions. *Nature Physics* **11**(11), 899–904 (2015). [27](#)
- [149] Berlow, R. B., Dyson, H. J., and Wright, P. E. Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature* **543**(7645), 447 (2017). [28](#)
- [150] Mylona, A., Theillet, F.-X., Foster, C., Cheng, T. M., Miralles, F., Bates, P. A., Selenko, P., and Treisman, R. Opposing effects of elk-1 multisite phosphorylation shape its response to erk activation. *Science* **354**(6309), 233–237 (2016). [28](#)
- [151] Bah, A., Vernon, R. M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., Sonenberg, N., Kay, L. E., and Forman-Kay, J. D. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* **519**(7541), 106 (2015). [28](#)
- [152] Uversky, V. N. Biophysical methods to investigate intrinsically disordered proteins: avoiding an elephant and blind men situation. In *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*, 215–260. Springer (2015). [28](#)
- [153] Uversky, V. N. and Dunker, A. K. Multiparametric analysis of intrinsically disordered proteins: looking at intrinsic disorder through compound eyes, (2012). [28](#)
- [154] Ferron, F., Longhi, S., Canard, B., and Karlin, D. A practical overview of protein disorder prediction methods. *Proteins: Structure, Function, and Bioinformatics* **65**(1), 1–14 (2006). [28](#)
- [155] Necci, M., Piovesan, D., Dosztányi, Z., Tompa, P., and Tosatto, S. C. A comprehensive assessment of long intrinsic protein disorder from the disprot database. *Bioinformatics* **34**(3), 445–452 (2017). [28](#)
- [156] Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., et al. Disprot: the database of disordered proteins. *Nucleic acids research* **35**(suppl.1), D786–D793 (2006). [28](#)
- [157] Di Domenico, T., Walsh, I., Martin, A. J., and Tosatto, S. C. Mobidb: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* **28**(15), 2080–2081 (2012). [28](#), [139](#)
- [158] He, X. and Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS genetics* **2**(6), e88 (2006). [28](#)
- [159] Djinovic-Carugo, K., Gautel, M., Ylänne, J., and Young, P. The spectrin repeat: a structural platform for cytoskeletal protein assemblies. *FEBS letters* **513**(1), 119–123 (2002). [28](#)
- [160] Collavin, L., Lunardi, A., and Del Sal, G. p53-family proteins and their regulators: hubs and spokes in tumor suppression. *Cell death and differentiation* **17**(6), 901 (2010). [29](#)
- [161] Ferreiro, D. U. and Wolynes, P. G. The capillarity picture and the kinetics of one-dimensional protein folding. *Proceedings of the National Academy of Sciences* **105**(29), 98539854 (2008). [29](#)
- [162] Burger, V. M., Gurry, T., and Stultz, C. M. Intrinsically disordered proteins: Where computation meets experiment. *Polymers* **6**(10), 2684–2719 (2014). [29](#)
- [163] Simon, M. and Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome biology* **10**(6), R59 (2009). [29](#)
- [164] Peng, Z., Yan, J., Fan, X., Mizianty, M. J., Xue, B., Wang, K., Hu, G., Uversky, V. N., and Kurgan, L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cellular and Molecular Life Sciences* **72**(1), 137–151 (2015). [29](#)
- [165] Frith, M. C. Gentle masking of low-complexity sequences improves homology search. *PLoS one* **6**(12), e28819 (2011). [29](#)
- [166] Mier, P., Alanis-Lobato, G., and Andrade-Navarro, M. A. Context characterization of amino acid homorepeats using evolution, position, and order. *Proteins: Structure, Function, and Bioinformatics* **85**(4), 709–719 (2017). [29](#)
- [167] Darling, A. L. and Uversky, V. N. Intrinsic disorder in proteins with pathogenic repeat expansions. *Molecules* **22**(12), 2027 (2017). [29](#)
- [168] Wootton, J. C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers & chemistry* **18**(3), 269–285 (1994). [30](#)
- [169] Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *Journal of structural biology* **179**(3), 279–288 (2012). [31](#)
- [170] Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K., and Zaslavsky, B. Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS letters* **589**(1), 15–22 (2015). [31](#)
- [171] Darling, A. L., Liu, Y., Oldfield, C. J., and Uversky, V. N. Intrinsically disordered proteome of human membrane-less organelles. *Proteomics* **18**(5-6), 1700193 (2018). [31](#)

REFERENCES

- [172] Lin, Y.-H., Forman-Kay, J. D., and Chan, H. S. Theories for sequence-dependent phase behaviors of biomolecular condensates. *Biochemistry* **57**(17), 2499–2508 (2018). [31](#)
