UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN: BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO: BIOCHIMICA E BIOFISICA
CICLO: XXVI

COMPUTATIONAL ANALYSIS AND ANNOTATION OF PROTEOME DATA: SEQUENCE, STRUCTURE, FUNCTION AND INTERACTIONS

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti
Coordinator d'indirizzo: Ch.mo Prof. Fabio Di Lisa
Supervisore: Ch.mo Prof. Silvio C.E. Tosatto

Dottorando: Tomás Di Domenico

A Constanza, por acompañarme en esta auténtica aventura transoceánica.

Sommario

Con l'avvento delle tecnologie di sequenziamento moderne, la quantità di dati biologici disponibili ha cominciato a sfidare la nostra capacità di elaborarli. È diventato quindi essenziale sviluppare nuovi strumenti e tecniche capaci di produrre dei risultati basati su grandi moli di informazioni. Questa tesi si concentra sullo sviluppo di tali strumenti computazionali e dei metodi per lo studio dei dati proteici.

Viene dapprima presento il lavoro svolto per la comprensione delle proteine intrinsecamente disordinate. Attraverso lo sviluppo di nuovi predittori di disordine, siamo stati in grado di sfruttare le fonti di dati attualmente disponibili per annotare qualsiasi proteina avente sequenza nota. Memorizzando queste predizioni, insieme ai dati provenienti da altre fonti, è stato creato MobiDB. Questa risorsa fornisce una visione completa sulle annotazioni di disordine disponibili per una qualsiasi proteina di interesse presente nel database UniProt. Sulla base delle osservazioni ottenute da questo strumento, è stato quindi creato un workflow di analisi dei dati con l'obiettivo di approfondire la nostra comprensione delle proteine intrinsecamente disordinate.

La seconda parte della tesi si concentra sulle proteine ripetute. Il metodo RAPHAEL è stato sviluppato per contribuire nell'identificazione di strutture proteiche ripetute all'interno dei file PDB. Le strutture selezionate da questo strumento sono state poi catalogate manualmente utilizzando uno schema formale di classificazione, e pubblicate quindi come parte del database RepeatsDB.

Infine, viene descritto lo sviluppo di strumenti basati su grafi per l'analisi di dati proteici. RING consente all'utente di visualizzare e studiare la struttura di una proteina come una rete di nodi collegati da tra loro da proprietà fisicochimiche. Il secondo metodo, PANADA, consente all'utente di creare reti di similarità di proteine e di valutare la trasferibilità delle annotazioni funzionali tra cluster diversi.

Abstract

With the advent of modern sequencing technologies, the amount of biological data available has begun to challenge our ability to process it. The development of new tools and methods has become essential for the production of results based on such a vast amount of information. This thesis focuses on the development of such computational tools and method for the study of protein data.

I first present the work done towards the understanding of intrinsic protein disorder. Through the development of novel disorder predictors, we were able to expand the available data sources to cover any protein of known sequence. By storing these predicted annotations, together with data from other sources, we created MobiDB, a resource that provides a comprehensive view of available disorder annotations for a protein of interest, covering all sequences in the UniProt database. Based on observations obtained from this resource, we proceeded to create a data analysis workflow with the goal of furthering our understanding of intrinsic protein disorder.

The second part focuses on tandem repeat proteins. The RAPHAEL method was developed to assist in the identification of tandem repeat protein structures from PDB files. Identified repeat structures were then manually classified into a formal classification schema, and published as part of the RepeatsDB database.

Finally, I describe the development of network-based tools for the analysis of protein data. RING allows the user to visualise and study the structure of a protein as a network of nodes, linked by physico-chemical properties. The second method, PANADA, enables the user to create protein similarity networks and to assess the transferability of functional annotations between clusters of proteins.

Acknowledgements

I would like to thank my advisor Prof. Silvio Tosatto for his guidance and support. I highly appreciate the friendly work environment under which this thesis was developed. I also thank him for the possibility and encouragement to expand my scientific network of contacts by participating in many high-profile conferences, and by getting involved in the activities of scientific societies and groups.

Thanks to all the members of the BioComputing group with whom I've shared these past years. In particular, I would like to thank to Manuel Giollo for many interesting and mind-opening discussions.

Special thanks to Dr. Maria Assunta Piano, Dr. Alberto Martin, and to the great Dr. Ian Walsh for their friendship, support and advise.

I would also like to thank Dr. Cristina Marino Buslje for giving me the opportunity to first come in contact with bioinformatics and thus opening the doors of my current career, and to Prof. Gustavo Parisi for his advise and support.

And finally, I would like to give very special thanks to my family. Their unconditional support in whatever enterprises I decided to pursue throughout my life, no matter how crazy they would seem, was essential in my path to discover my passion for research.

List of original publications

1. Walsh I, **Di Domenico T**, Tosatto SCE (2013) RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. Amino Acids: 1–10. doi:10.1007/s00726-013-1645-3.

2. **Di Domenico T**, Potenza E, Walsh I, Gonzalo Parra R, Giollo M, et al. (2013) RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res.

3. Martin AJM, Walsh I, **Di Domenico T**, Mičetić I, Tosatto SCE (2013) PANADA: Protein Association Network Annotation, Determination and Analysis. PLoS ONE 8: e78383.

4. Bateman A, Kelso J, Mietchen D, Macintyre G, **Di Domenico T**, et al. (2013) ISCB Computational Biology Wikipedia Competition. PLoS Comput Biol 9: e1003242.

5. **Di Domenico T**, Walsh I, Tosatto SC (2013) Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. BMC Bioinformatics 14: S3.

6. Walsh I, Sirocco FG, Minervini G, **Di Domenico T**, Ferrari C, et al. (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. Bioinformatics 28: 3257–3264.

7. **Di Domenico T**, Walsh I, Martin AJM, Tosatto SCE (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28: 2080–2081.

8. Walsh I, Martin AJM, **Di Domenico T**, Tosatto SCE (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28: 503–509.

9. Martin AJM, Vidotto M, Boscariol F, **Di Domenico T**, Walsh I, et al. (2011) RING: networking interacting residues, evolutionary information and energetics in protein structures. Bioinformatics 27: 2003–2005.

10. Walsh I, Martin AJM, **Di Domenico T**, Vullo A, Pollastri G, et al. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Research 39: W190–W196.

Contents

List of Figures

List of Tables

Part I

Introduction

1. Overview

During the last two decades, with the advent of new sequencing technologies, the biological sciences have joined other fields in generating massive amounts of data at ever increasing rates [1–3]. While this *data deluge* offers outstanding opportunities for novel research efforts and the advancement of the field, it also represents a great challenge in terms of our ability to store and analyse the generated data. If we add to the mix the tendency of data availability to decline as time progresses [4], it is hard to ignore the fact that we are currently faced with a truly complex and multifaceted issue.

Bioinformatics is a field that combines elements from biology, computer science, engineering, mathematics, and statistics (among others), to tackle biological problems. It involves the development of tools and methods to aid on the storage, processing and analysis of biological data. Even though the origins of bioinformatics can be traced as far back as the 1970s [5], it is undeniable that the field has experienced an exponential growth that is coupled to the emergence of the aforementioned data deluge.

Owing its heterogeneous nature to its relationship to a field as varied as biology, the field of Bioinformatics deals with a wide range of topics. Three main groups of datasets have been proposed as being the main focus of the field: macromolecular structures, genome sequences, and results of functional genomics experiments [6]. This is of course a rather *coarse grained* definition, as the complexity of biological data defies the possibility of achieving a unique an precise categorisation. Figure 1.1 offers a schematic view of the breadth and depth of bioinformatics, featuring examples of the types of data that can be of interest for bioinformatics analysis, and of the methods that can be applied to such data.

1.1 Outline

This thesis is divided into five parts, themselves subdivided into chapters. Part I consists of an introduction and a general overview of the topics that are present in the different sections. Part II describes the work done on the analysis of intrinsic protein disorder. Part III presents results obtained from the study of tandem repeat proteins. Part IV describes the use of networks for the analysis of protein data. To conclude, Part V presents conclusions and outlook of the obtained results.

It should be noted that most chapters included in parts II, III and IV have been previously published in peer-reviewed journals, as specified at on their corresponding headings. It is also worth noting that I have used both “I” and “We”

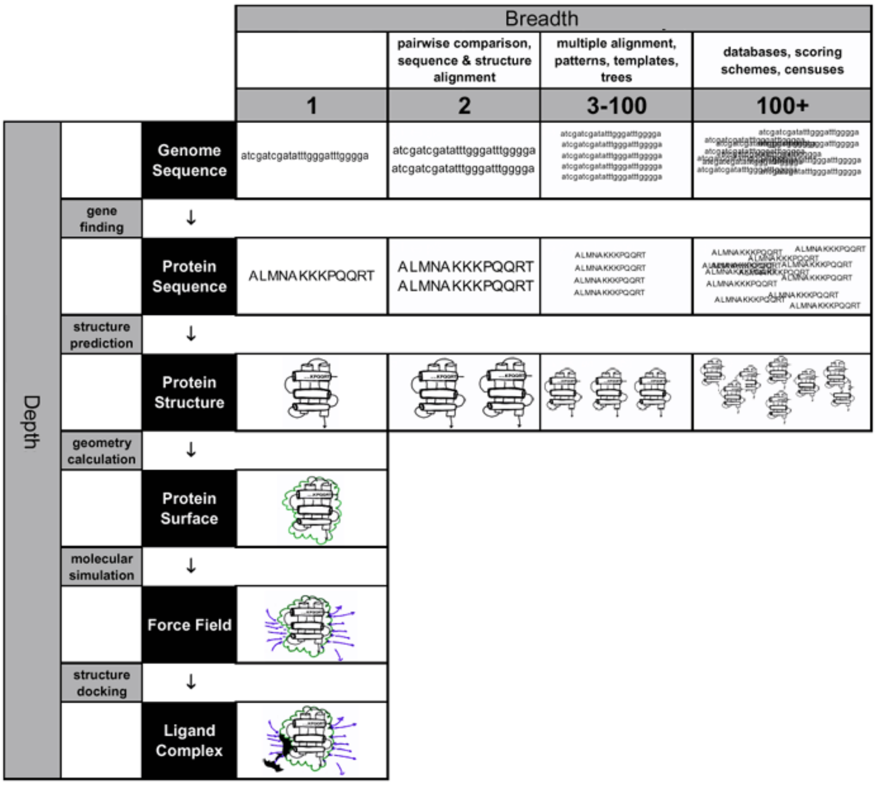


Figure 1.1: The Bioinformatics Spectrum. A summarised, schematic view of the breadth and depth of bioinformatics [6, 7].

when describing the work done in different sections. This has the double objective of distinguishing the sections where work was done in collaboration with other researchers, and to differentiate my personal opinions from those of the group where the work was done.

1.2 Computational methods and tools for the analysis and annotation of proteins

The conversion of raw data into useful information, which will allow a researcher to arrive to valuable conclusions, is far from being a trivial task. Adding the fact that in the biological sciences this data has overwhelming volumes, it becomes clear how researchers face a formidable challenge.

From a data-centric point of view, research can be though of as the execution of four main, intertwined steps: the generation of raw data, the conversion of raw data into processed data, the storage (and publication) of data, and the analysis of data. A research project may require anything from a single one to all of these steps, depending on previous availability, data complexity, research goals, etc. With the exception of the generation of raw data, all of the previously enumerated steps were covered during the development of this thesis.

The following overall description will provide some general insight on the topics of research that constitute the core of the research done throughout this thesis, and that will be described in detail in the upcoming chapters.

1.2.1 Intrinsic protein disorder

During the last few years, the traditional "lock-and-key" paradigm of structural biology, where proteins interact with each other by the complementary nature of their well-defined structures, has been increasingly challenged. It is now generally accepted that there exists a group of proteins which can lack (in whole or in part) stable three-dimensional structure under certain conditions. These proteins are known as *intrinsically disordered proteins*[1].

Chapter 2 and Chapter 3 describe newly developed protein disorder predictors which, at the time of writing, constitute the state-of-the-art in the field. These predictors leverage machine learning methods to generate high-quality predictions for intrinsic protein disorder. Both predictors are based on bi-directional recursive neural networks (BRNNs). Unlike traditional neural networks, BRNNs do not learn the context of the sequence by applying a "sliding window" (i.e. a fixed-length piece of sequence). Instead, they extract information implicitly through the recursive dynamics of the network.

Building on the results obtained from these predictors, Chapter 4 describes MobiDB, a database for intrinsic disorder annotation. Currently in its second major release, MobiDB uses state-of-the-art technology and a carefully developed modular design to provide a fast and comprehensive resource to the public. Its goal is to provide the best possible intrinsic protein disorder annotation for each of its entries, and it achieves this by aggregating information from different sources and by combining this information in a clever way. Covering the full set of proteins available in the UniProt knowledge base (roughly 50 million at the time of writing), it is the largest available repository of intrinsic protein disorder annotations.

Chapter 5 provides a more detailed description of the annotations available in the original release of MobiDB. The analysis of agreement and disagreement between different sources motivated many of the improvements that were later introduced into the second major release of the database[2].

Based on the content of MobiDB, Chapter 6 presents an attempt to critically analyse and expand our knowledge of intrinsic protein disorder. By leveraging the tools and methods described in previous chapters as well as external resources, we study the similarities and differences between different disorder annotation sources to study a particular type of disorder: folding upon binding. Proteins that undergo folding upon binding are known to be disordered until binding a partner. In addition to detecting such cases, we were able to develop an initial classification schema for subtypes of intrinsic disorder. This schema should prove useful for the further refinement of tools that deal with intrinsic protein disorder.

[1] For consistency, I have chosen to use the term *intrinsically disordered protein* throughout this thesis. The variety of names under which the phenomenon can be found in literature evidences how even reaching an agreement on a naming convention is far from a simple task [8].

[2] Given that the second release of MobiDB was finished shortly before the time of writing, this thesis features a description and screen captures of the previous, original release. Chapter 6 and the final conclusions presented in Chapter 12, however, are based on the datasets that were eventually included in the new release

Post-translational modifications have often been associated with intrinsically disordered regions. Chapter 7, perhaps the least organic to this second part of the thesis, makes use of disorder annotations in an attempt to improve the accuracy when detecting ubiquitination sites.

1.2.2 Tandem Repeat Proteins

Like intrinsically disordered proteins, tandem repeat proteins fall into the category of "non-globular proteins": proteins that instead of having a globular-shaped structure tend to be rather elongated. Unlike intrinsically disordered proteins, however, tandem repeat proteins have a well defined structure. Their elongation is due to their structure being formed by adjacent structural subunits highly similar to each other. In most cases of tandem repeats, these subunits depend on one another to fold.

Following a logic similar to that used for the study of intrinsic disorder, Chapter 8 presents RAPHAEL, a tool for detecting tandem repeats from structure form the Protein Data Bank. Originally designed for the detection of solenoid repeats, RAPHAEL is actually capable of detecting a wide range of tandem repeat classes. By manually curating the set of candidates obtained from running RAPHAEL on the full Protein Data Bank, we generated a highly detailed dataset of repeat protein annotations.

The results of the previously described annotation effort were made available through RepeatsDB. Described on Chapter 9, RepeatsDB presents a classification of repeat structures into an organised schema. Entries contained in RepeatsDB present different levels of annotations, ranging from being automatically detected as repeated, to highly detailed manual annotations that include start and end positions for each repeat unit.

1.2.3 Network-based tools for the analysis of protein data

More user-oriented than the previously described ones, this part presents tools aimed at the end user. Chapter 10 describes RING, a tool for the construction and analysis of residue interaction networks (RINs). RING allows the user to construct physico-chemically valid RINs interactively, and to analyse them using the Cytoscape software. PANADA, described in Chapter 11, is a tool for the visualisation of protein similarity networks. The tool permits the user to visualise clusters of related proteins, and to assess the transferability of functional annotations.

Part II

Intrinsic Protein Disorder

2. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs

This chapter was first published in Walsh I, Martin AJM, Domenico TD, Vullo A, Pollastri G, et al. (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucl Acids Res 39: W190–W196.

2.1 Summary

CSpritz is a web server for the prediction of intrinsic protein disorder. It is a combination of previous Spritz with two novel orthogonal systems developed by our group (Punch and ESpritz). Punch is based on sequence and structural templates trained with support vector machines. ESpritz is an efficient single sequence method based on bidirectional recursive neural networks. Spritz was extended to filter predictions based on structural homologues. After extensive testing, predictions are combined by averaging their probabilities. The CSpritz website can elaborate single or multiple predictions for either short or long disorder. The server provides a global output page, for download and simultaneous statistics of all predictions. Links are provided to each individual protein where the amino acid sequence and disorder prediction are displayed along with statistics for the individual protein. As a novel feature, CSpritz provides information about structural homologues as well as secondary structure and short functional linear motifs in each disordered segment. Benchmarking was performed on the very recent CASP9 data, where CSpritz would have ranked consistently well with a Sw measure of 49.27 and AUC of 0.828. The server, together with help and methods pages including examples, are freely available at URL: http://protein.bio.unipd.it/cspritz/.