- [173] Paladin, L., Hirsh, L., Piovesan, D., Andrade-Navarro, M. A., Kajava, A. V., and Tosatto, S. C. E. RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Res.* **45**, D308–D312 (2017). [33](#), [34](#), [37](#), [43](#), [89](#), [90](#), [91](#), [92](#)
- [174] Necci, M., Piovesan, D., Dosztányi, Z., and Tosatto, S. C. Mobidb-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**(9), 1402–1404 (2017). [33](#), [35](#), [102](#), [103](#), [105](#)
- [175] Paladin, L., Tosatto, S. C., and Minervini, G. Structural in silico dissection of the collagen v interactome to identify genotype-phenotype correlations in classic ehlers-danlos syndrome (EDS). *FEBS Letters* **589**(24), 3871–3878 (2015). [34](#), [68](#), [70](#), [74](#)
- [176] Hirsh, L., Piovesan, D., Paladin, L., and Tosatto, S. C. E. Identification of repetitive units in protein structures with ReUPred. *Amino Acids* **48**(6), 1391–1400 (2016). [34](#), [37](#), [40](#), [42](#), [43](#), [45](#), [79](#), [80](#), [81](#), [82](#), [83](#), [84](#), [85](#), [87](#), [96](#)
- [177] Hirsh, L., Paladin, L., Piovesan, D., and Tosatto, S. C. Repeatsdb-lite: a web server for unit annotation of tandem repeat proteins. *Nucleic Acids Research* **46**(W1), W402–W407 (2018). [34](#), [37](#), [93](#), [95](#), [96](#), [97](#), [98](#)
- [178] Paladin, L., Piovesan, D., and Tosatto, S. C. Soda: prediction of protein solubility from disorder and aggregation propensity. *Nucleic acids research* **45**(W1), W236–W240 (2017). [36](#), [46](#), [131](#), [132](#), [133](#)
- [179] Zhang, Y. and Skolnick, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research* **33**(7), 2302–2309 (2005). [39](#), [41](#), [45](#)
- [180] Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**(5), 680–682 (2010). [41](#), [44](#)
- [181] Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. Mustang: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics* **64**(3), 559–574 (2006). [44](#)
- [182] Walsh, I., Seno, F., Tosatto, S. C. E., and Trovato, A. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res* **42**, W301–W307 (2014). [46](#), [141](#)
- [183] Walsh, I., Martin, A. J. M., Di Domenico, T., and Tosatto, S. C. E. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**(4), 503–509 (2012). [46](#), [104](#), [141](#)
- [184] Kyte, J. and Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**(1), 105–132 (1982). [46](#), [141](#)
- [185] Piovesan, D., Walsh, I., Minervini, G., and Tosatto, S. C. Fells: fast estimator of latent local structure. *Bioinformatics* **33**(12), 1889–1891 (2017). [46](#), [47](#), [105](#), [141](#)
- [186] Yang, Y., Niroula, A., Shen, B., and Vihinen, M. PONsol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics* **32**(13), 2032–2034 (2016). [47](#)
- [187] Sormanni, P., Aprile, F. A., and Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology* **427**(2), 478–490 (2015). [48](#), [129](#)
- [188] Trevino, S. R., Scholtz, J. M., and Pace, C. N. Measuring and increasing protein solubility. *Journal of Pharmaceutical Sciences* **97**(10), 4155–4166 (2008). [48](#), [129](#)
- [189] Miklos, A., Kluwe, C., Der, B., Pai, S., Sircar, A., Hughes, R., Berrondo, M., Xu, J., Codrea, V., Buckley, P., Calm, A., Welsh, H., Warner, C., Zacharko, M., Carney, J., Gray, J., Georgiou, G., Kuhlman, B., and Ellington, A. Structure-based design of supercharged, highly thermoresistant antibodies. *Chemistry & Biology* **19**(4), 449–455 (2012). [48](#), [129](#)
- [190] Tan, P. H., Chu, V., Stray, J. E., Hamlin, D. K., Pettit, D., Wilbur, D. S., Vessella, R. L., and Stayton, P. S. Engineering the isoelectric point of a renal cell carcinoma targeting antibody greatly enhances scFv solubility. *Immunotechnology* **4**(2), 107–114 (1998). [48](#), [129](#)
- [191] Dudgeon, K., Rouet, R., Kokmeijer, I., Schofield, P., Stolp, J., Langley, D., Stock, D., and Christ, D. General strategy for the generation of human antibody variable domains with increased aggregation resistance. *PNAS* **109**(27), 10879–10884 (2012). [48](#), [129](#)
- [192] Brinckmann, J. Collagens at a glance. *Topics in Current Chemistry* **247**, 1–6 (2005). [56](#)
- [193] Mason, J. M. and Arndt, K. M. Coiled coil domains: Stability, specificity, and biological implications. *ChemBioChem* **5**(2), 170–176 (2004). [56](#)
- [194] Fournier, D., Palidwor, G. a., Shcherbinin, S., Szengel, A., Schaefer, M. H., Perez-Iratxeta, C., and Andrade-Navarro, M. a. Functional and genomic analyses of alpha-solenoid proteins. *PLoS ONE* **8**(11) (2013). [57](#)
- [195] Ng, A. C. Y., Eisenberg, J. M., Heath, R. J. W., Huett, A., Robinson, C. M., Nau, G. J., and Xavier, R. J. Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proceedings of the National Academy of Sciences of the United States of America* **108** Suppl, 4631–4638 (2011). [57](#)
- [196] Smith, T. F., Gaitatzes, C., Saxena, K., and Neer, E. J. The WD repeat: A common architecture for diverse functions. *Trends in Biochemical Sciences* **24**(5), 181–185 (1999). [58](#)
- [197] Stirnimann, C. U., Petsalaki, E., Russell, R. B., and Mller, C. W. WD40 proteins propel cellular networks. *Trends in Biochemical Sciences* **35**(10), 565–574 (2010-10). [58](#)
- [198] Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. Protein repeats: structures, functions, and evolution. *J Struct Biol* **134**(2), 117 – 131 (2001). [60](#)

REFERENCES

- [199] Lowe, A. R. and Itzhaki, L. S. Rational redesign of the folding pathway of a modular protein. *PNAS* **104**(8), 2679–2684 (2007). 60
- [200] Exposito, J.-Y., Valcourt, U., Cluzel, C., and Lethias, C. The fibrillar collagen family. *International Journal of Molecular Sciences* **11**(2), 407–426 (2010). 60
- [201] Chaudhuri, I., Sding, J., and Lupas, A. N. Evolution of the -propeller fold. *Proteins* **71**(2), 795–803 (2008). 60
- [202] Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P., and Plckthun, A. Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *Journal of Molecular Biology* **332**(2), 489–503 (2003). 60
- [203] Park, K., Shen, B. W., Parmeggiani, F., Huang, P.-S., Stoddard, B. L., and Baker, D. Control of repeat-protein curvature by computational protein design. *Nat Struct Mol Biol* **22**(2), 167–174 (2015). 60
- [204] Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E., and Regan, L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Current Opinion in Structural Biology* **15**(4), 464–471 (2005). 60, 136, 137
- [205] Bjrkklund, . K., Light, S., Sagit, R., and Elofsson, A. Nebulin: A study of protein repeat evolution. *Journal of Molecular Biology* **402**(1), 38–51 (2010). 60
- [206] Magliery, T. J. and Regan, L. Beyond consensus: Statistical free energies reveal hidden interactions in the design of a TPR motif. *Journal of Molecular Biology* **343**(3), 731–745 (2004). 61
- [207] Jha, S. and Ting, J. P.-Y. Inflammasome-associated nucleotide-binding domain, leucine-rich repeat proteins and inflammatory diseases. *The Journal of Immunology* **183**(12), 7623–7629 (2009). 61
- [208] Grove, T. Z., Cortajarena, A. L., and Regan, L. Ligand binding by repeat proteins: natural and designed. *Current Opinion in Structural Biology* **18**(4), 507–515 (2008). 61, 137
- [209] Gavin, A. C. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002). 63
- [210] Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O’Shea, E. K. Global analysis of protein localization in budding yeast. *Nature* **425**(6959), 686–691 (2003). 63
- [211] Jobling, R., DSouza, R., Baker, N., Lara-Corrales, I., Mendoza-Londono, R., Dupuis, L., Savarirayan, R., Ala-Kokko, L., and Kannu, P. The collagenopathies: Review of clinical phenotypes and molecular correlations. *Curr Rheumatol Rep* **16**(1), 1–13 (2013). 64
- [212] Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**(6833), 41–42 (2001). 66, 137
- [213] De Paepe, A. and Malfait, F. The ehlers-danlos syndrome, a disorder with many faces. *Clinical Genetics* **82**(1), 1–11 (2012). 66, 137
- [214] Symoens, S. Col5a1 signal peptide mutations interfere with protein secretion and cause classic ehlers-danlos syndrome. *Human Mutation* **30**(2) (2008). 67
- [215] Wenstrup, R. J. Murine model of the ehlers-danlos syndrome. *Journal of Biological Chemistry* **281**(18), 1288812895 (2006). 67, 71
- [216] Ritelli, M. Clinical and molecular characterization of 40 patients with classic ehlers-danlos syndrome: Identification of 18 col5a1 and 2 col5a2 novel mutations. *Orphanet Journal of Rare Diseases* **8**(1), 58 (2013). 67
- [217] Grond-Ginsbach, C. Sequence analysis of the col5a2 gene in patients with spontaneous cervical artery dissections. *Neurology* **58**(7), 11031105 (2002). 67