2.2 Introduction

The 3D native structure of proteins has been considered the major determinant of function for many years. Over the last decade there has been a growing realization of an alternative mechanism whereby non-folding regions are both widespread and also carry functional significance [9, 10]. These non-folding regions within a protein, coming in various guises ranging from fully extended to molten globule-like and partially folded structures [11], are collectively known as intrinsically disordered regions [12]. Such regions often become structured upon binding to a target molecule and have been shown to be involved in various biological processes such as cell signaling or regulation [13], DNA binding [14] and molecular recognition in general [11, 15]. An interesting observation is that the amount of disorder within a proteome seems to correlate with complexity of the organism, with an apparent increase in disorder for eukaryotic organisms [16, 17]. The conservation of disorder [18, 19] and specific amino acid patterns [20, 21] (e.g. PxPxP) have also been studied. Indeed, there is a growing realization that intrinsically disordered regions are widely used as hubs for protein–protein interactions [22], for which structural data can be accessed in the ComSin database [21]. Functional linear motifs [23, 24], which are mostly hidden in disordered regions [25], have been characterized in resources such as ELM [26], an online repository of linear motifs.

The experimental determination of native disorder, once considered an anomaly, can be time consuming, difficult and expensive. As a result, computational approaches have largely driven our understanding of disorder over the last decade [22]. The bi-yearly Critical Assessment of Techniques for protein Structure Prediction (CASP) experiment has included a disorder category since CASP5 in 2002 [27]. Previously published methods can be roughly divided into biophysical and machine learning approaches. The former rely on the unique amino acid distribution associated with protein disorder [28–30]. Machine learning methods use either neural networks [31–33] or support vector machines [17, 34] and are commonly based on sequence profiles, predicted secondary structure and more recently template structures [35]. More recently, meta servers combining several biophysical and machine learning methods have been published [36–38]. All these methods have shown promising results, possibly for two reasons: (i) as the amino acid sequence contains all the information to determine structure it is reasonable to assume that unstructured regions have specific amino acid propensities and (ii) disorder is important in many biological functions and therefore unstructured protein segments should be conserved by evolution. Knowing that disordered segments have a biased sequence, machine learning techniques should excel. In this paper we describe and benchmark CSpritz, an extension of our previous Spritz server [34] based on three distinct modules for the prediction of intrinsically disordered regions in proteins. The performance of the method will be benchmarked on the latest available data for short and long disordered segments. A novel addition to the CSpritz server is information about homologous structures found from PSI-BLAST searches, secondary structure and linear motifs contributing to the functional annotation of disordered segments.

2.3 Materials and methods

CSpritz predicts intrinsic disorder from protein sequences through a combination of three machine learning systems, which will be described in the following sections. Most methods consider short and long disorder separately, as they have different characteristics. Short disorder can be derived from residues missing backbone atoms in X-ray crystallographic structures deposited in the Protein Data Bank (PDB) [39]. Long disorder is taken from the Disprot database [40] because it is largely missing from the PDB. All data sets used throughout training are appropriately redundancy reduced using UniqueProt [41] and in all cases contain only sequences available before May 2008 (i.e. the start of CASP8).

2.3.1 Spritz

The original Spritz [34] is based on PSI-BLAST [42] multiple sequence profiles and predicted secondary structure. Support Vector Machines (SVMs) were used on a local sequence window to train two specialized binary classifiers, for long and short regions of disorder. A description of the data sets can be found in the previous publication [34]. In addition to the original ab initio version of Spritz, a filter removing PDB structural homologues from predicted disorder is implemented. This works by performing a PSI-BLAST search against a redundancy reduced sequence database. The generated sequence profile is then used in a final PSI-BLAST round against a filtered PDB. Residues matching a structural template are assigned a Spritz score below the disorder threshold.

2.3.2 Punch

Punch is a SVM based predictor extending Spritz. Sequence and structural homologues are detected as in Spritz. In addition, Porter secondary structure [43] and PaleAle relative solvent accessibility [44] are also included. Unlike Spritz, information about structural templates is encoded and fed directly to the SVM together with the other inputs. The two data sets used for learning (see A.1) are a large set of disordered X-ray chains derived from the PDB (December 2007) and a publicly available data set [31] based on disordered X-ray segments from the PDB (May 2004). The assignment of disorder is different in both data sets and does not necessarily intersect.

2.3.3 ESpritz

ESpritz is a fast predictor using bidirectional recursive neural networks (BRNNs) [45]. BRNNs do not require contextual windows because they extract this information dynamically from the sequence. ESpritz consists of 20 inputs where each unit is allocated for one of the 20 amino acids. Although the method is very simple, the BRNN is useful for extracting relevant patterns required for disorder without the use of PSI-BLAST sequence alignments (results not shown). Like Spritz, two types of data based on long and short disorder types are designed (see A.1). The short disorder set is built from X-ray PDB structures (May 2008). Long disorder segments are extracted from Disprot (version 3.7) with identical sequences removed.

2.3.4 Linear motifs and secondary structure

It can be useful to unify the following information for disordered segments: (i) amino acids involved; (ii) secondary structure; and (iii) important linear motifs. CSpritz offers this predicted information in various forms (see output section). Secondary structure propensities are predicted from Porter [43]. Linear motifs (LMs) are selected from ELM [26] as the ligand binding subset (names starting with LIG). ELM is a resource for predicting functional sites in eukaryotic proteins where functional sites are identified by patterns. These motifs are supposed to be representative of the more studied LM–protein binding examples. The selected LMs are returned when sub-sequences are matched by their regular expressions in ELM.

2.4 Performance evaluation

2.4.1 Combination

Experiments were carried out for the best procedure to combine Punch, Spritz and ESpritz. After trying majority voting, unanimous votes and combination with neural networks, the simplest method of averaging the probabilities produced by each system was found to be the best (data not shown). The optimal decision threshold was determined on data independent from the benchmarking set by maximizing the Sw measure [46]. CASP8 data [46] was used for short and Disprot (version 3.7) for long disorder. Regular expressions are incorporated to fill disordered regions separated by less than three residues. The Pearson correlation of the probabilities produced on CASP9 disorder targets was calculated to test how different the three predictors are. Table 2.1 shows this correlation and proves that the three systems are indeed sufficiently different. This is important for combining the three systems since it is well known that ensembling predictions which are different or uncorrelated improve generalization performance considerably [47]. In particular, combination is especially beneficial when the wrongly predicted residues for each predictor do not correlate (i.e. their probabilities do not correlate) [48, 49].

	ESpritz	Spritz	Punch
ESpritz	1.00	0.51	0.59
Spritz		1.00	0.42
Punch			1.00

Table 2.1: Pearson correlation of the three systems on CASP9 targets. The probabilities are produced by each component on all residues for 117 CASP9 targets. Since the correlations are low, combining the three systems improves performance over the individual systems.

2.4.2 Benchmarking sets

Validation of short disorder segments is performed on the 117 CASP9 targets[1], comparing with other groups taking part in the disorder category experiment according to their official CASP results. In order to validate the long disorder segments we choose DisProt entries enriched with PDB annotation from the SL data set defined in [49]. Unfortunately, selecting sequences with <40% sequence identity to our training set leaves only 29 proteins. We also define a set of 569 X-ray sequences (Xray569) deposited in the PDB (resolution at most 2.5 Å and R-free <0.25) between May 2008 and September 2010 reduced by sequence identity using UniqueProt [41] to an HSSP value of 0 to our training data and among each other. Supplementary Table A.1.1 shows the size and composition of the validation data sets. Note that to ensure a fair comparison to other methods on our benchmarking sets, CSpritz was in all cases run with sequence and PDB databases frozen prior to May 2008.

2.4.3 CASP short disorder

To assess the performance of our server for the short disorder option, we rank all groups participating in the CASP9 experiment. Table 2.2 shows the top 5 (out of 32) groups plus CSpritz and Spritz ranked by Sw, a commonly used measure at CASP. For Sw, as in the CASP8 assessment [46] the statistical significance of the evaluation scores was determined by bootstrapping: 80% of the targets were randomly selected 1000 times, and the standard error of the scores was calculated (i.e. 1.96 * standard error gives 95% confidence around mean for normal distributions). For a full list of rankings see the online methods page. Our results suggest a consistently good performance of our server, especially when taking into account that some of the top five are meta-servers and some are not publicly available.

GroupID: Name	Sw (±SE)	ACC	AUC
291: PRDOS2	50.44 (±1.08)	75.22	0.852
119: MULTICOM-REFINE	49.53 (±1.00)	74.77	0.818
000: CSpritz	49.27 (±1.02)	74.64	0.828
351: BIOMINE_DR_PDB	48.21 (±1.25)	74.11	0.818
374: GSMETADISORDERMD	47.13 (±0.96)	73.57	0.815
193: MASON	45.98 (±1.17)	73.00	0.740
000: Spritz	24.91 (±1.18)	62.46	0.716

Table 2.2: Results for the top five performing groups at the CASP9 experiment, CSpritz and the original Spritz. Disordered segments of less than three residues were removed (results unchanged if included, see Supplementary Table A.1.3). The standard error (SE) for Sw is shown in brackets. ACC is the accuracy, i.e. (sensitivity + specificity)/2, and AUC the area under the receiver operator curve. A total of 32 groups participated in CASP9 disorder prediction category.

[1] http://www.predictioncenter.org/casp9/

2.4.4 DisProt long disorder

The long disorder type performance of CSpritz was benchmarked by comparing Sw, accuracy and AUC with the original Spritz and state-of-the-art predictors PONDR-FIT [37], Disopred [17] and IUPred [30]. Table 2.3 shows CSpritz performing significantly better than the other predictors for this type of disorder. In addition CSpritz improves over the long disorder predictions made by our previous server Spritz.

Method	Sw (±SE)	ACC	AUC
CSpritz (short)	54.64 (±3.58)	77.32	0.837
CSpritz (long)	65.70 (±3.52)	82.85	0.891
Spritz (short)	12.12 (±6.16)	56.06	0.685
Spritz (long)	35.55 (±3.58)	67.78	0.734
PONDR-FIT	51.53 (±4.34)	75.77	0.817
Disopred2	46.20 (±4.00)	73.10	0.806
IUPred (short)	37.65 (±4.77)	68.83	0.814
IUPred (long)	42.57 (±4.75)	71.29	0.818

Table 2.3: Comparison for DisProt disordered regions. Spritz is compared with the original Spritz, PONDR-FIT, Disopred and IUPred. Where applicable both short and long options are reported. The standard error (SE) for Sw is shown in brackets. ACC is the accuracy, i.e. (sensitivity + specificity)/2, and AUC the area under the receiver operator curve. The decision threshold and best Sw was found to be 0.26 and 51.85 on the training set.

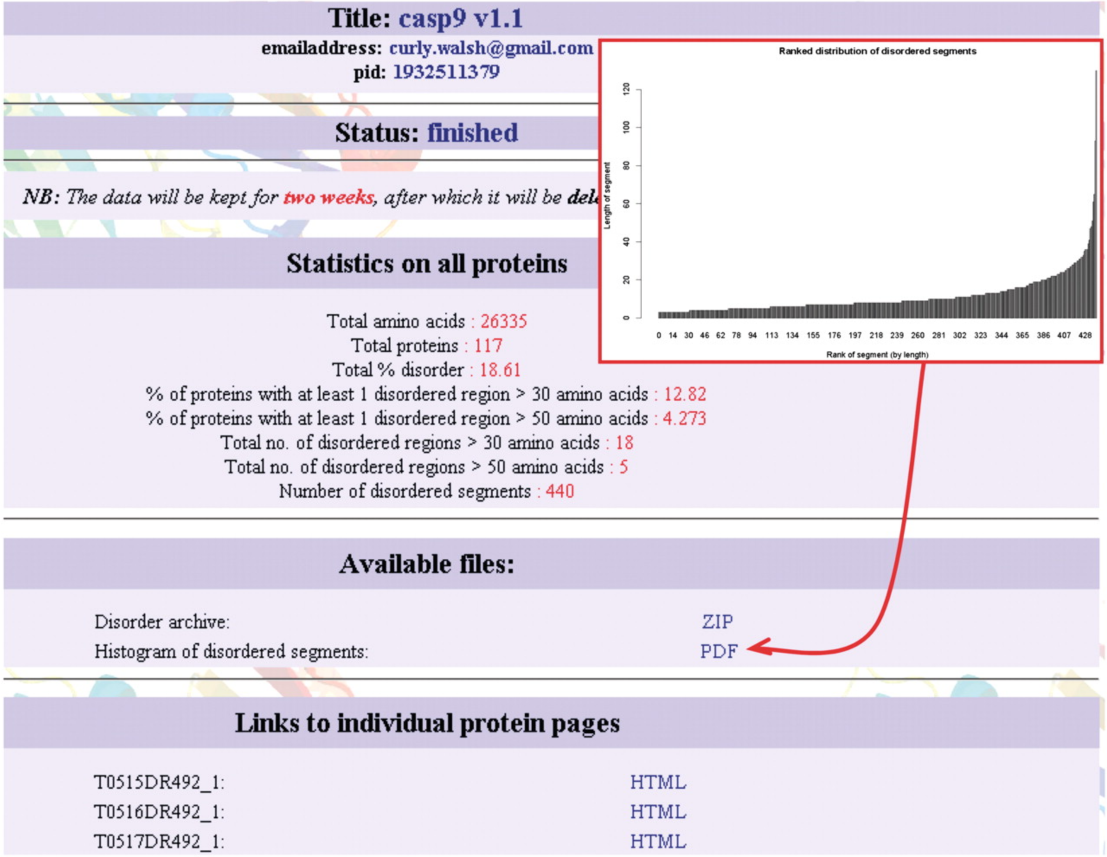
2.4.5 Large-scale performance

To estimate the run time of CSpritz compared to others and validate the predictions on a larger set of PDB structures we use the Xray569 set. The results (Supplementary Table A.1.2) are similar to the DisProt set and confirm the performance of CSpritz compared to the other methods. As can be expected, all methods are better at predicting disorder at the N- and C-termini than in the central part of the protein sequences. The execution time for CSpritz is largely determined by the PSI-BLAST search and comparable to the original Spritz and Disopred2, with ca. 15 min for an average protein. When executing multiple predictions, the CSpritz web server will run up to five proteins in parallel, reducing the overall time significantly.

2.5 Server description

The CSpritz input page is designed with simplicity in mind. A single or multiple sequences in FASTA format are the only input required and can be either pasted or uploaded as a file. Pasting is limited to 32000 characters but uploading has no restrictions. User email address and a query title are optional. Either short (default) or long disorder options can be selected, with the appropriate decision thresholds determined on data not involved in the benchmarking. To facilitate navigation, help and methods pages are available at the top of the interface.

The CSpritz output is presented in two main pages. The first page, displaying statistics, links to individual pages and a downloadable archive for all user supplied proteins, is present only if more than one sequence was submitted. A histogram of disordered segments and an archive for download containing all generated data are also available. Figure 2.1 shows a sample global page for the 117 CASP9 targets.



Title: casp9 v1.1

emailaddress: curly.walsh@gmail.com
pid: 1932511379

Status: finished

NB:** The data will be kept for **two weeks**, after which it will be **del

Statistics on all proteins

Total amino acids : 26335
Total proteins : 117
Total % disorder : 18.61
% of proteins with at least 1 disordered region > 30 amino acids : 12.82
% of proteins with at least 1 disordered region > 50 amino acids : 4.273
Total no. of disordered regions > 30 amino acids : 18
Total no. of disordered regions > 50 amino acids : 5
Number of disordered segments : 440

Available files:

Disorder archive: ZIP
Histogram of disordered segments: PDF

Links to individual protein pages

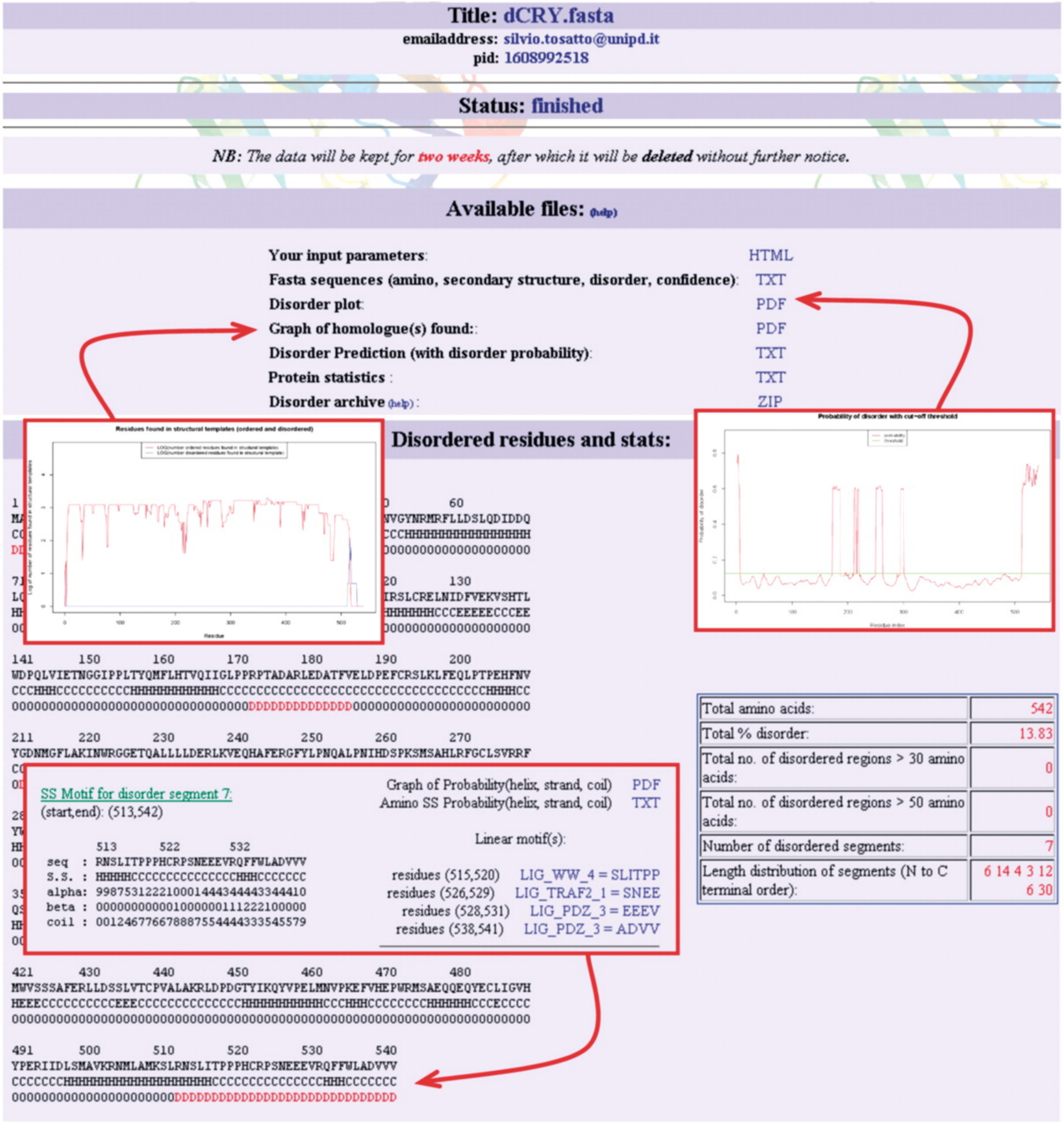
T0515DR492_1: HTML
T0516DR492_1: HTML
T0517DR492_1: HTML

Figure 2.1: Global output page for multiple sequences. Summary statistics are displayed for some interesting values about the disorder segments of all query sequences. An archive is offered for download containing all disorder predictions, linear motifs and statistics for each protein the user supplied. The inset shows a graph displaying the length distribution of disorder segments among all proteins.

The second output displays predicted disorder and annotation for individual proteins. In addition to showing the sequence with predicted secondary structure and disorder, several statistics regarding the distribution of disorder are presented. An extensive description of the output is available as part of the online help page. Two graphs plot the probability of disorder and the number of available structural templates versus disordered regions in homologous PDB structures. The last part of the output concerns the presence of putative linear motifs and secondary structure propensity for disordered segments. This can be a useful source of functional annotation, as shown in Figure 2.2 for Drosophila melanogaster Cryptochrome (dCRY). Following computational analysis, functional linear motifs were experimentally confirmed in the disordered C-terminus of dCRY [50]. CSpritz aims to speed up this type of analysis by providing additional clues. In dCRY the putative linear motifs (Figure 2.2) match the disordered residues having a favorable alpha helical propensity. It is known that many such interactions involve disorder to secondary structure transitions upon binding [51].

2.6 Conclusions

We have described CSpritz, a novel web server for the prediction of intrinsically disordered protein segments from sequence. It allows the batch prediction of many sequences simultaneously, providing overview statistics. The single protein sequence is annotated with disorder and useful information regarding local secondary structure and possible interaction motifs, providing a first step towards the functional interpretation of disorder. Future work will concentrate on improving the functional description of disordered regions by including other types of related information such as repeats [52] and aggregation [53].



Residues found in structural templates (ordered and disordered)
LOG(number ordered residues found in structural templates)
LOG(number disordered residues found in structural templates)
Residue
Probability of disorder with cut-off threshold
probability
threshold
Residue index

Title: dCRY.fasta

emailaddress: silvio.tosatto@unipd.it
pid: 1608992518

Status: finished

NB: *The data will be kept for* ***two weeks****, after which it will be* ***deleted*** *without further notice.*

Available files: (help)

Your input parameters:	HTML
Fasta sequences (amino, secondary structure, disorder, confidence):	TXT
Disorder plot:	PDF
Graph of homologue(s) found::	PDF
Disorder Prediction (with disorder probability):	TXT
Protein statistics :	TXT
Disorder archive (help) :	ZIP

Disordered residues and stats:

```
1                                                    0         60
MA                                                   NVGYNRMRFLLDSLQDIDDQ
CC                                                   CCHHHHHHHHHHHHHHHHHH
DD                                                   00000000000000000000

71                                                   20        130
LQ                                                   IRSLCRELNIDFVEKVSHTL
HH                                                   HHHHHHHCCCEEEEECCCEE
00                                                   00000000000000000000

141       150       160       170       180       190       200
WDPQLVIETNGGIPPLTYQMFLHTVQIIGLPPRPTADARLEDATFVELDPEFCRSLKLFEQLPTPEHFNV
CCCHHHCCCCCCCCCHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCC
0000000000000000000000000000000000DDDDDDDDDDDDDD0000000000000000000000

211       220       230       240       250       260       270
YGDNMGFLAKINWRGGETQALLLLDERLKVEQHAFERGFYLPNQALPNIHDSPKSMSAHLRFGCLSVRRF
CC
OD

28
YW
HH
00

35
QS
HH
00

421       430       440       450       460       470       480
MWVSSSAFERLLDSSLVTCPVALAKRLDPDGTYIKQYVPELMNVPKEFVHEPWRMSAEQQEQYECLIGVH
HEEECCCCCCCCCEEECCCCCCCCCCCCCHHHHHHHHHHHHCCCHHHCCCCCCCCHHHHHHCCCECCCC
0000000000000000000000000000000000000000000000000000000000000000000000

491       500       510       520       530       540
YPERIIDLSMAVKRNMLAMKSLRNSLITPPPHCRPSNEEEVRQFFWLADVVV
CCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCHHHCCCCCCC
0000000000000000000000DDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
```

SS Motif for disorder segment 7:
(start,end): (513,542)

```
        513       522        532
seq  : RNSLITPPPHCRPSNEEEVRQFFWLADVVV
S.S. : HHHHHCCCCCCCCCCCCCCHHHCCCCCCC
alpha: 998753122210001444344443344410
beta : 000000000001000000111222100000
coil : 001246776678887554444333545579
```

Graph of Probability(helix, strand, coil) PDF
Amino SS Probability(helix, strand, coil) TXT

Linear motif(s):

residues (515,520) LIG_WW_4 = SLITPP
residues (526,529) LIG_TRAF2_1 = SNEE
residues (528,531) LIG_PDZ_3 = EEEV
residues (538,541) LIG_PDZ_3 = ADVV

Total amino acids:	542
Total % disorder:	13.83
Total no. of disordered regions > 30 amino acids:	0
Total no. of disordered regions > 50 amino acids:	0
Number of disordered segments:	7
Length distribution of segments (N to C terminal order):	6 14 4 3 12 6 30

Figure 2.2: Individual output page for D. melanogaster Cryptochrome. The main figure shows the list of available files and actual disorder prediction. The latter is composed of the amino acid sequence, its predicted secondary structure and the CSpritz disorder classification, with disordered residues highlighted in red font. Disorder statistics about the protein is presented on the right. Two insets show the graphs for the disorder propensity plot (top right) and number of available structural coordinates versus disordered segments in homologous sequences. The inset on the bottom part shows the annotated disordered segment covering the C-terminus of Cryptochrome (residues 513–542). The propensities for secondary structure and location of putative functional motifs are shown. Links to the ELM description of the motif amino acids involved in the motif are supplied on the right. A graph and probabilities secondary structure propensity are also supplied.

3. ESpritz: accurate and fast prediction of protein disorder

This chapter was first published in Walsh I, Martin AJM, Di Domenico T, Tosatto SCE (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28: 503–509.

3.1 Summary

Intrinsically disordered regions are key for the function of numerous proteins, and the scant available experimental annotations suggest the existence of different disorder flavors. While efficient predictions are required to annotate entire genomes, most existing methods require sequence profiles for disorder prediction, making them cumbersome for high-throughput applications.

In this work, we present an ensemble of protein disorder predictors called ESpritz. These are based on bidirectional recursive neural networks and trained on three different flavors of disorder, including a novel NMR flexibility predictor. ESpritz can produce fast and accurate sequence-only predictions, annotating entire genomes in the order of hours on a single processor core. Alternatively, a slower but slightly more accurate ESpritz variant using sequence profiles can be used for applications requiring maximum performance. Two levels of prediction confidence allow either to maximize reasonable disorder detection or to limit expected false positives to 5%. ESpritz performs consistently well on the recent CASP9 data, reaching a Sw measure of 54.82 and area under the receiver operator curve of 0.856. The fast predictor is four orders of magnitude faster and remains better than most publicly available CASP9 methods, making it ideal for genomic scale predictions.

ESpritz predicts three flavors of disorder at two distinct false positive rates, either with a fast or slower and slightly more accurate approach. Given its state-of-the-art performance, it can be especially useful for high-throughput applications.

3.2 Introduction

Protein function has been traditionally thought to be determined by tertiary structure. More recently, an alternative view is emerging with respect to non-folding regions, which suggests a reassessment of the structure-to-function paradigm [9, 54–56]. Flexible segments lacking a unique native structure within a protein are known as disordered regions [12]. Disorder has been shown to be widespread within known natural proteins, especially in eukaryotic organisms [16, 55, 56]. It also plays a key role in human disease [57] where it is thought that 79% of all cancer-associated proteins are at least in part unstructured/disordered [55]. Proteins with disordered segments are frequently associated with molecular recognition [11, 15]. They have also been observed to be common among hub proteins, i.e. those with a large number of interaction partners [58]. In addition, protein disorder is also important for protein expression, purification and crystallization since difficulties often arise when long disordered regions are present, as happens frequently at the N and C termini.

Protein disorder is experimentally determined with an assortment of indirect biochemical methods collected in the DisProt database [40] currently containing ∼640 proteins. Alternatively, missing residues in X-ray crystallographic structures from ∼70 000 structures deposited in the Protein Data Bank (PDB) [39] can be used. The analysis of ∼6000 nuclear magnetic resonance (NMR) ensembles from the PDB is also possible [59], but to the best of our knowledge, has never been used to train a prediction method. It is assumed that different flavors of protein disorder exist [55]. The most common distinction is between long (DisProt) and short (X-ray) segments [56]. Alternatively, there has also been an attempt to distinguish flavors based on enrichment for certain amino acid types [60]. The characteristically skewed amino acid distribution of disordered segments, lacking in hydrophobic and enriched in polar and charged residues [61], can be easily exploited for sequence-based predictions. Available prediction methods can be broadly divided into three classes. Biophysical methods [28–30, 62–64] exploit the sequence distribution to derive pseudo-energy propensities to adopt a disordered state. Of these, IUPred [30] is probably the most widely used due to its availability and efficiency, as it does not require multiple sequence alignments. Machine learning techniques have been widely used for the prediction of protein disorder [17, 31, 33–35, 65–67]. In most cases, PSI-BLAST sequence profiles [42] are combined with additional features, e.g. predicted secondary structure in the widely used Disopred [17]. On average, these methods are slower but somewhat more accurate than biophysical predictors. The last, and most recent, category of disorder predictors use a consensus of various biophysical and machine learning methods [36–38, 68]. Here, a further improvement in accuracy is obtained at the cost of running several predictors in parallel and averaging their output.

Among the applications of disorder prediction, we can distinguish at least two different scenarios. The first is represented by the Critical Assessment of techniques for protein Structure Prediction (CASP) experiment, where the methods are used to predict disorder on a relatively small number of proteins with maximum accuracy [46]. Here, clearly consensus predictors aiming for maximum accuracy should excel. However, a more practical scenario is represented by high-throughput analysis of protein disorder, e.g. on entire genomes [56]. In this case, the focus is shifted toward fast predictors producing a minimum number of false positives [49]. Over the years, most prediction methods have addressed the first problem, with comparatively little attention to the practicalities of large-scale predictions. This has led to a relative paucity of accurate fast predictors, as adding more prediction layers has produced slightly more accurate but increasingly cumbersome methods (e.g. [36, 37, 56, 68]).

Here we present the ESpritz methods for determining disorder based solely on sequence, aimed for high-throughput applications. The predictor is tested on large datasets of different disorder types, including a novel NMR mobility definition. The sole input for the ESpritz method is the amino acid. It does not require sliding windows to capture local contextual information or any complex sources of information and is shown to be both state-of-the-art and efficient.

3.3 Methods

ESpritz uses bidirectional recurrent neural networks (BRNN) (Baldi et al., 1999) to predict disorder from sequence information. A BRNN can be likened to an ensemble of three neural networks, learning the N-terminal sequence context, the sequence and the C-terminal sequence context, respectively. Where regular neural networks use a sliding window of predetermined size, BRNNs learn this context information through the recursive dynamics of the network, reducing the number of parameters and extracting information implicitly from the surrounding local context. Another important feature is a top layer filter which takes as input 'semi-global' information from the bottom layer in analogy to [43]. Parameter learning proceeds by gradient descent and the back-propagation algorithm and the output contains two units producing the probability of order and disorder. The total number of parameters depends on the number of neuronal units in the various network layers. ESpritz never exceeds 5886 which is acceptable and very unlikely to overfit considering the amount of training examples which is between 30 and 100 times the number of parameters (Supplementary Table A.2.1). BRNNs with a similar number of parameters have already been applied to various prediction problems, e.g. secondary structure [43]. In the following, we introduce the four basic variants and their consensus (Table 3.1) before describing the datasets used and evaluation measure.

Acronym	Sequence	Profile	Consensus
ESpritz	Both	Yes	Four-way
ESpritzP	Both	Yes	Two-way
ESpritzS	Both		Two-way
ASpritzP	Attributes	Yes	
ASpritzS	Attributes		
SSpritzP	Identities	Yes	
SSpritzS	Identities		

Table 3.1: Definitions for the Spritz variants. Sequence relates to the input information with attributes (five Atchley scales) or identities (20 residue types). Two-way consensus is calculated for the two sequence coding schemes with and without profile. Four-way consensus is calculated among all four basic variants.

3.3.1 Sequence-only predictions

In analogy to our work on repeat proteins [52], ASpritz uses the five Atchley sequence metrics [69] as numerical sequence attributes for BRNN input. Each scale, listed in Table 2 of [69], was obtained by clustering almost 500 different amino acid scales from the AAindex database [70]. The scales were shown to reflect polarity, secondary structure, molecular volume, codon diversity and electrostatic charge [69] and may allow for a richer amino acid representation. As the five scales have different, asymmetric, ranges they require normalization in order to be useful as neural network inputs. As in our previous work, normalization is performed so that the squares of the scales sum to 1 [52]:

$$\sum_X [A_t(X)]^2 = 1 \quad (t = 1, 2, \ldots, 5) \tag{3.1}$$

where X=[A, C, D, E,. . . , W, Y] is the one letter code corresponding to each of the 20 amino acids, and At(X) is the sequence metric for amino acid X. ASpritz has five inputs i to the neural network for each sequence position k, each representing one normalized Atchley scale. If position k in the sequence contains amino acid X then the five inputs to this system are as follows:

$$i_k^t = A_t(X) \quad (t = 1, 2, ..., 5) \tag{3.2}$$

Alternatively, SSpritz considers the 20 amino acids in ‘one-hot’ encoding. It consists of 20 inputs i where each unit for sequence position k is allocated for 1 of the 20 amino acids:

$$i_k^{1-20} = R_k(X) \tag{3.3}$$

where X is the residue at position k, $R_k(X)\epsilon R20$ is an alphabetically ordered vector of positions R_kj corresponding to the 20 amino acids (i.e. [A, C, D, E,. . . , W, Y]). $R_kj = 1$ if the position amino acid is in the sequence at position k and $R_kj = 0$ otherwise. ASpritz and SSpritz are combined into a consensus score ESpritzS. As previously shown for CSpritz [68], simply averaging the two scores proved most effective (data not shown).