- [218] Castori, V. neurological manifestations of ehlers-danlos syndrome(s): a review. *iran. J. Neurol* **13**(4), 190208 (2014). 67
- [219] Giunta, C., , and Steinmann, B. Compound heterozygosity for a disease-causing g1489d and disease-modifying g530s substitution incol5a1 of a patient with the classical type of ehlers-danlos syndrome: An explanation of intrafamilial variability? *American Journal of Medical Genetics* **90**(1), 7279 (2000). 67
- [220] Giunta, C. Homozygous gly530ser substitution incol5a1 causes mild classical ehlers-danlos syndrome. *American Journal of Medical Genetics* **109**(4), 284290 (2002). 67
- [221] Symoens, S. comprehensive molecular analysis demonstrates type. *Human Mutation* **Vvol. 33, no. 10,** 14851493 (2012). 67, 69
- [222] Sherry, S. Dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research* **29**(1), 308311 (2001). 67, 69
- [223] de Leeuw, K., Goorhuis, J. F., Tielliu, I. F. J., Symoens, S., Malfait, F., de Paepe, A., van Tintel, J. P., and Hulscher, J. B. F. Superior mesenteric artery aneurysm in a 9-year-old boy with classical ehlers-danlos syndrome. *Am. J. Med. Genet. A* **158A**(3), 626–629 (2012-03). 67
- [224] Mitchell, A. L. Molecular mechanisms of classical ehlers-danlos syndrome (eds). *Human Mutation* **30**(6), 9951002 (2009). 69
- [225] Symoens, S. identification of binding partners interacting with the 1-n-propeptide of type. *Biochemical Journal* **Vvol. 433, no. 2,** 371381 (2011). 69
- [226] Ricard-Blum, S. The collagen family. *Cold Spring Harb Perspect Biol* **3**(1), a004978 (2011). 69
- [227] Adzhubei, A. A., Sternberg, M. J., and Makarov, A. A. Polyproline-II helix in proteins: Structure and function. *Journal of Molecular Biology* **425**(12), 2100–2132 (2013). 69
- [228] Consortium, U. Activities at the universal protein resource (uniprot). *Nucleic Acids Research* **42**(11), 74867486 (2014). 70, 87

REFERENCES

- [229] Stark, C. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research* **34**(90001) (2006). [71](#)
- [230] Franceschini, A. String v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41**(1) (2012). [71](#)
- [231] Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N., and Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Research* **39**, D235–D240 (2011). [71](#)
- [232] Nuytink, L. Classical ehlers-danlos syndrome caused by a mutation in type i collagen. *The American Journal of Human Genetics* **66**(4), 13981402 (2000). [71](#)
- [233] Klaassens, M., Reinstein, E., Hilhorst-Hofstee, Y., Schrandt, J., Malfait, F., Staal, H., ten Have, L., Blaauw, J., Roggeveen, H., Krakow, D., De Paepe, A., van Steensel, M., Pals, G., Graham, J., and Schrandt-Stumpel, C. Ehlers-danlos arthrochalasia type (VIIAb) expanding the phenotype: from prenatal life through adulthood. *Clinical Genetics* **82**(2), 121–130 (2012). [71](#)
- [234] Abdelaziz, D. M. Behavioral signs of pain and functional impairment in a mouse model of osteogenesis imperfecta. *Bone* **81**(2015), 400406 (2015). [71](#)
- [235] Kadler, K. E., Baldock, C., Bella, J., and Boot-Handford, R. P. Collagens at a glance. *Journal of Cell Science* **120**(12), 1955–1958 (2007). [71](#)
- [236] Viglio, S., Zoppi, N., Sangalli, A., Gallanti, A., Barlati, S., Mottes, M., Colombi, M., and Valli, M. Rescue of migratory defects of ehlers-danlos syndrome fibroblasts in vitro by type v collagen but not insulin-like binding protein-1. *J. Invest. Dermatol.* **128**(8), 1915–1919 (2008). [72](#)
- [237] Schalkwijk, J. A recessive form of the ehlers-danlos syndrome caused by tenascin-x deficiency. *New England Journal of Medicine* **345**(16), 11671175 (2001). [72](#)
- [238] Ritelli, M., Dordoni, C., Venturini, M., Chiarelli, N., Quinzani, S., Traversa, M., Zoppi, N., Vascellaro, A., Wischmeijer, A., Manfredini, E., Garavelli, L., Calzavara-Pinton, P., and Colombi, M. Clinical and molecular characterization of 40 patients with classic ehlers-danlos syndrome: identification of 18 COL5A1 and 2 COL5A2 novel mutations. *Orphanet J Rare Dis* **8**, 58 (2013). [72](#), [73](#)
- [239] Malfait, F., Wenstrup, R. J., and De Paepe, A. Clinical and genetic aspects of ehlers-danlos syndrome, classic type. *Genet. Med.* **12**(10), 597–605 (2010). [72](#)
- [240] Giunta, C. and Steinmann, B. Compound heterozygosity for a disease-causing g1489e [correction of g1489d] and disease-modifying g530s substitution in COL5A1 of a patient with the classical type of ehlers-danlos syndrome: an explanation of intrafamilial variability? *Am. J. Med. Genet.* **90**(1), 72–79 (2000). [72](#)
- [241] Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., et al. Pdbe: improved accessibility of macromolecular structure data from pdb and emdb. *Nucleic acids research* **44**(D1), D385–D395 (2015). [87](#)
- [242] Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research* **41**, D483–D489 (2013-01-01). [87](#)
- [243] Kubiak, R. L. and Holden, H. M. Structural studies of antd: an n-acyltransferase involved in the biosynthesis of d-anthrose. *Biochemistry* **51**(4), 867–878 (2012). [94](#)
- [244] Piovesan, D. and Tosatto, S. C. Mobi 2.0: an improved method to define intrinsic disorder, mobility and linear binding regions in protein structures. *Bioinformatics* **34**(1), 122–123 (2017). [101](#), [103](#), [104](#)
- [245] Potenza, E., Domenico, T. D., Walsh, I., and Tosatto, S. C. Mobidb 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic acids research* **43**(D1), D315–D320 (2014). [101](#), [139](#)
- [246] Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S. D., Koike, R., Hiroaki, H., and Ota, M. Ideal in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic acids research* **42**(D1), D320–D325 (2013). [102](#)
- [247] Schad, E., Fichó, E., Pancsa, R., Simon, I., Dosztányi, Z., and Mészáros, B. Dibs: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**(3), 535–537 (2017). [102](#)
- [248] Fichó, E., Reményi, I., Simon, I., and Mészáros, B. Mfib: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **33**(22), 3682–3684 (2017). [102](#)
- [249] Dinkel, H., Van Roey, K., Michael, S., Davey, N. E., Weatheritt, R. J., Born, D., Speck, T., Krüger, D., Grebnev, G., Kubaň, M., et al. The eukaryotic linear motif resource elm: 10 years and counting. *Nucleic acids research* **42**(D1), D259–D266 (2013). [102](#)
- [250] Consortium, U. Uniprot: the universal protein knowledgebase. *Nucleic acids research* **45**(D1), D158–D169 (2016). [102](#), [105](#)
- [251] Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., et al. Disprot 7.0: a major update of the database of disordered proteins. *Nucleic acids research* **45**(D1), D219–D227 (2016). [102](#), [105](#)
- [252] Miskei, M., Antal, C., and Fuxreiter, M. Fuzdb: database of fuzzy complexes, a tool to develop stochastic structure-function relationships for protein complexes and higher-order assemblies. *Nucleic Acids Research* **45**(D1), D228–D235 (2017). [102](#)
- [253] Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* **19**(2), 327–335 (2009). [103](#)
- [254] Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., and Orengo, C. Gene3d: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic acids research* **40**(D1), D465–D471 (2011). [103](#)

REFERENCES

- [255] Monzon, A. M., Rohr, C. O., Fornasari, M. S., and Parisi, G. Codnas 2.0: a comprehensive database of protein conformational diversity in the native state. *Database* **2016** (2016). [103](#)
- [256] Martin, A. J., Walsh, I., and Tosatto, S. C. Mobi: a web server to define and visualize structural mobility in nmr protein ensembles. *Bioinformatics* **26**(22), 2916–2917 (2010). [103](#), [104](#)
- [257] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., et al. Biomagresbank. *Nucleic acids research* **36**(suppl.1), D402–D408 (2007). [103](#)
- [258] Piovesan, D., Minervini, G., and Tosatto, S. C. The ring 2.0 web server for high quality residue interaction networks. *Nucleic acids research* **44**(W1), W367–W374 (2016). [104](#)
- [259] Sormanni, P., Piovesan, D., Heller, G. T., Bonomi, M., Kukic, P., Camilloni, C., Fuxreiter, M., Dosztanyi, Z., Pappu, R. V., Babu, M. M., et al. Simultaneous quantification of protein order and disorder. *Nature chemical biology* **13**(4), 339 (2017). [104](#)
- [260] Camilloni, C., De Simone, A., Vranken, W. F., and Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* **51**(11), 2224–2231 (2012). [104](#)
- [261] Berjanskii, M. V. and Wishart, D. S. A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society* **127**(43), 14970–14971 (2005). [104](#)
- [262] Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**(16), 3433–3434 (2005). [104](#)
- [263] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC bioinformatics* **7**(1), 208 (2006). [104](#)