3.3.2 Multiple sequence alignment-based methods

Evolutionary information in the form of multiple sequence alignments is commonly used to improve predictor performance. Here, the two sequence encodings are extended to accommodate sequence profiles. Let a sequence profile pk(X) give the probability of finding amino acid X in the multiple sequence alignment at position k along the sequence (gaps not considered). For the Atchley scales, the profile-based predictor (ASpritzP) contains six inputs i for sequence position k, one for each scale plus gaps:

$$i_k^t = \sum_{X \epsilon C_k(X)} A_t(X) p_k(X) \quad (t = 1, ..., 5) \quad i_k^6 = \frac{g}{n+l} \tag{3.4}$$

Predictor	Sens	Spec	Sw	AUC
CSpritz	79.63	85.05	**64.68**	0.8993
SSpritzP	76.48	87.02	**63.50**	0.8893
ESpritz	77.23	85.63	**62.85**	**0.8912**
ESpritzP	77.49	85.29	**62.77**	0.8876
MULTICOM	81.99	80.37	62.37	0.8879
ASpritzP	76.06	84.81	60.86	0.8766
ESpritzS	73.67	86.23	59.89	0.8748
SSpritz	73.98	85.39	59.37	0.8699
ASpritz	73.03	86.23	59.25	0.8721
NN (w=23)	69.39	87.74	57.12	0.8645
PONDR-FIT	69.20	86.73	55.92	0.8609
Disopred	56.48	93.87	50.33	0.8391
IUPred (short)	54.00	94.95	48.92	0.8475
Spritz (old)	41.63	93.09	34.69	0.7884
DisEMBL465	31.91	97.67	29.61	0.8320

Table 3.2: Performance measured on X-ray disorder for 569 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined.

where Ck(X) is the set of amino acids for position k, g is the number of gaps, n is the number of non-gaps and l is the total number of sequences involved in the multiple sequence alignment. Alternatively, when considering the 20 amino acids, the sequence profile pk(X) is multiplied against the input vector from SSpritz:

$$i_k^{1-20} = \sum_{X \epsilon C_k(X)} R_k(X) p_k(X) \tag{3.5}$$

The fraction of gaps in the sequence profile is included as an additional input (ik21) and the system is called SSpritzP. Averaging the ASpritzP and SSpritzP output into a consensus prediction will be termed ESpritzP, while ESpritz is the ensemble combination of all four single predictors by averaging their probabilities (Table 3.1). In order to assess the effect of using the BRNN architecture, we also train a standard feed-forward neural network (NN) using ‘one-hot’ encoding [Equation (3)] and a fixed window size of 23 residues (6484 parameters) found to be the best combination on the X-ray training set.

3.3.3 Datasets

To train and measure the performance of the predictors, we created several datasets from structures deposited at the PDB [39] and from experimental data as deposited in the Disprot database [40]. Training and testing data are strictly separated, with appropriate separation and redundancy reduction (maximum ∼25% for X-ray and NMR, 40% for DisProt) to present truly unseen data during testing. Unless stated otherwise, all alignments were calculated using PSI-Blast [42] using options -b 3000 -e 0.001 -h 1e-10. The alignment sequence database for PSI-Blast was non-redundant (NR) at a 90% sequence identity level. Low complexity sequences, transmembrane helices and coiled-coil regions were filtered from the sequence database using Pfilt [71]. All new datasets used for training and testing are available for download from URL: http://protein.bio.unipd.it/espritz/.

X-ray disorder

The X-ray training set was constructed from crystallographic structures deposited in the PDB until May 1, 2008, restricted to X-ray protein chains of length between 25 and 2000 amino acids, with resolution at most 2.5 Å and R-factor up to 25%. Disordered residues are defined as those with missing backbone C-alpha atoms. All proteins were classified into those containing at least three consecutive disordered amino acids and those with no disordered regions. Both subsets were sorted by decreasing quality and reduced by sequence identity using UniqueProt [41] to an HSSP value of 0 (%25% over 100 aligned residues) giving priority to proteins with better quality (-m option). The resulting lists were merged and redundancy reduced in a similar manner leaving proteins with disordered regions as a priority. The training set contains 3244 proteins with 660 120 residues of which 5.68% are disordered. The test set was created using the same procedure for proteins released by the PDB between May 1, 2008 and September 13, 2010. The test set contains 569 proteins with 94 520 residues of which 7.34% are disordered.

DisProt disorder

The training set is based on DisProt version 3.7 (January 28, 2008) in order to ensure that we have sufficient testing data, i.e. proteins annotated between 2008 and 2010. It contains 484 proteins with 219 424 residues where we consider 25.71% disordered. Here we define a residue as disordered if the DisProt curators consider the residue to be disordered at least once while all other residues (including unannotated) are considered structured. Since many residues are unannotated in DisProt, and this could be a potential source of bias in testing, we extend the coverage of the test set by annotating DisProt version 5.7 with PDB structures in analogy to the work of [49]. Briefly, all sequences from DisProt 5.7 with ¿40% sequence identity to the training set are removed. The remainder was matched to PDB entries through the UniProt accession code from DisProt and linked through the SIFTS database [72]. Entries were annotated for disorder considering DisProt definitions where available. Unannotated residues are deemed structured if they exist in both the SEQRES and ATOM sections, and as disordered for regions of length at least five residues missing from the latter. The DisProt and PDB sequences were then aligned to take into account possible variations, and PDB disorder annotations transferred to the DisProt sequence with at least 95% sequence identity. The new test set contains 52 proteins where 49.72% of the residues are unannotated, 41.04% are disordered and 9.24% are ordered for a total of 18 096 residues.

NMR mobility

NMR mobility datasets are calculated using the Mobi server [59]. Mobi is based on a simple algorithm to find regions with different conformations among all models in an NMR ensemble. Briefly put, residues with a variation of atomic coordinates and torsion angles between models above a fixed threshold are marked as mobile. The threshold was optimized to replicate the NMR disorder definition used in CASP8 [46]. The extraction and redundancy reduction is identical to the X-ray datasets (see above) except that PDB NMR structures are considered (no quality filter). The training set consists of 2187 proteins with 173 154 residues of which 16.90% are considered disordered. The testing set contains 671 proteins with 59 384 residues and 18.70% disorder.

Other datasets

The MxD dataset [36], sharing ¡25% sequence identity with CASP8, was downloaded from the website at URL: http://biomine-ws.ece.ualberta.ca/MxD.txt. Note that the 5-fold cross-validation used here might differ from the one used in the paper. The Homo sapiens protein sequences were downloaded from URL: ftp://ftp.ncbi.nih.gov/genomes/. The total number of proteins as of September 2010 for the human genome was 39 151. The time comparison was calculated for 1% of the human genome (i.e. 391 proteins).

3.3.4 Comparison with available methods

ESpritz is compared with several other methods which were either downloaded (Disopred, MULTICOM, DisEMBL, IUpred) or used as web server (PONDR-FIT). The original Spritz method [34] and our recently published improvement CSpritz [68] are also shown for comparison. In all cases, the methods were used with default parameters. Multiple sequence alignments for Disopred and ESpritz were calculated on the 90% reduction of the May 2008 non-NR database and pre-processed with the pfilt program. MULTICOM alignments were calculated using an internal database. The CASP9 data was downloaded from the official website (URL: http://predictioncenter.org/casp9/). Note that, in contrast to our previous paper [68], 252 residues marked as 'x' in CASP9 were not considered in the analysis and disordered segments of 1 or 2 residues are considered. ESpritz is available both as a web server and as a pre-compiled executable for Linux machines from URL: http://protein.bio.unipd.it/espritz/.

3.3.5 Measuring performance

The assessment of our predictions use similar measures as used in CASP8 and previous CASPs [46]. There are two types of measures. Binary measures are calculated once the probability decision threshold is found. All our disorder probability thresholds were found on the corresponding training sets. We define the binary measures sensitivity ($Sens = TP/N_{dis}$), specificity ($Spec = TN/N_{ord}$), selectivity ($Sel = TP/(TP + FP)$), F-measure ($F = 2 * Sen * Sel/(Sen + Sel)$), Matthews correlation coefficient (MCC), accuracy ($Acc = (Sens + Spec)/2$) and the score ($Sw = Sens + Spec - 1$) [73]. TP, TN, FN and FP are the number of true positives, true negatives, false negatives and false positives, respectively (positive is disorder, negative is order). N_dis, Nord are the number of disorder and ordered residues, respectively. We also use area under the receiver operator curve (AUC), calculated between false positive rate (FPR = 1 - specificity; x axis) and true positive rate (TPR = sensitivity; y axis), as a measure of the quality of the probabilities. As in CASP8 [46], the statistical significance of the evaluation scores was determined by bootstrapping (see A.2): 80% of the targets were randomly selected 1000 times, and the standard error of the scores was calculated (i.e. 1.96 x standard_error gives 95% confidence around mean for normal distributions).

3.4 Results

3.4.1 Training and consensus building

The different Spritz variants (Table 3.1) have been trained on pre-CASP8 data from all three flavors of protein disorder (X-ray, DisProt, NMR). A comparison between training and test set performance can be found in Supplementary Table A.2.1, which also show how the training examples is 30–100 times the number of parameters for each predictor. The consistent performance on independent training and test sets is an indication of the good generalization capability of ESpritz. Table 3.2 and Supplementary Table A.2.2 show the results for the X-ray disorder definition in comparison with various other methods and Table 3.3 and Supplementary Table A.2.3 for analogous data on DisProt. From the data, it is apparent that the Spritz variants present high specificity and clearly outperform the methods used for comparison. The difference is more pronounced for the DisProt dataset, probably because it contains longer disordered segments on which fewer methods have been trained. This appears to confirm the hypothesis that long (i.e. DisProt) and short (i.e. X-ray) disorder are different flavors [56]. As expected, the profile-based predictors perform slightly better than the sequence-only ones, although the latter are still competitive. The effect of the BRNN architecture becomes apparent in comparison to the standard neural network (NN), which performs significantly worse. It is also interesting to note that the differences between sequence encoding schemes appear minimal. Each Spritz variant remains, nevertheless, competitive against other state-of-the-art methods such as IUPred and DisoPred. These results are also verified using 5-fold cross-validation on an independent set provided by the MxD database [36] (Supplementary Table A.2.4).

Once the performance has been established, the question becomes whether the same disorder information is detected to a different degree or slightly distinct signals are picked up by the Spritz variants. This information could then be used to create a consensus predictor. Figure 3.1 shows how the different methods represent somewhat different predictions and an implicit confidence estimate. Whenever the four variants agree, as they do for ∼80% of all residues, the accuracy is close to 100% for order and ∼xs40% for disorder (see Supplementary Table A.2.5 for Pearson's correlation coefficients). The relative rarity of inter-

Predictor	Sens	Spec	Sw	AUC
ESpritz	77.51	80.37	**70.58**	**0.892**
ESpritzS	73.78	93.66	**67.44**	**0.901**
ESpritzP	75.47	91.69	67.15	0.888
PONDR-FIT	68.89	93.18	62.08	0.885
IUPred (long)	61.57	96.83	58.4	0.878
Disopred	64.19	93.9	58.54	0.824
CSpritz	79.07	78.02	57.09	0.877
MULTICOM	77.35	78.89	56.23	0.853
NN (w=23)	69.05	82.66	51.71	0.815
IUPred (short)	49.17	97.61	46.77	0.855
Spritz (old)	81.74	59.86	41.6	0.770
DisEMBL465	32.51	98.03	30.53	0.792
DisEMBL	46.74	82.89	29.63	0.692

Table 3.3: Performance on enhanced Disprot disorder for 52 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined.

mediate cases should allow a simple averaging of the probabilities (ESpritz) to outperform each individual method, as shown in Table 3.2. In order to maintain the efficiency of the sequence-only variants, the two partial combinations between SSpritz/ASpritz (ESpritzS) and SSpritzP/ASpritzP (ESpritzP) are also shown. In the following, for simplicity we will only show results for the ESpritz variants. Full data is available in A.2.

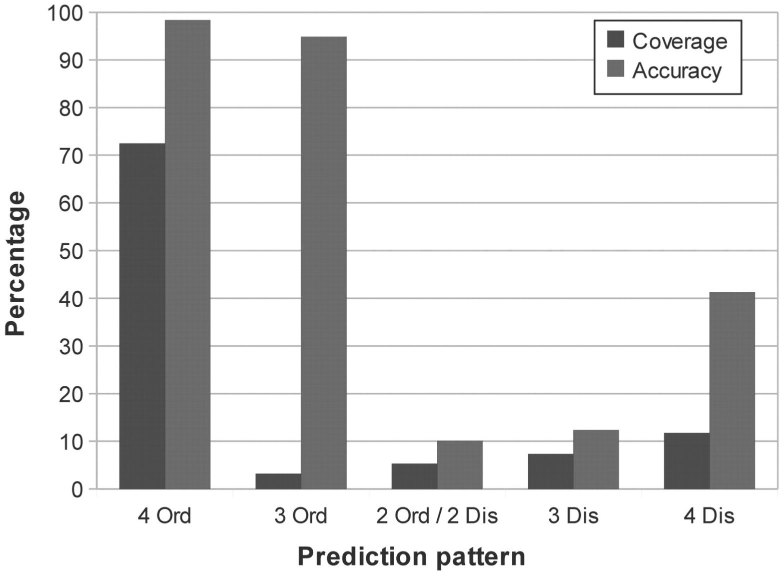


Figure 3.1: Agreement between the four Spritz variants. The relative frequency (coverage) of each state distribution for the four predictors is plotted together with the accuracy for that case.

3.4.2 Novel NMR mobility flavor

A unique feature of ESpritz is the explicit prediction of NMR mobility through a dedicated predictor. To the best of our knowledge, no other method has been developed to predict disorder defined on mobile residues in NMR structural ensembles, although this has been benchmarked since CASP8 in 2008 [46]. One of the problems was the unique automatic definition of NMR mobility, which we have recently addressed [59]. As can be seen from the results shown in Table 3.4 (and Supplementary Table A.2.6), ESpritz has a strong performance and even the sequence-based predictors outperform existing methods. NMR mobility appears to harbor a distinct signal that is somewhere between short (X-ray) and long (DisProt) disorder. ESpritz is particularly useful to detect this novel flavor of disorder, although the specificity values remain below those of other variants. The latter may be speculatively attributed to the variability of NMR structures, which combine greater structural flexibility than crystal structures with a wider range of experimental conditions. In general, the NMR flavor appears to predict more disorder than the X-ray and DisProt ones, with segments of length somewhere between the other two. Supplementary Figure A.2.1 shows as example ESpritz predictions for the human p53 protein using the three different flavors.

Predictor	Sens	Spec	Sw	AUC
ESpritzP	72.83	79.19	**52.01**	**0.8366**
ESpritz	72.53	79.33	**51.85**	**0.8401**
ESpritzS	66.94	80.77	47.71	0.8179
CSpritz	71.93	74.74	46.67	0.7964
MULTICOM	75.14	69.55	44.69	0.7976
PONDR-FIT	63.74	75.55	39.29	0.7533
Disopred	48.69	89.19	37.88	0.7556
IUPred (short)	45.07	90.73	35.79	0.7505
NN (w=23)	51.73	80.27	31.99	0.7301
Spritz (old)	29.29	96.49	26.28	0.7481
DisEMBL HL	25.38	82.08	7.46	0.6329

Table 3.4: Performance on NMR disorder for 671 structures. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined. IUpred long performs worse than IUPred short and is not shown. DisEMBL HL is the hot loops predictor which performs better than DisEMBL465 on this dataset.

3.4.3 Comparison on CASP9 data

In order to fully compare our method to the state-of-the-art, we use data from the recent CASP9 experiment. Table 3.5 and Supplementary Table A.2.7 show the results for all targets, while Supplementary Table A.2.8 shows only the NMR targets. ESpritz is significantly more accurate than all methods using both the Sw and AUC criteria. This strong performance can be partially explained by the use of a dedicated NMR prediction mode. Perhaps not unexpectedly, the ESpritz variants excel on NMR targets thanks to the novel NMR prediction mode, where they outperform the best CASP9 methods by at least 15% on Sw and 7% on AUC (Supplementary Table A.2.8). ESpritz also outperforms our recent consensus-based CSpritz method [68], which combines three different predictors including a preliminary version of SSpritz but lacks an explicit NMR mode.

Predictor	Sens	Spec	Sw	AUC
ESpritz	67.41	87.52	**54.82**	**0.8558**
PRDOS2(291)	60.78	90.03	50.65	0.8544
CSpritz	63.66	86.37	49.91	0.8316
Multicom-refine(119)	64.98	85.02	49.89	0.8217
Biomine(351)	59.63	89.01	48.48	0.8213
ESpritzS	59.75	88.83	48.43	0.8308
GSMETADISORDERMD(374)	65.72	81.93	47.57	0.8184
MASON (193)	53.70	92.76	46.25	0.7438

Table 3.5: Performance on 117 CASP9 targets (19 NMRs and 98 X-rays). The top five performing CASP9 groups are shown with their official group name and number in brackets. Methods performing at least as well as or not statistically different from ESpritz are highlighted in bold. Methods performing at least as well as or not statistically different from ESpritzS, our best fast predictor, are in italics and underlined. Note that group 351 was missing 10 proteins.

3.4.4 Large-scale predictions

The large-scale analysis across entire genomes is an important application of disorder predictors [17, 49, 56], both to further our understanding of disorder as a biological phenomenon and to help establishing protein function. The efficiency in terms of CPU time versus AUC of different methods on a randomly selected 1% of the human genome is shown in Figure 3.2. As can be seen, the field is divided between fast methods with somewhat lower accuracy and much slower methods using multiple sequence alignment information from PSI-BLAST. The latter improve AUC by up to four percentage points at the cost of four orders of magnitude of computation time. ESpritzS combines the best of both worlds, by maximizing performance for a fast method that does not require multiple sequence alignments.