- [264] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., and Vranken, W. F. From protein sequence to dynamics and disorder with dynamine. *Nature communications* **4**, 2741 (2013). [105](#)
- [265] Mészáros, B., Simon, I., and Dosztányi, Z. Prediction of protein binding regions in disordered proteins. *PLoS computational biology* **5**(5), e1000376 (2009). [105](#)
- [266] Smith, M. H. The amino acid composition of proteins. *Journal of Theoretical Biology* **13**, 261–282 (1966). [109](#), [119](#)
- [267] Wootton, J. C. and Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & chemistry* **17**(2), 149–163 (1993). [111](#)
- [268] Kreil, D. P. and Ouzounis, C. A. Comparison of sequence masking algorithms and the detection of biased protein sequence regions. *Bioinformatics* **19**(13), 1672–1681, Sep (2003). [111](#)
- [269] Huntley, M. A. and Golding, G. B. Simple sequences are rare in the Protein Data Bank. *Proteins* **48**(1), 134–140, Jul (2002). [111](#)
- [270] Hao, J., Zou, B., Narayanan, K., and George, A. Differential expression patterns of the dentin matrix proteins during mineralized tissue formation. *Bone* **34**(6), 921–932, Jun (2004). [112](#)
- [271] Hao, J., Ramachandran, A., and George, A. Temporal and spatial localization of the dentin matrix proteins during dentin biomineralization. *J. Histochem. Cytochem.* **57**(3), 227–237, Mar (2009). [112](#)
- [272] Suzuki, S., Sreenath, T., Haruyama, N., Honeycutt, C., Terse, A., Cho, A., Kohler, T., Muller, R., Goldberg, M., and Kulkarni, A. B. Dentin sialoprotein and dentin phosphoprotein have distinct roles in dentin mineralization. *Matrix Biol.* **28**(4), 221–229, May (2009). [112](#)
- [273] Jadowiec, J., Koch, H., Zhang, X., Campbell, P. G., Seyedain, M., and Sfeir, C. Phosphorylation regulates the gene expression and differentiation of NIH3T3, MC3T3-E1, and human mesenchymal stem cells via the integrin/MAPK signaling pathway. *J. Biol. Chem.* **279**(51), 53323–53330, Dec (2004). [112](#)
- [274] Jadowiec, J. A., Zhang, X., Li, J., Campbell, P. G., and Sfeir, C. Extracellular matrix-mediated signaling by dentin phosphoprotein involves activation of the Smad pathway independent of bone morphogenetic protein. *J. Biol. Chem.* **281**(9), 5341–5347, Mar (2006). [112](#)
- [275] Eapen, A., Ramachandran, A., and George, A. Dentin phosphoprotein (DPP) activates integrin-mediated anchorage-dependent signals in undifferentiated mesenchymal cells. *J. Biol. Chem.* **287**(8), 5211–5224, Feb (2012). [112](#)
- [276] Eapen, A. and George, A. Dentin phosphoprotein in the matrix activates AKT and mTOR signaling pathway to promote preodontoblast survival and differentiation. *Front Physiol* **6**, 221 (2015). [112](#)
- [277] Promponas, V. J., Enright, A. J., Tsoka, S., Kreil, D. P., Leroy, C., Hamodrakas, S., Sander, C., and Ouzounis, C. A. CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* **16**(10), 915–922, Oct (2000). [113](#)
- [278] Tautz, D., Trick, M., and Dover, G. A. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**(6080), 652–656 (1986). [114](#)
- [279] Alba, M. M., Laskowski, R. A., and Hancock, J. M. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **18**(5), 672–678, May (2002). [114](#)
- [280] Simon, M. and Hancock, J. M. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol.* **10**(6), R59 (2009). [115](#)
- [281] Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. Sequence complexity of disordered protein. *Proteins* **42**(1), 38–48, Jan (2001). [117](#)

REFERENCES

- [282] Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**(16), 3433–3434, Aug (2005). [118](#), [126](#)
- [283] Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**(4), 827–839, Apr (2005). [118](#)
- [284] Gavira, J. A. Current trends in protein crystallization. *Arch. Biochem. Biophys.* **602**, 3–11, Jul (2016). [120](#)
- [285] Guo, Q., Huang, B., Cheng, J., Seefelder, M., Engler, T., Pfeifer, G., Oeckl, P., Otto, M., Moser, F., Maurer, M., et al. The cryo-electron microscopy structure of huntingtin. *Nature* **555**(7694), 117 (2018). [121](#)
- [286] Piovesan, D., Walsh, I., Minervini, G., and Tosatto, S. C. E. FIELDS: fast estimator of latent local structure. *Bioinformatics* **33**(12), 1889–1891, Jun (2017). [121](#)
- [287] Walsh, I., Seno, F., Tosatto, S. C., and Trovato, A. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* **42**(Web Server issue), W301–307, Jul (2014). [121](#)