When analyzing large numbers of sequences, it can be especially useful to be able to limit the number of expected false positives to avoid drawing false conclusions on the prevalence of disorder [49]. Figure 3.3 shows a typical receiver-operating characteristic curve plot on the X-ray dataset. As can be seen, ESpritzS is particularly good at low FPRs up to around 5% FPR, after which the relative FPR increases. The optimal binary Sw decision threshold can be found around 13.5% FPR. An alternative 5% expected FPR threshold was derived on the training dataset for all ESpritz variants. On the testing data this yields ∼63% sensitivity and 45% selectivity at ∼6.5% FPR. The more stringent decision threshold provides a simple way to limit the number of false positives at the expense of somewhat lower sensitivity. At this low FPR threshold, the ESpritz variants perform better than the other tested methods using similar decision thresholds. For the full analysis on all datasets, including F-measure and MCC values, please refer to A.2. The 5% expected FPR threshold should prove useful for high-throughput

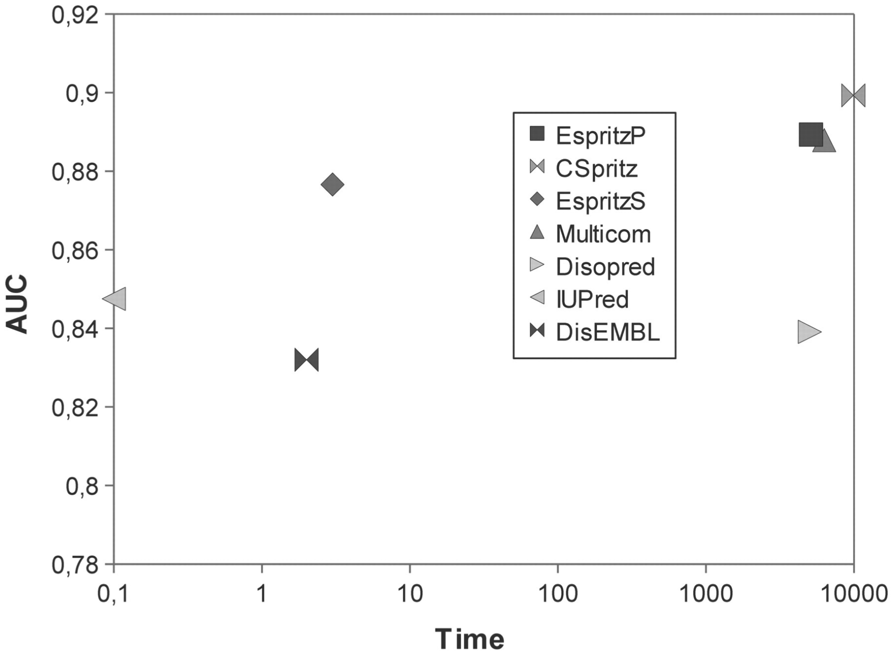


Figure 3.2: Time versus performance plot for different predictors. The time in minutes for pedicting 1% of the human genome on a single Intel Xeon processor core is plotted against the AUC for each locally installed method. Note that the time axis uses a logarithmic scale.

applications requiring low FPRs. It is, therefore, expected that ESpritzS can provide a valid alternative in applications requiring the high-throughput analysis of thousands of sequences or entire genomes.

3.5 Conclusions

We have presented a new ensemble of disorder predictors, called ESpritz, having state-of-the-art performance on three different flavors of disorder. Compared with our previous methods Spritz [34] and CSpritz [68], ESpritz combines a more sophisticated BRNN architecture with enhanced definitions of disorder flavors. The BRNN improves performance slightly on X-ray data but substantially on the other two disorder datasets. The comparatively larger improvement on DisProt data may be related to our usage of an enhanced re-annotation of ordered segments in DisProt [49], providing a clearer distinction between the two states. Unsurprisingly, where ESpritz really excels is on NMR mobility. This is a novel definition which, to the best of our knowledge, was never incorporated before in a disorder predictor. Our comparison with existing methods, and the strong performance on CASP9 data, suggest that NMR flexibility is encoded by a somewhat different but related signal to the other two flavors. The NMR flavor appears to capture a larger fraction of amino acids at the borderline between the ordered and disordered states, perhaps at the expense of more false positive predictions. Nevertheless, the differences between disorder datasets support the hypothesis of different flavors being encoded by somewhat different sequence features as suggested by [56]. The second major improvement in ESpritz is the creation of a sequence-only predictor which is four orders of magnitude faster than multiple

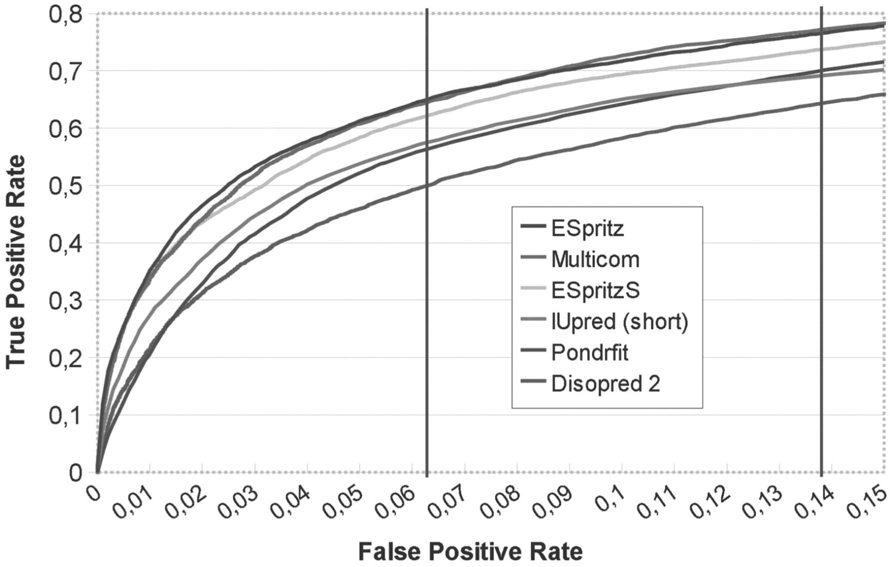


Figure 3.3: Receiver-operating characteristic curve for X-ray test set data. The plot shows the FPR in the region from 0% to 15% false positives for various methods. The two vertical lines represent the ESpritz decision thresholds corresponding to a predicted 5% FPR (left) and the optimal Sw threshold (right).

sequence alignment-based methods at the expense of a slight reduction in accuracy. This allows the user to choose between highly accurate predictions for single proteins or high-throughput predictions at genomic scale. The third, and final, improvement in ESpritz is the definition of an alternative, more stringent, disorder threshold limiting the expected FPR to 5%. This allows the user to choose between detection of more disorder or highly selective predictions depending on the data being analyzed. The very high specificity of ESpritz also ensures a low rate of false positives on high-throughput problems, making it even more valuable for this task. This scenario is typically overlooked when developing disorder prediction methods, but accounts for a large part of the biological problems to be addressed. We believe that ESpritz offers an accurate and efficient way to address many biologically relevant problems encountered with disordered proteins.

4. MobiDB: a comprehensive database of intrinsic protein disorder annotations

This chapter was first published in Di Domenico T, Walsh I, Martin AJM, Tosatto SCE (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28: 2080–2081.

4.1 Summary

Motivation: Disordered protein regions are key to the function of numerous processes within an organism and to the determination of a protein's biological role. The most common source for protein disorder annotations, DisProt, covers only a fraction of the available sequences. Alternatively, the Protein Data Bank (PDB) has been mined for missing residues in X-ray crystallographic structures. Herein, we provide a centralized source for data on different flavours of disorder in protein structures, MobiDB, building on and expanding the content provided by already existing sources. In addition to the DisProt and PDB X-ray structures, we have added experimental information from NMR structures and five different flavours of two disorder predictors (ESpritz and IUpred). These are combined into a weighted consensus disorder used to classify disordered regions into flexible and constrained disorder. Users are encouraged to submit manual annotations through a submission form. MobiDB features experimental annotations for 17 285 proteins, covering the entire PDB and predictions for the SwissProt database, with 565 200 annotated sequences. Depending on the disorder flavour, 6–20% of the residues are predicted as disordered.

4.2 Introduction

During the last decade, strong evidence has surfaced indicating that many proteins function in a natively unfolded or intrinsically disordered state [9, 74]. These regions have been shown to play important roles in various biological processes [75]. The amount of disorder within a proteome seems to correlate with the complexity of the organism, especially in eukaryotes [17]. The existence of different flavours of disorder has been proposed [60], and disordered regions have been categorized according to their function with a suggested coupling between disorder conservation and protein function [56, 76].

The main repository for experimentally determined disorder is the DisProt database [40], containing manually curated information on currently ca. 650 proteins from the literature. Although invaluable as a gold standard, DisProt represents only a fraction of the known protein sequences posing a bottleneck for large-scale analysis of intrinsic protein disorder. Many prediction methods have long resorted to considering the lack of coordinates in X-ray protein structures as a proxy for intrinsic disorder [17, 68]. This increases the number of available sequences by an order of magnitude for mostly short disordered segments. Recently, our group has also developed a method to define intrinsic disorder by looking at mobile regions in NMR structures [59]. Herein we describe MobiDB, a centralized resource for disorder annotation in protein sequences.

4.3 Implementation

MobiDB is a relational PostgreSQL database consisting of 11 tables. The data are divided into two subsets: MobiDB-xp, containing only proteins with experimental annotation and MobiDB-full, for proteins with predictions. Annotations are extracted from different sources, currently yielding eight-different flavours. The PDB-X-ray data are obtained by considering as disordered residues whose Cα atoms are missing from X-ray crystallographic structures deposited in the PDB [39]. The novel PDB-NMR is generated by processing NMR structures in the PDB with MOBI [59] and DisProt data [40] are obtained directly. Predictions are obtained by running ESpritz [77] (long, X-ray and NMR) and IUPred [30] (short and long) on all SwissProt sequences. Sensitivity and specificity values of each predictor on a common benchmark can be found online. Sequences are linked to UniProt [78] and Pfam [79] through SIFTS [72] for PDB structures and DisProt. A consensus disorder score assigns higher weights to experimental annotations over predictions (see online documentation). Disorder is divided into constrained and flexible based on conservation [76]. Secondary structure in PDB files is identified with DSSP [80]. Manual data curation is also supported and users are encouraged to submit annotations through a feedback submission form.

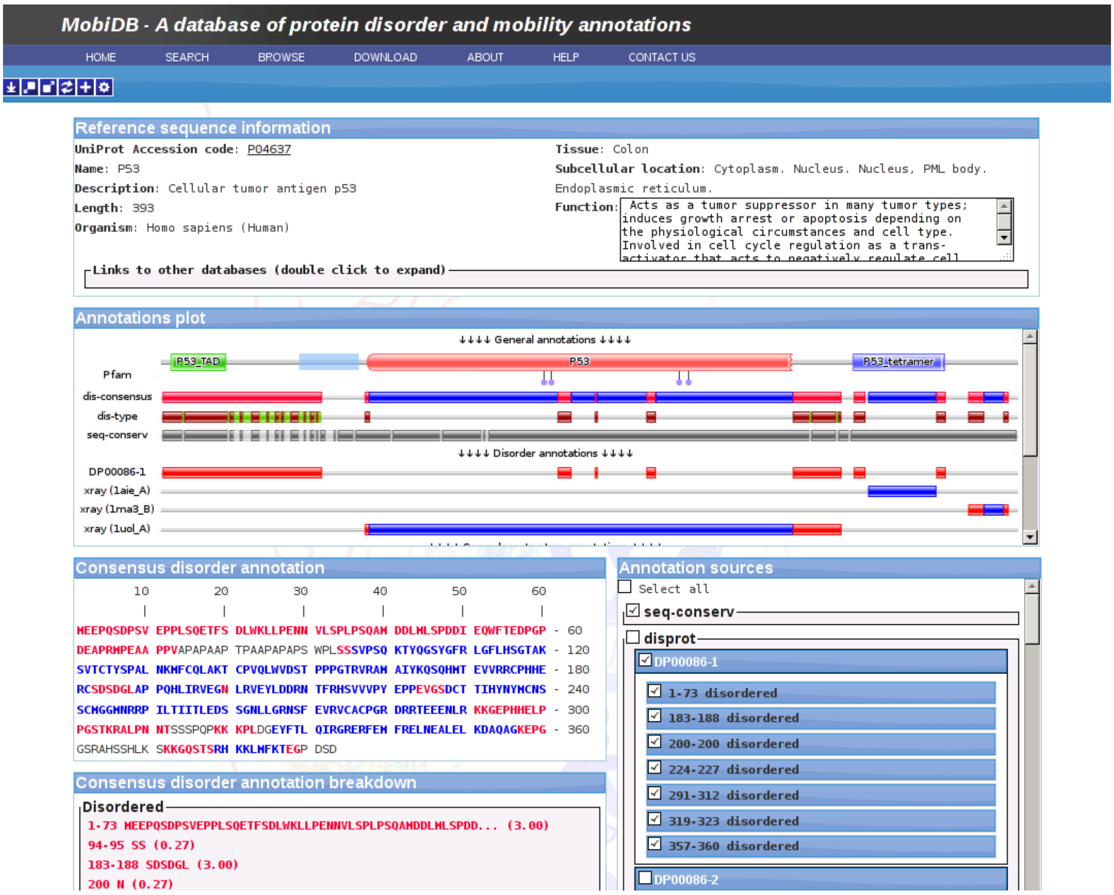
4.4 Usage

MobiDB was designed with two main scenarios in mind. First, a user wishes to analyse a particular protein of interest and dynamically access all the available disorder information, with the option to generate (and download) a consensus annotation. Second, the user would like to obtain a dataset of disorder information for a protein ensemble with certain characteristics, downloading it for offline usage and analysis with other tools. MobiDB offers two options to access the information. The user may either browse the different MobiDB-xp flavours (PDB-X-ray, PDB-NMR or DisProt) or use the search function. The latter offers three options: by identifier, standard and BLAST [42]. For a full explanation please refer to the online documentation. After selecting a browse option or performing a search, the user will be presented with the results page. In this page, it is possible to either select a single entry and proceed to the protein visualization interface or to generate a dataset containing disorder annotations for all selected proteins. This dataset will consist of two FASTA formatted files for each protein: one containing an alignment of the reference sequence and all the annotating sequences and the second containing annotations associated to these sequences.

The protein visualization interface (Fig. 4.1 was designed as an annotation sandbox for dynamical protein annotation. The interface is composed of a variety of widgets or boxes that can be dragged, expanded or collapsed, allowing for the optimization of the available workspace. The 'reference sequence information' widget displays data for the chosen reference sequence from UniProt. The 'annotation sources' widget allows selecting or deselecting annotating sequences, and/or their corresponding regions. The 'annotations plot' widget offers a graphical representation of the reference sequence and the chosen annotating sequences, while also displaying Pfam and secondary structure annotations (where available). The 'dynamic annotation' widget displays the colour coded sequence for the reference protein, according to whether a region is annotated as ordered or disordered. A second set of three colours for predicted disorder annotations is provided (in lighter shades). Consensus disorder predictions are provided together with a classification into flexible and constrained regions [76].

As an example, we show the annotations for the human p53 tumour suppressor protein in Figure 4.1. p53 contains structured tetramerization and core domains linked together and flanked by intrinsically disordered regions. The structure of p53 (or lack thereof) has been widely studied, and a comprehensive model has been built [81]. The MobiDB entry for p53 summarizes this situation well (Fig. 4.1.

MobiDB provides the means to obtain disorder annotations for an extensive set of proteins as a centralized and up-to-date source of information on various available disorder flavours. We are planning on providing manual annotations and integrating more data generated from other predictors to better characterize different disorder flavours and their functional implications.


MobiDB - A database of protein disorder and mobility annotations
HOME
SEARCH
BROWSE
DOWNLOAD
ABOUT
HELP
CONTACT US

Reference sequence information

UniProt Accession code: P04637
Name: P53
Description: Cellular tumor antigen p53
Length: 393
Organism: Homo sapiens (Human)

Tissue: Colon
Subcellular location: Cytoplasm. Nucleus. Nucleus, PML body. Endoplasmic reticulum.
Function: Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell

Links to other databases (double click to expand)

Annotations plot

Consensus disorder annotation

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP - 60
DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK - 120
SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE - 180
RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS - 240
SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP - 300
PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG - 360
GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD
```

Consensus disorder annotation breakdown

Disordered

1-73 MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDD... (3.00)
94-95 SS (0.27)
183-188 SDSDGL (3.00)
200 N (0.27)

Annotation sources

☐ Select all
☑ **seq-conserv**
☐ **disprot**
☑ DP00086-1
☑ 1-73 disordered
☑ 183-188 disordered
☑ 200-200 disordered
☑ 224-227 disordered
☑ 291-312 disordered
☑ 319-323 disordered
☑ 357-360 disordered
☐ DP00086-2

Figure 4.1: Sample MobiDB output for human p53. The top part contains the UniProt description and database links. The sequence plot (central part) summarizes the disorder information graphically, showing the protein sequence horizontally annotated with Pfam domains and the different disorder flavours. Experimental data are shown in stronger colours (ordered in blue and disordered in red) than predictions. Consensus disorder (blue to red colour gradient) and conservation annotations are also shown. The detailed annotations (bottom part) allow the dynamic selection of annotating sequences and show the relevant sequence stretches.

5. Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database

This chapter was first published in Di Domenico T, Walsh I, Tosatto SC (2013) Analysis and consensus of currently available intrinsic protein disorder annotation sources in the MobiDB database. BMC Bioinformatics 14: S3.

5.1 Summary

Intrinsic protein disorder is becoming an increasingly important topic in protein science. During the last few years, intrinsically disordered proteins (IDPs) have been shown to play a role in many important biological processes, e.g. protein signalling and regulation. This has sparked a need to better understand and characterize different types of IDPs, their functions and roles. Our recently published database, MobiDB, provides a centralized resource for accessing and analysing intrinsic protein disorder annotations.

Here, we present a thorough description and analysis of the data made available by MobiDB, providing descriptive statistics on the various available annotation sources. Version 1.2.1 of the database contains annotations for ca. 4,500,000 UniProt sequences, covering all eukaryotic proteomes. In addition, we describe a novel consensus annotation calculation and its related weighting scheme. The comparison between disorder information sources highlights how the MobiDB consensus captures the main features of intrinsic disorder and correlates well with manually curated datasets. Finally, we demonstrate the annotation of 13 eukaryotic model organisms through MobiDB's datasets, and of an example protein through the interactive user interface.

MobiDB is a central resource for intrinsic disorder research, containing both experimental data and predictions. In the future it will be expanded to include additional information for all known proteins.

5.2 Background

Intrinsic protein disorder is becoming an increasingly important topic in protein science [56, 74, 82]. Protein function has been traditionally thought to be determined by tertiary structure. Over the last decade, intrinsically disordered proteins (IDPs) have been found to be important in many important biological processes [9, 12, 54]. IDPs are widespread in natural proteins, especially in eukaryotic organisms [17, 83], and are frequently associated with molecular recognition [11, 84]. They have been observed to be common among hub proteins, i.e. those with many interaction partners [58] and also to play a key role in human disease [57]. In addition, protein disorder is important for experimental protein characterization since difficulties often arise when long disordered regions are present, which frequently happens at the N and C termini [85]. IDPs represent a heterogeneous concept with many different and elusive definitions [8] which can be traced back to different indirect experimental methods.

5.2.1 Sources of disorder information

Currently available sources for intrinsic disorder annotations can be divided in two main groups. The first group includes annotations inferred from experiment, with evidence in publications. The second group includes annotations automatically extracted by computational tools. The latter can be further subdivided into automatic annotations derived from experimental sources, and automatic annotations obtained from software predictors.

There are currently two available sources of intrinsic protein disorder information with evidence in publications. The DisProt [40] database, a manually curated repository, features disorder and structure annotations for 667 proteins (version 6.00). The IDEAL [86] database, also manually curated, contains information on 209 proteins. The Protein Data Bank (PDB) [39] constitutes the main source of available experimentally-based disorder annotations with over 70,000 different structures. It is widely accepted that missing residues from X-ray structures have a good correlation with intrinsically disordered residues [87]. These missing regions can easily be extracted from structure files deposited in the PDB. Some 6,000 structures solved by NMR experiments are generally deposited as structural ensembles in a single file. These can be used to detect residue mobility [59] which, in a way that is analogous to the missing X-ray regions, are a good indicator of intrinsic disorder. NMR structures were only recently considered in disorder prediction [77], demonstrating the long held belief of different flavours of disorder [56, 74, 88].

A great number of intrinsic disorder predictors have been developed over the last few years [89], with two main scenarios emerging for their application. The first is represented by predictions of disorder on a relatively small number of proteins with maximum accuracy, such as in the CASP experiment [90]. Most existing prediction methods, such as Disopred [17], VSL1 [29] and CSpritz [68], have been trained for this scenario. A more practical scenario is however represented by the genome-scale analysis of disorder [56, 83], where some performance is sacrificed to achieve results in a reasonable time frame. This usually entails using a method that does not require a multiple-sequence alignment, thereby speeding up computation by several orders of magnitude [77]. DisEMBL [33], IUPred [91] and, more recently, ESpritz [77] have been all developed with this scenario in mind.

In the following, we will describe the construction of the MobiDB database of experimental and predicted disorder annotations in proteins [92]. In particular, we will compare the different annotation sources and how they are integrated. A coherent consensus disorder definition will be derived and used to annotate the proteomes of a set of representative model organisms.

5.3 Materials and methods

5.3.1 Database structure

MobiDB [92] data is stored and queried using the PostgreSQL database engine. The database schema is composed of 11 tables and shown in Figure 5.1. The main idea in the database is to have a set of reference protein sequences, which will be annotated by associating as many annotating sequences as possible to them. The reference sequences represent distinct biological objects, e.g. proteins, which can be obtained with unique identifiers from a reference collection such as UniProt. Annotating sequences are obtained from the various sources mentioned in the previous section. They can be mapped at residue-level to the reference sequences, and provide information such as e.g. disorder, secondary structure, and sequence conservation. In principle, the database schema can be used for any sequence-based annotation. The data is partially normalized, although some exceptions to the normal forms have been introduced with the aim of improving efficiency when inserting and querying data.

Data loading is performed as a three-step process. In the first step, annotations are extracted from each annotation source and stored as two Fasta files. One of these files contains the annotating sequences, and the other the annotations extracted from those. An extra comma-separated file is generated which links the annotating sequences to their corresponding reference sequences. In the second step, a script takes the first step output files and generates tab-separated files compatible with the database engine's batch-loading mechanism. During this step, if an annotating sequence covers only part of its corresponding reference sequence, an alignment between the two is performed. The potential resulting gaps introduced in the annotating sequence are also transferred to the extracted annotation. The third and final step consists simply of loading the data in batch to the database. To maximize the loading performance, the affected database indices are dropped before the insertion begins. The resulting database constitutes the backend of the application, which will then be accessed by the user interface.

The middle tier of MobiDB is composed of Java Servlets. These receive a query from the front-end, submit it to the backend, and translate the results into hierarchical Java objects. These objects are then transformed into JSON objects, and made available for further processing by the front-end. MobiDB's user interface makes extensive use of modern internet browser features to provide

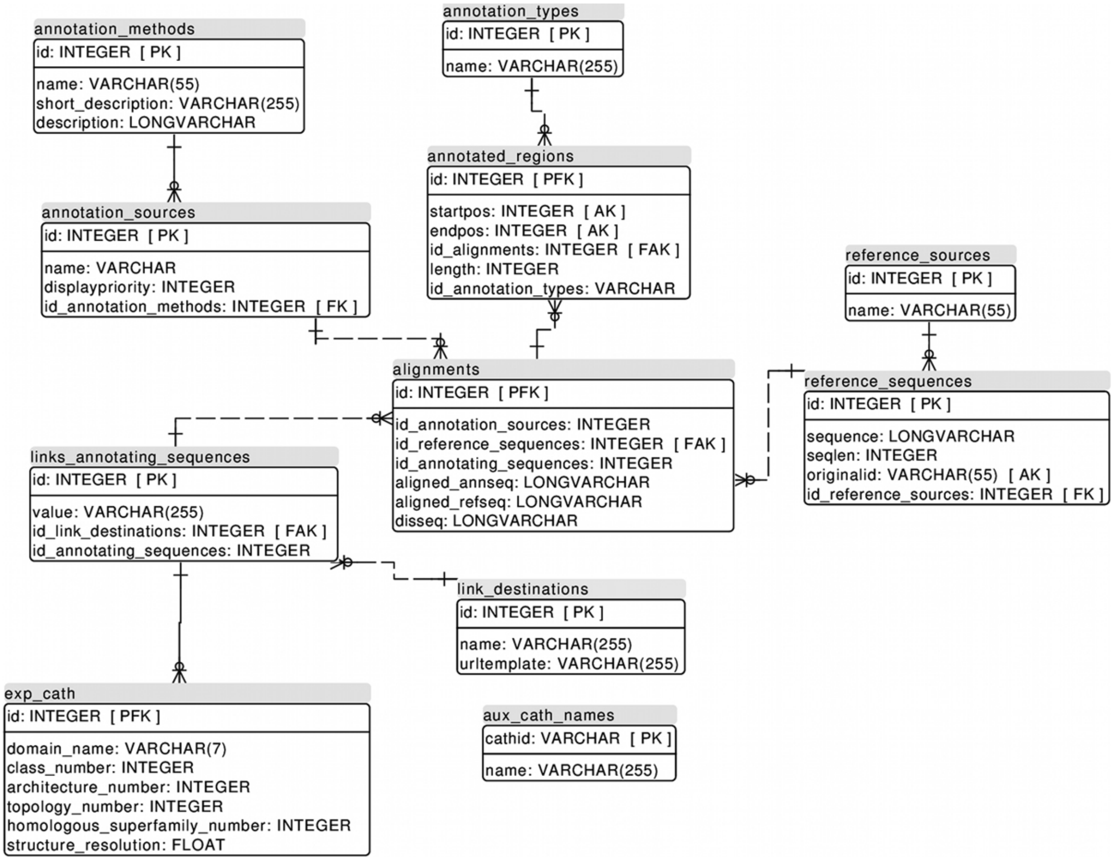


Figure 5.1: Database schema for MobiDB

a flexible user experience. The results provided by the middle-tier are processed and displayed using the JQuery JavaScript library. A widget system was developed which allows for the display of information in independent UI subunits that can be rearranged throughout the screen by the user to fit its needs.

5.3.2 Disorder data and resources in MobiDB

All of the aforementioned disorder sources are integrated into the MobiDB database. XML files from the DisProt and IDEAL databases are parsed for annotations. Information on the corresponding UniProt entries to be linked to those sources is included in also included in the XML files. Annotating sequences from PDB files are extracted by means of custom scripts (X-ray) and the MOBI server (NMR). These annotations are then linked to their corresponding UniProt protein sequences by means of the SIFTS database [93]. In order to capture different flavours of disorder, seven in silico disorder predictors are run against all the reference sequences: Three Espritz [77] flavours (X-ray, NMR, DisProt) and two flavours each for IUpred [91] (short, long) and DisEMBL [33] (remark465 and hot loops).

MobiDB version 1.2.1 integrates the latest versions of its data sources at the time of writing. It features a total of 4,662,776 proteins and covers all complete proteomes for eukaryotic species, as present in the UniProt database [94]. Table 5.1 provides a detailed list of such sources and their corresponding versions.

Database	Reference	Information extracted	Version/Date
UniProt	[94]	Reference sequences	2012-07
DisProt	[40]	Disorder and structure	6.00
IDEAL	[86]	Disorder and structure	2012-05-09
PDB	[39]	Disorder and structure	2012-08-15
SIFTS	[93]	UniProt-PDB links	2012-08-15
Pfam	[95]	Functional domain annotations	Web service
OMA Browser	[96]	Protein orthologs	2012-03
CATH	[97]	Structural classification	3.4
DSSP	[80]	Secondary structure	2012-08-15

Table 5.1: Overview of the databases used in ModiDB 1.2.1. The databases used and relevant references are listed with the description of extracted information and the version or download date included in MobiDB.

5.3.3 Disorder consensus and weighting

For each protein with experimental annotations, the average of annotating sequences in the MobiDB database is seven. The annotations from in silico predictors are excluded from this average, since they require only the protein sequence as input and can therefore provide annotations for all proteins in the database. Furthermore, different disorder annotation sources may reflect different types of disorder phenomena. Given these facts, a simple method to combine annotations would allow for a more integral vision of disorder information. With this in mind, we developed a novel consensus disorder annotation that integrates all available disorder annotations for a protein. The consensus is calculated for each position of the reference sequence, by taking into account the corresponding positions in the annotating sequences whenever they are available. It is composed of two values: disorder level, and annotation score. The disorder level evidences how much the selected annotations agree on whether a given position of the reference sequence is structured or disordered. It is an integer value ranging from 0 to 9, with 0 meaning full agreement on a region being structured, and 9 meaning full agreement on a structure being disordered. It is calculated by the following formula:

$$l = min\left[round\left(\frac{10 * \sum dw}{\sum dw + \sum sw}\right), 9\right] \tag{5.1}$$