- [288] Palidwor, G. A., Shcherbinin, S., Huska, M. R., Rasko, T., Stelzl, U., Arumughan, A., Foulle, R., Porras, P., Sanchez-Pulido, L., Wanker, E. E., and Andrade-Navarro, M. A. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput. Biol.* **5**(3), e1000304, Mar (2009). [121](#)
- [289] Jorda, J. and Kajava, A. V. Protein homorepeats sequences, structures, evolution, and functions. *Adv Protein Chem Struct Biol* **79**, 59–88 (2010). [122](#)
- [290] Kim, M. W., Chelliah, Y., Kim, S. W., Otwinowski, Z., and Bezprozvanny, I. Secondary structure of Huntingtin amino-terminal region. *Structure* **17**(9), 1205–1212, Sep (2009). [122](#)
- [291] Zhemkov, V. A., Kulminskaya, A. A., Bezprozvanny, I. B., and Kim, M. The 2.2-Angstrom resolution crystal structure of the carboxy-terminal region of ataxin-3. *FEBS Open Bio* **6**(3), 168–178, 03 (2016). [122](#)
- [292] Bennett, M. J., Huey-Tubman, K. E., Herr, A. B., West, A. P., Ross, S. A., and Bjorkman, P. J. A linear lattice model for polyglutamine in CAG-expansion diseases. *Proc. Natl. Acad. Sci. U.S.A.* **99**(18), 11634–11639, Sep (2002). [122](#)
- [293] Li, P., Huey-Tubman, K. E., Gao, T., Li, X., West, A. P., Bennett, M. J., and Bjorkman, P. J. The structure of a polyQ-anti-polyQ complex reveals binding according to a linear lattice model. *Nat. Struct. Mol. Biol.* **14**(5), 381–387, May (2007). [122](#)
- [294] Baias, M., Smith, P. E., Shen, K., Joachimiak, L. A., Zerko, S., Kozminski, W., Frydman, J., and Frydman, L. Structure and Dynamics of the Huntingtin Exon-1 N-Terminus: A Solution NMR Perspective. *J. Am. Chem. Soc.* **139**(3), 1168–1176, 01 (2017). [122](#)
- [295] Urbanek, A., Morato, A., Allemand, F., Delaforge, E., Fournet, A., Popovic, M., Delbecq, S., Sibille, N., and Bernado, P. A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. *Angew. Chem. Int. Ed. Engl.* **57**(14), 3598–3601, Mar (2018). [122](#)
- [296] Eftekharzadeh, B., Piai, A., Chiesa, G., Mungianu, D., Garcia, J., Pierattelli, R., Felli, I. C., and Salvatella, X. Sequence Context Influences the Structure and Aggregation Behavior of a PolyQ Tract. *Biophys. J.* **110**(11), 2361–2366, Jun (2016). [122](#)
- [297] Masino, L., Kelly, G., Leonard, K., Trottier, Y., and Pastore, A. Solution structure of polyglutamine tracts in GST-polyglutamine fusion proteins. *FEBS Lett.* **513**(2-3), 267–272, Feb (2002). [122](#)
- [298] Totzeck, F., Andrade-Navarro, M. A., and Mier, P. The Protein Structure Context of PolyQ Regions. *PLoS ONE* **12**(1), e0170801 (2017). [122](#)
- [299] Mier, P. and Andrade-Navarro, M. A. dAPE: a web server to detect homorepeats and follow their evolution. *Bioinformatics* **33**(8), 1221–1223, 04 (2017). [123](#)
- [300] Ahmed, A. B., Znassi, N., Chateau, M. T., and Kajava, A. V. A structure-based approach to predict predisposition to amyloidosis. *Alzheimers Dement* **11**(6), 681–690, Jun (2015). [123](#)
- [301] Fan, H. C., Ho, L. I., Chi, C. S., Chen, S. J., Peng, G. S., Chan, T. M., Lin, S. Z., and Harn, H. J. Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant* **23**(4-5), 441–458 (2014). [123](#)
- [302] Spink, B. J., Sivaramakrishnan, S., Lipfert, J., Doniach, S., and Spudich, J. A. Long single alpha-helical tail domains bridge the gap between structure and function of myosin VI. *Nat. Struct. Mol. Biol.* **15**(6), 591–597, Jun (2008). [123](#), [125](#)
- [303] Suveges, D., Gaspari, Z., Toth, G., and Nyitray, L. Charged single alpha-helix: a versatile protein structural motif. *Proteins* **74**(4), 905–916, Mar (2009). [123](#)
- [304] Dobson, L., Nyitray, L., and Gaspari, Z. A conserved charged single -helix with a putative steric role in paraspeckle formation. *RNA* **21**(12), 2023–2029, Dec (2015). [123](#)
- [305] Simm, D. and Kollmar, M. Waggawagga-CLI: A command-line tool for predicting stable single -helices (SAH-domains), and the SAH-domain distribution across eukaryotes. *PLoS ONE* **13**(2), e0191924 (2018). [123](#)
- [306] Dudola, D., Toth, G., Nyitray, L., and Gaspari, Z. Consensus Prediction of Charged Single Alpha-Helices with CSAHserver. *Methods Mol. Biol.* **1484**, 25–34 (2017). [123](#), [125](#)
- [307] Martinez, S. R. and Miranda, J. L. CTCF terminal segments are unstructured. *Protein Sci.* **19**(5), 1110–1116, May (2010). [125](#)