where $\sum dw$ is the sum of weights of annotations considering the region disordered, and $\sum sw$ is the sum of weights of annotations considering the region structured. The annotation score evidences the strength of a given consensus annotation. It is the sum of the weights of every annotation that agrees with the final consensus for a certain region. Its objective is to allow the classification of regions according to the amount of data backing up the resulting annotation. This amount is also dependent on the relative weight of each annotation. In all cases, the sums are calculated over all the annotations corresponding to a certain position of the reference sequence. This may be visualized as the columns in an alignment between the reference sequence and its corresponding annotating sequences. In the case where an annotating sequence has no annotation for a certain reference sequence position, its contribution to the sum is zero. In all cases the minimum value of the sums is zero, and the maximum will depend on the number of annotations available, and the weight assigned to each of them.

Empirical weight factors have been derived for each disorder annotation source. Intuitively, the rationale is to favour manually curated annotations (DisProt and IDEAL) over experimental structures from the PDB, and the latter over all predictors. The weighting factors were thus chosen to resemble this situation, with X-ray structures judged depending on resolution and preferred over NMR models. The weights for annotations obtained from the DisProt and IDEAL databases were chosen so that having a few high resolution X-ray structures can tilt the disorder consensus towards ambiguity as these are may represent regions of alternating structure. DisProt and IDEAL annotations are assigned a weight of 3, to reflect the quality of the manually curated data. Each X-ray annotation is given a weight according to the following formula, which increases the weight as the resolution of the experiment improves:

$$W_{xray} = 1 - \frac{log(r)}{log(rT)} \tag{5.2}$$

where r is the resolution of the experiment, and rT is a user-defined maximum resolution threshold. This threshold allows the user to set a baseline in the form of a minimum resolution required for a structure to provide a significant annotation. In the case where the resulting weight is smaller than 0.2, a fixed value of 0.2 is assigned. PDB NMR structures are assigned a fixed weight of 0.2 each, to reflect the usually higher uncertainty in coordinates obtained by NMR experiments when compared to their X-ray counterparts. Finally, predictor-generated annotations are given a weight of 0.05, which allows experimentally obtained data to prevail whenever it is available.

5.3.4 Sequence conservation and disorder classification

In order to provide information regarding the sequence conservation of disorder, MobiDB [92] also annotates sequence conservation on groups of orthologous protein sequences. For each reference sequence in the database, a search is performed in the OMA Browser database [96] to look for a corresponding group of orthologs. If such a group is found and contains at least 10 members, a multiple sequence alignment is constructed with CLUSTALW [98]. A position in the alignment is considered conserved if the same residue is present in at least 50% of the sequences. Whenever such sequence conservation annotations are available, disordered regions in reference sequences are classified in a way analogous to the definitions introduced by Bellay and co-workers [76]. If the region is disordered and its sequence conserved, it is defined as ”constrained disorder”. If, on the other hand, the region is disordered but the sequence not conserved, it is termed ”flexible disorder”.

5.4 Results and discussion

In order to assess the available information on disorder, it was first necessary to create a new database. MobiDB was thus designed with three main goals in mind: performance, scalability and usability. The database had to maintain good performance both when loading, so it can be updated frequently, and querying, so as to be useful for the public by providing fast response times. It had to be scalable, meaning that performance levels can be maintained when expanding with further information. Last but not least, it had to provide high levels of usability, giving the user a centralized, flexible and useful way to access intrinsic disorder information in an intuitive way. Updates for MobiDB are carried out through a three-step loading process integrated into a single, automated pipeline (see Methods). This allows for the easy regeneration of the entire database with up-to-date information in less than a week's time. Enabled by this fact, and based on the update frequencies of the different sources integrated into MobiDB, we have set a quarterly update interval. Every three months MobiDB will be updated to keep up with recent additions to its information sources.

5.4.1 Use cases for MobiDB

There are two main use cases for MobiDB. The first one is the analysis of a single protein by means of the user interface. The second one is the generation of a custom dataset for offline analysis. Both actions are available after performing a database search, or after accessing one of the browse options. MobiDB supports the UniProt complex search syntax, through a web service call to the UniProt server. This allows to build sophisticated queries with various filters, e.g. organisms and subcellular localizations. All proteins matching the search parameters will be listed along with relevant information for each entry in the Search results page.

From the search results, the user can click on a protein name and be directed to the Protein analysis page. This page features four interactive widgets, each containing different pieces of information regarding the selected protein. The Reference sequence information widget contains general information related to the chosen reference protein, extracted from the UniProt database. The Annotation sources widget contains the different annotated regions from each annotating sequence that has been linked to the reference sequence. The Annotations plot widget provides a graphical representation of the available annotations associated to the reference sequence. This contains general annotations such as Pfam annotations and disorder consensus, as well as all available disorder annotations sources.

Instead of analysing a single protein via the graphical interface, the user can opt to download a dataset containing multiple entries. This can be done by pressing the download button in the top left of the search results page. The exported dataset will is composed of two fasta files. One of them containing all relevant reference and annotating sequences and the other one containing all the corresponding annotations. Pre-computed datasets are available in the download section of the MobiDB website for the different experimental data sources, as well as for each of the 297 complete proteomes.

5.4.2 Analysis

Given the unifying concept of MobiDB, where different disorder data sources are collected and serve to annotate the same sequences, it is interesting to note how these sources relate to each other. In an effort to quantify the differences and similarities and to allow for the comparison, Table 5.2 provides a variety of descriptive statistics. The left half of the table contains residue-level information, while the right half contains region-level information. We define a disordered region as a consecutive stretch of residues annotated as disordered. The residue-level data gives a quick picture on the amount of information each source contributes. It also shows how generous each of them tends to be when annotating a residue as disordered or structured. The region-level data evidences the length distribution of the regions detected by each source. As can be expected, the different disorder sources contain data with different characteristics. There appears to be two well-defined clusters and some outliers. The PDB-xray, ESpritz-xray and IDEAL annotations appear to concentrate on few residues with somewhat longer disordered regions. This can be rationalized as sequence segments which probably do not crystallize. DisProt tends to annotate regions of similar length to the previously mentioned sources, but mostly contains only disorder annotations, yielding more disordered residues. On the other hand, the PDB-nmr, ESpritz-nmr, DisEMBL-hl and, to a lesser degree, the IUPred sources tend to annotate a larger amount of residues, but grouped in shorter regions. This likely can be explained as flexible regions fluctuating in space, which may or may not be entirely disordered. ESpritz-disprot is an outlier which predicts comparatively few residues as disordered, but when disorder is predicted it is for very long segments.

A second test was carried out to better understand the relationship between the different disorder data sources, as defined in Table 5.3, and manually curated disorder definitions. Figure 5.2 shows the agreement between each source and the DisProt and IDEAL annotations used as gold standard. Here, matches or mismatches are only considered when a curated annotation exists. The first striking result is that the two gold standards, DisProt and IDEAL, are rather different. In fact, for proteins with both annotations, the reproducibility of one from the other is around 20%. This is rather puzzling, given how both strive to describe the same phenomenon. Upon closer inspection, it becomes apparent that IDEAL focuses more on shorter disordered regions, which more readily correspond to missing X-ray residues. DisProt on the other hand contains more longer disorder segments. In general, it is harder to reproduce the DisProt annotation than IDEAL.

5.4.3 Consensus

From the various disorder data sources it is a logical step to derive a consensus annotation. The protocol for this is described in Methods and a set of variants defined in Table 5.3 are also tested in Figure 5.2. The predictor consensus agrees ca. 50% of the time with DisProt, but covers almost 90% of the IDEAL annotations, again reinforcing the impression about the differences between these two gold standards. In general, the full consensus was designed to closely reproduce the manually curated data whenever available. This analysis can be taken one

		Residues			Disordered region lengths			
Source	Entries	Annotated	Disordered	Fraction disordered	1st quartile	Median	Mean	3rd quartile
DisProt	794	84,671	79,820	0.943	8	20	58.88	63
IDEAL	207	47,967	6,077	0.127	5	16	49.95	62.25
PDB-nmr	7,556	642,252	120,117	0.187	4	7	19.72	22
PDB-xray	180,373	47,309,921	2,400,507	0.050	5	20	88.96	132
DisEMBL-465	4,662,776	2,070,982,327	238,367,624	0.115	13	26	81.34	85
DisEMBL-HL	4,662,776	2,070,982,327	529,251,895	0.255	12	20	41.92	45
ESpritz-disprot	4,662,776	2,070,982,327	165,087,066	0.080	18	102	239.1	347
ESpritz-nmr	4,662,776	2,070,982,327	621,164,099	0.300	6	14	34.71	36
ESpritz-xray	4,662,776	2,070,982,327	252,651,437	0.122	6	25	106.1	128
IUPred-long	4,662,776	2,070,982,327	462,197,994	0.223	1	4	28.97	22
IUPred-short	4,662,776	2,070,982,327	404,986,684	0.195	2	5	32.69	25

Table 5.2: Comparison between disorder data sources. The different disorder data sources are compared in terms of available sequence entries and distribution of ordered and disordered residues. The distribution of disordered regions is also shown in terms of the lowest (1st) and highest (3rd) quartiles, median and mean.

Label	Definition	Type
DisProt	DisProt database annotations	Exp
IDEAL	IDEAL database annotations	Exp
NMR	PDB NMR annotations	Exp
Xray-2.5	PDB Xray annotations, resolution threshold of 2,5 Å	Exp
Xray-5.0	PDB Xray annotations, resolution threshold of 5 Å	Exp
PDB-2.5	PDB-xray and PDB-nmr annotations, resolution threshold of 2,5 Å	Cons, Exp
PDB-5.0	PDB-xray and PDB-nmr annotations, resolution threshold of 5 Å	Cons, Exp
DisEMBL-465	DisEmbl remark 465 predictions	Pred
DisEMBL-HL	DisEmbl hot loops predictions	Pred
Espritz-disprot	ESpritz DisProt predictions	Pred
Espritz-nmr	Espritz NMR predictions	Pred
Espritz-xray	Espritz XRay predictions	Pred
IUpred-long	IUPred long predictions	Pred
IUpred-short	IUPred short predictions	Pred
Preds	All predictors	Cons, Pred
Nodisprot	Full MobiDB consensus without DisProt	Cons, Exp, Pred
Noideal	Full MobiDB consensus without IDEAL	Cons, Exp, Pred
Nomanual	Full MobiDB consensus without manually curated data (DisProt and IDEAL)	Cons, Exp, Pred
Full	Full MobiDB consensus (all sources)	Cons, Exp, Pred

Table 5.3: Overview of the disorder definitions used. The labels used for disorder data sources throughout the paper are defined. The type column lists whether the source contains experimental information (Exp), predictions (Pred) or consensus (Cons).

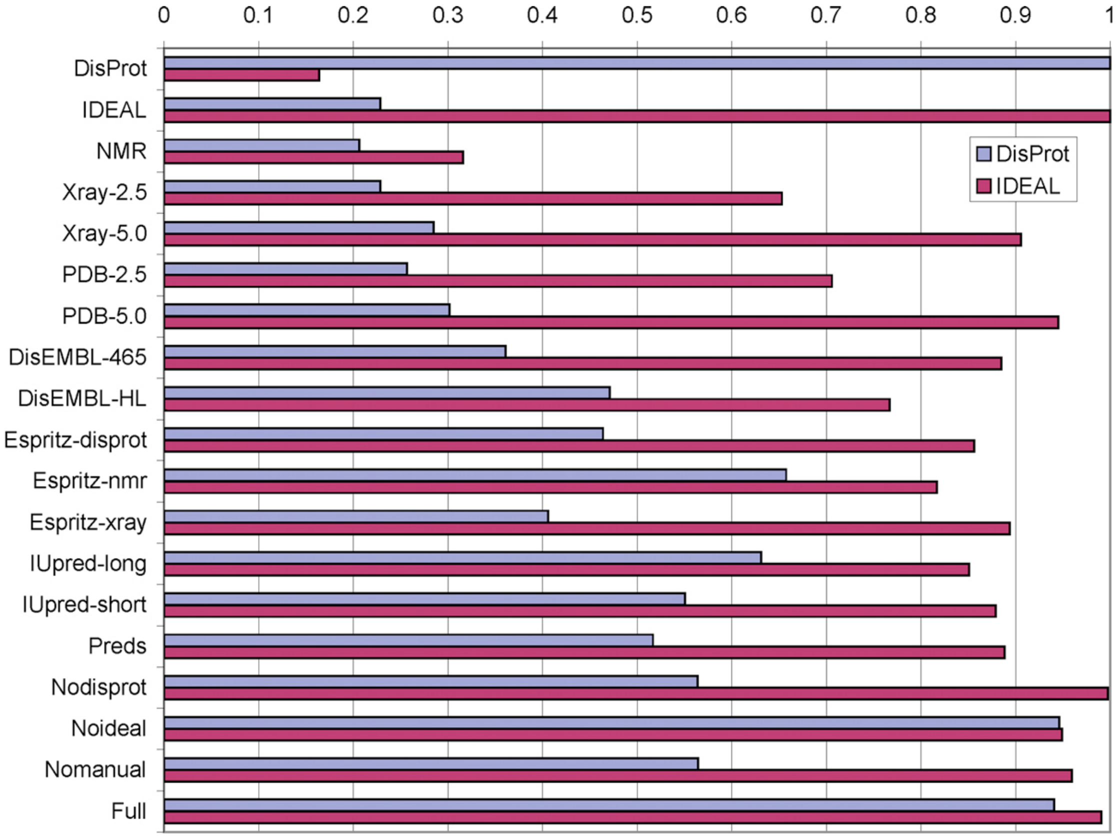
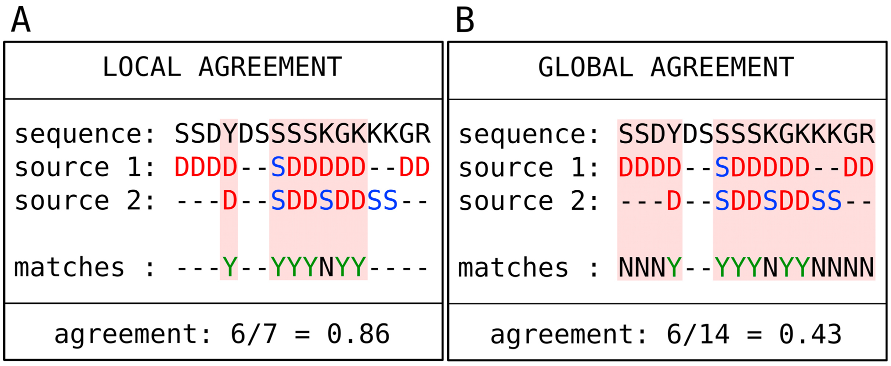


Figure 5.2: Agreement of disorder sources and consensus with the DisProt and IDEAL annotations. Results are shown as agreement of each data source with the DisProt and IDEAL reference datasets. Notice the difference between DisProt and IDEAL, and how the latter is mainly similar to PDB information. Overall, it is interesting to see that IDEAL is much easier to replicate than DisProt, suggesting a relative lack of long disordered regions in the former.

level further, by showing the level of agreement between each possible combination of annotations, by building ”restricted” consensus annotations that include only a subset of the disorder information sources. Three different ways to calculate the agreement are defined in Figure 5.3, defining whether gaps in one annotation should be considered or not. Figure 5.4 shows the results for these definitions, which are broadly similar with the baseline agreement (Figure 5.4c) being perhaps the most representative. Most sources are rather similar, with the notable exceptions of DisProt, PDB-nmr and, to a lesser degree, IDEAL. The former two have a low agreement with the other sources and among themselves, reinforcing the notion of their unique contribution to disorder. IDEAL confirms its rather good agreement with PDB-xray data.

5.4.4 Proteome analysis

As an example of the potential of MobiDB, we present an analysis of disorder in 13 eukaryotic model organisms (see Figure 5.5). Our analysis is in broad agreement with previous data suggesting a correlation between organism complexity and disorder [17]. The overall fraction of disordered residues is lower than in previous publications, with an average or only 15% for the disorder consensus. Due