- [308] Gaspari, Z., Suveges, D., Perczel, A., Nyitray, L., and Toth, G. Charged single alpha-helices in proteomes revealed by a consensus prediction approach. *Biochim. Biophys. Acta* **1824**(4), 637–646, Apr (2012). [125](#)

REFERENCES

- [309] Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., and Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**(3), 573–584, Oct (2002). [125](#)
- [310] Szappanos, B., Suveges, D., Nyitray, L., Perczel, A., and Gaspari, Z. Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils. *FEBS Lett.* **584**(8), 1623–1627, Apr (2010). [125](#)
- [311] Gaspari, Z. Is Five Percent Too Small? Analysis of the Overlaps between Disorder, Coiled Coil and Collagen Predictions in Complete Proteomes. *Proteomes* **2**(1), 72–83, Feb (2014). [125](#)
- [312] Smithers, B., Oates, M. E., Tompa, P., and Gough, J. Three reasons protein disorder analysis makes more sense in the light of collagen. *Protein Sci.* **25**(5), 1030–1036, May (2016). [125](#)
- [313] Bosshard, H. R., Durr, E., Hitz, T., and Jelesarov, I. Energetics of coiled coil folding: the nature of the transition states. *Biochemistry* **40**(12), 3544–3552, Mar (2001). [125](#)
- [314] Bachmann, A., Kiefhaber, T., Boudko, S., Engel, J., and Bachinger, H. P. Collagen triple-helix formation in all-trans chains proceeds by a nucleation/growth mechanism with a purely entropic barrier. *Proc. Natl. Acad. Sci. U.S.A.* **102**(39), 13897–13902, Sep (2005). [125](#)
- [315] Obradović, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61** Suppl 7, 176–182 (2005). [126](#)
- [316] Lupas, A., Van Dyke, M., and Stock, J. Predicting coiled coils from protein sequences. *Science* **252**(5009), 1162–1164, May (1991). [126](#)
- [317] McDonnell, A. V., Jiang, T., Keating, A. E., and Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**(3), 356–358, Feb (2006). [126](#)
- [318] Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., Bateman, A., and Eddy, S. R. HMMER web server: 2015 update. *Nucleic Acids Res.* **43**(W1), W30–38, Jul (2015). [126](#)
- [319] Meszaros, B., Simon, I., and Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **5**(5), e1000376, May (2009). [127](#)
- [320] Sormanni, P., Aprile, F. A., and Vendruscolo, M. The camsol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology* **427**(2), 478490 (2015). [129](#)
- [321] Magnan, C. N., Randall, A., and Baldi, P. Solpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**(17), 2200–2207 (2009). [129](#)
- [322] Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. Proso ii—a new method for protein solubility prediction. *The FEBS journal* **279**(12), 2192–2200 (2012). [129](#)
- [323] Teplyakov, A., Obmolova, G., Malia, T. J., Luo, J., Muzammil, S., Sweet, R., Almagro, J. C., and Gilliland, G. L. Structural diversity in a human antibody germline library. *mAbs* **8**(6), 1045–1063 (2016). [130](#)
- [324] Ingram, V. M. et al. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* **180**(4581), 326–328 (1957). [132](#)
- [325] Luzzatto, L. Sickle cell anaemia and malaria. *Mediterranean journal of hematology and infectious diseases* **4**(1) (2012). [132](#)
- [326] Diggs, L. and Walker, R. A solubility test for sickle cell hemoglobin: I. aggregation and separation of soluble and insoluble components without centrifugation. *Laboratory Medicine* **4**(10), 27–31 (1973). [133](#)
- [327] Monplaisir, N., Merault, G., Poyart, C., Rhoda, M.-D., Craescu, C., Vidaud, M., Galacteros, F., Blouquit, Y., and Rosa, J. Hemoglobin s antilles: a variant with lower solubility than hemoglobin s and producing sickle cell disease in heterozygotes. *Proceedings of the National Academy of Sciences* **83**(24), 9363–9367 (1986). [133](#)
- [328] Pellegrini, M. Tandem repeats in proteins: prediction algorithms and biological role. *Frontiers in bioengineering and biotechnology* **3**, 143 (2015). [136](#)
- [329] Schaper, E., Gascuel, O., and Anisimova, M. Deep conservation of human protein tandem repeats within the eukaryotes. *Molecular Biology and Evolution* **31**(5), 1132–1148 (2014). [137](#)
- [330] Tompa, P. Intrinsically disordered proteins: a 10-year recap. *Trends in biochemical sciences* **37**(12), 509–516 (2012). [139](#)
- [331] Salemi, S., Markovic, M., Martini, G., and D’Amelio, R. The expanding role of therapeutic antibodies. *International reviews of immunology* **34**(3), 202–264 (2015). [141](#)