```
A                                   B
LOCAL AGREEMENT                     GLOBAL AGREEMENT

sequence: SSDYDSSSSKGKKKGR          sequence: SSDYDSSSSKGKKKGR
source 1: DDDD--SDDDDD--DD          source 1: DDDD--SDDDDD--DD
source 2: ---D--SDDSDDSS--          source 2: ---D--SDDSDDSS--

matches : ---Y--YYYNYY----          matches : NNNY--YYYNYYNNNN

agreement: 6/7 = 0.86               agreement: 6/14 = 0.43
```

Figure 5.3: Agreement definitions. Schematic representation of two alternative agreement definitions between two disorder annotation sources mapping to the same sequence stretch. In the source lines, D is used for disorder and S for structured. The match line shows the agreement between sources with Y (yes) used for agreement and N (no) for disagreement.

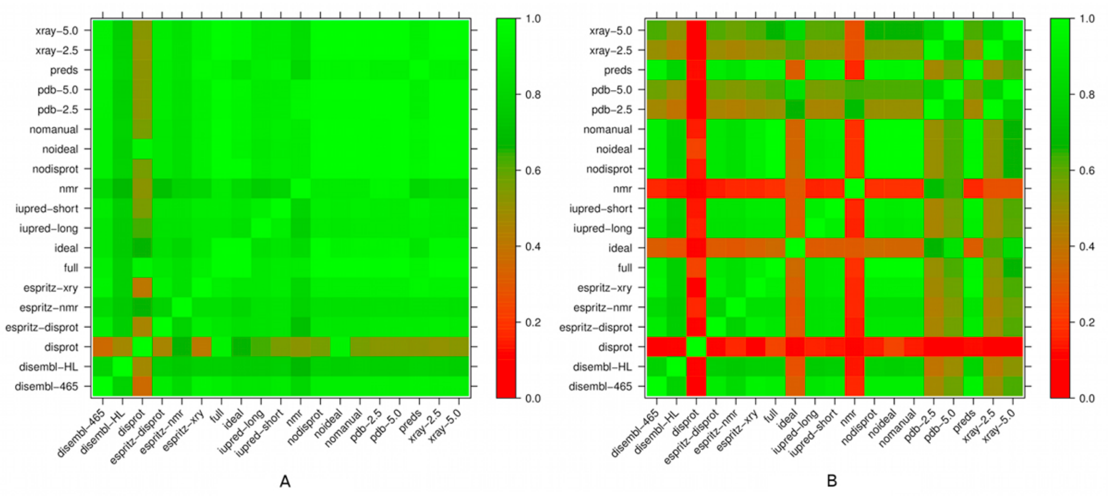


Figure 5.4: Agreement among disorder sources. The agreement among pairs of disorder data sources is plotted from red (0.0) to green (1.0) using the two definitions local (A) and global (B) agreement from Figure 5.3. Even though the different sources have a high level of local agreement, the number of times two sources annotate at the same time is relatively low. The global agreement evidences this by showing a drastic drop in agreement when the situation of one of the sources annotating, and the other not, is considered a negative. The MobiDB consensus aims at combining different sources to maximize the coverage when annotating a reference sequence, trying to overcome this issue.

to different disorder sources covering slightly different sequence stretches, effectively cancelling out each other, this estimate should be considered a lower bound only. Somewhat surprisingly, a few simple organisms are predicted to have more disordered residues than more complex ones (Figure 5.5). A similar observation was recently made for a larger set of eukaryotic proteomes, leading the authors to speculate about an organism's lifestyle [83]. In any case, MobiDB provides the means necessary to easily carry out proteome-wide comparisons of disorder distributions.

5.4.5 Single protein analysis

For the use case of analysing a single protein MobiDB provides an interactive user interface. In this interface the user can customize the resulting consensus by se-

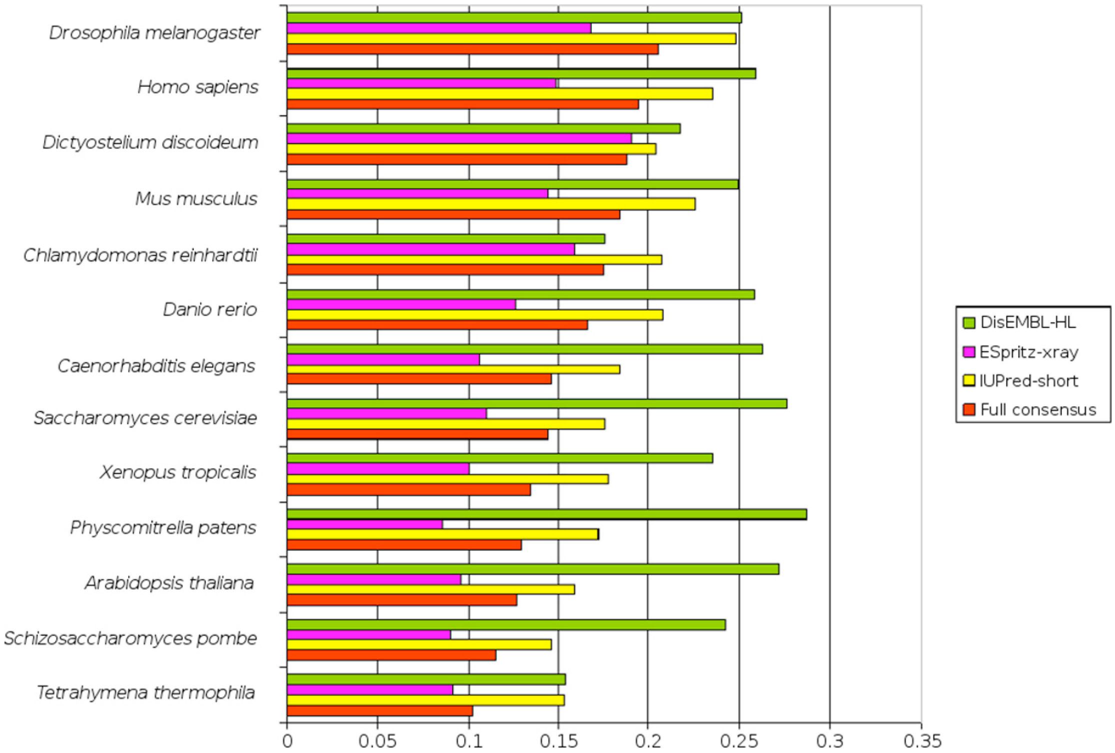
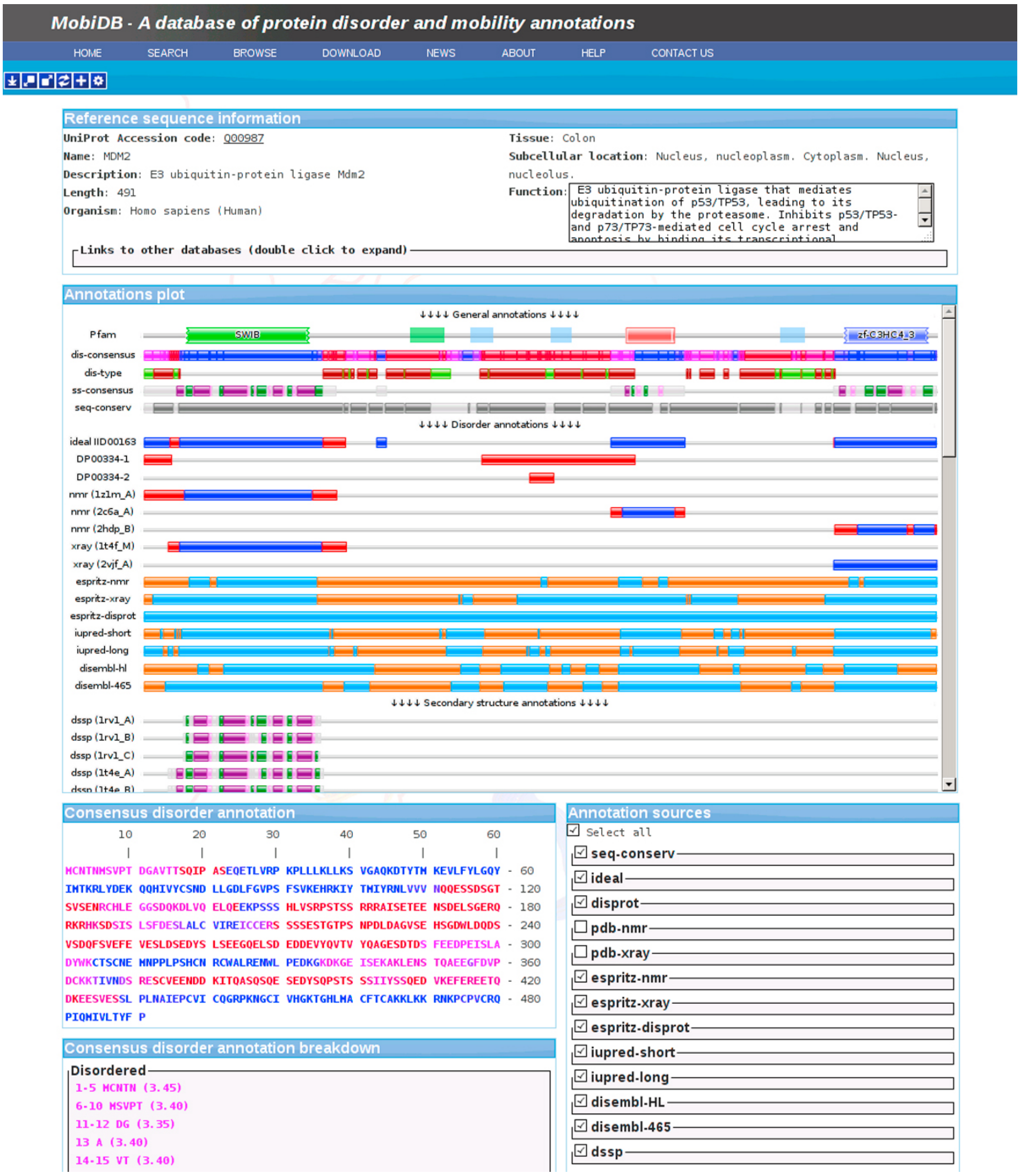


Figure 5.5: Percentage of disordered residues on all proteins encoded by a selection of model organisms. The fraction of disordered residues is recorded for a group of model organisms according to the full MobiDB consensus and three chosen predictors. Data is and sorted increasingly according to MobiDB consensus. Notice how DisEMBL-HL is the only predictor to break the broad trend for more disorder in higher organisms. The effect is however smoothed by the MobiDB full consensus. See main text for an explanation.

lecting only those sources of information relevant to the analysis being performed. In the case of the E3 ubiquitin-protein ligase Mdm2 (Figure 5.6), experimental annotations are available from the IDEAL and DisProt databases, and from PDB X-Ray and NMR experiments. None of these, however, provide coverage for the full protein sequence. The MobiDB consensus provides the means to elegantly combine the available annotations, allowing the user to quickly understand how disorder is distributed in his protein of interest.

5.4.6 Conclusions

We have presented a detailed description of MobiDB, a database of experimental and predicted disorder in proteins, and its main features, disorder consensus and weighting. The database is highly modular and extensible, allowing inclusion of a growing amount of information. A comparison between different disorder data sources highlights how the MobiDB consensus captures the main features of intrinsic disorder and correlates well with the manually curated datasets from DisProt and IDEAL. In more detail, the DisProt curation is best approximated with a combination of disorder predictors, allowing a robust estimation of the presence of disorder in eukaryotic genomes, roughly confirming the higher incidence of disorder in higher organisms. In the future we plan to expand MobiDB to include additional information for all known proteins, both from experimental sources and new predictors, with the goal of making it an increasingly useful, centralized source of data for intrinsic disorder research.


MobiDB - A database of protein disorder and mobility annotations
HOME
SEARCH
BROWSE
DOWNLOAD
NEWS
ABOUT
HELP
CONTACT US

Reference sequence information

UniProt Accession code: Q00987

Name: MDM2

Description: E3 ubiquitin-protein ligase Mdm2

Length: 491

Organism: Homo sapiens (Human)

Tissue: Colon

Subcellular location: Nucleus, nucleoplasm. Cytoplasm. Nucleus, nucleolus.

Function: E3 ubiquitin-protein ligase that mediates ubiquitination of p53/TP53, leading to its degradation by the proteasome. Inhibits p53/TP53- and p73/TP73-mediated cell cycle arrest and apoptosis by binding its transcriptional

Links to other databases (double click to expand)

Annotations plot

Consensus disorder annotation

```
        10        20        30        40        50        60
         |         |         |         |         |         |
MCNTNMSVPT DGAVTTSQIP ASEQETLVRP KPLLLKLLKS VGAQKDTYTM KEVLFYLGQY - 60
IMTKRLYDEK QQHIVYCSND LLGDLFGVPS FSVKEHRKIY TMIYRNLVVV NQQESSDSGT - 120
SVSENRCHLE GGSDQKDLVQ ELQEEKPSSS HLVSRPSTSS RRRAISETEE NSDELSGERQ - 180
RKRHKSDSIS LSFDESLALC VIREICCERS SSSESTGTPS NPDLDAGVSE HSGDWLDQDS - 240
VSDQFSVEFE VESLDSEDYS LSEEGQELSD EDDEVYQVTV YQAGESDTDS FEEDPEISLA - 300
DYWKCTSCNE MNPPLPSHCN RCWALRENWL PEDKGKDKGE ISEKAKLENS TQAEEGFDVP - 360
DCKKTIVNDS RESCVEENDD KITQASQSQE SEDYSQPSTS SSIIYSSQED VKEFEREETQ - 420
DKEESVESSL PLNAIEPCVI CQGRPKNGCI VHGKTGHLMA CFTCAKKLKK RNKPCPVCRQ - 480
PIQMIVLTYF P
```

Consensus disorder annotation breakdown

Disordered

1-5 MCNTN (3.45)

6-10 MSVPT (3.40)

11-12 DG (3.35)

13 A (3.40)

14-15 VT (3.40)

Annotation sources

☑ Select all

☑ **seq-conserv**

☑ **ideal**

☑ **disprot**

☐ **pdb-nmr**

☐ **pdb-xray**

☑ **espritz-nmr**

☑ **espritz-xray**

☑ **espritz-disprot**

☑ **iupred-short**

☑ **iupred-long**

☑ **disembl-HL**

☑ **disembl-465**

☑ **dssp**

Figure 5.6: The E3 ubiquitin-protein ligase Mdm2 in the MobiDB interactive user interface. The interactive user interaface of MobiDB allows the user to build a customized consensus based on annotations of interest. The example shows how the database facilitates the easy integration of different data sources to maximize coverage of disorder annotations. In the example, annotations extracted from the IDEAL and DisProt databases and from X-ray and NMR experiments of the PDB are complemented by predictions to provide accurate annotations covering the full extent of the protein's sequence.

6. The many faces of intrinsic protein disorder

6.1 Introduction

6.1.1 Structure and function of intrinsically disordered proteins

As has already been discussed in previous chapters, intrinsic protein disorder has been a very active field of research over the last few decades. Experimental determination of the phenomenon, however, is rare, and it often involves conditions that are not natural to the protein. For example, large complexes may be broken down into parts to facilitate X-Ray diffraction experiments. These alterations can likely disrupt a protein's natural structure and/or function. It is possible that results influenced by such modifications have led us to the misinterpretation of many cases currently thought to be evidence of intrinsic protein disorder. In fact, it has recently been suggested that intrinsic protein disorder may be largely caused by artifacts of current methods for protein production [99]. Instead of being intrinsically disordered and functional while in that state, the authors of the study suggest that, when in vivo, most disordered proteins are actually *proteins waiting for a partner* (PWPs). A PWP's disordered state is only transient and non-functional, and it will undergo folding upon binding its partner. Upon reaching its folded state, it will then be able to perform its function. The concept of folding upon binding is by no means novel, and it has been extensively documented [100–103]. The ability of intrinsically disordered proteins to promiscuously bind many partners allows them to act as hub proteins, which are often involved in cellular regulatory and feedback mechanisms [103–105]. Independently of the chosen name, and of the possible existence false positives due to the previously mentioned artifacts, folding upon binding seems to play an important role in the function of proteins that are highly relevant from the biological point of view.

6.1.2 The current situation of intrinsic disorder annotations

Currently available intrinsic protein disorder annotations can be grouped into three categories: experimental, indirect, and predicted. As is usually the case, the higher the quality of the data, the harder it is to obtain. The quality/coverage tradeoff between the different categories is illustrated in Figure 6.1.

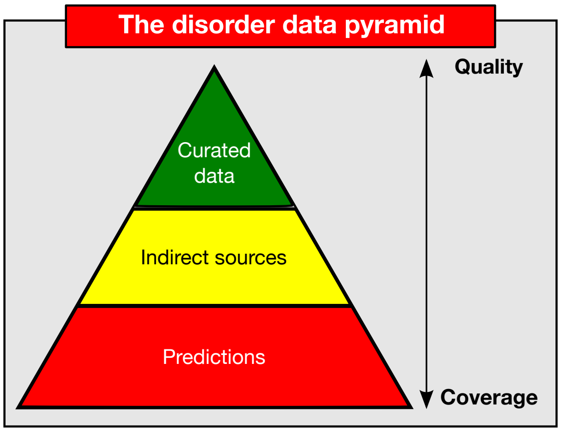


Figure 6.1: The intrinsic protein disorder data pyramid. Available intrinsic protein disorder data sources present a serious tradeoff between quality and coverage. While, on one extreme, well documented annotations from literature are counted in the hundreds, automated predictors allow us to obtain annotations for any protein with a known sequence. For comparison purposes, the UniProt database features -at the time of writing- almost 50 million protein sequences.

Experimental annotations

The de-facto standard for experimental annotations of disorder is the DisProt database [40]. DisProt is a highly valuable resource, containing curated annotations manually extracted from literature. A quick analysis of its contents, however, makes it evident that both the frequency of its updates and the number of new annotations on each release are seriously lagged when compared to the growth of protein data repositories such as the Protein Data Bank [39] or the SwissProt database [78].

Indirect annotations

In order to make up for the limited coverage of experimentally obtained annotations, many research efforts turn to indirect methods for the generation of datasets. The most popular of these methods is X-Ray diffraction. When a protein structure is resolved by X-Ray diffraction, there are often regions where its backbone can not be accurately observed. One of the possible reasons for this is the fact that the region is disordered, and its position is therefore not easily determined by a method that has been devised for the observation of stable protein structures [87, 99]. An advantage of using this definition of disorder is that one can take advantage of the resources from the Protein Data Bank, and thus obtain a dataset of tens of thousands of proteins. A great disadvantage, however, is that missing regions in X-ray diffraction experiments are rather short, since by definition an X-Ray experiment strives to resolve as much structure as possible. The presence of very long, not observed regions, would be considered evidence of an unsuccessful experiment. Figure 6.2 provides an overview on the distribution of lengths that can be obtained by extracting the missing regions from PDB X-Ray experiments. It is clear that potentially interesting, longer disordered regions are obtained much less frequently with this approach. A second disadvantage is the fact that missing regions could be due to causes other than naturally occurring intrinsic disorder (for example low resolution of the X-Ray experiment, high B-factor, modifications of conditions (e.g. pH, temperature, or pressure) during the experiment) [99].

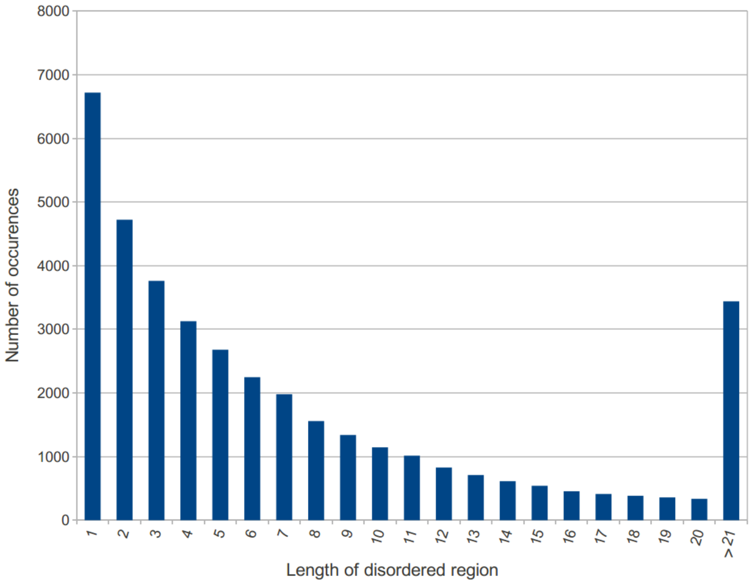


Figure 6.2: Length distribution of not observed regions in PDB structures, clearly showing the over-representation of short segments.

Predicted annotations

By using the previously described experimental and/or indirect annotations as training data, many automated predictors have been developed. Based mainly on mathematical/statistical models (e.g. machine learning methods), these tools present the immediate advantage of expanding the available dataset of annotations to all proteins of known sequence. As shown in Chapter 2 and Chapter 3, these predictors can very accurately detect disorder as defined by the available experimental and indirect annotations. It is worth mentioning, however, that the limitations of these annotations (scarcity of the experimental ones and fuzziness of the indirect ones) are inherently present in the predictors' output. For this reason, while they represent an extremely useful tool to guide research, predicted annotations should always be thought of as an approximation to disorder, and not as a precise determination of it.

6.1.3 Dissecting available intrinsic protein disorder annotations

While analysing some entries from the original release of the MobiDB database, which provides a quick overview of all available disorder annotations for any given protein (see Chapter 4), we noticed several cases where a protein's disorder annotation from the DisProt database would be in conflict with a structure from the

PDB. While DisProt would annotate a certain region as disordered, one could find a deposited structure on the PDB which covered the same region. Our hypothesis was that these conflicts could be caused by annotations of proteins that undergo folding upon binding: while DisProt would most likely be annotating the unfolded conformation, the PDB structure would have captured its bound and folded state. In order to test this hypothesis, we set out to analyse all cases where DisProt and the PDB exhibit such a conflict. By reviewing the literature associated with the proteins exhibiting this behaviour, we hoped to gain some understanding on the types of disorder that can be extracted from currently available disorder annotations, and to understand whether the observed conflicts could be an indication of folding upon binding.

6.2 Materials and methods

In order to obtain an overview of the relationship between DisProt and PDB annotations, we started by plotting the percentage of agreement between DisProt "disorder" annotations and PDB "missing residue" annotations in the 287 cases where there was at least one residue in agreement. Results are shown in Figure 6.3. Out of all the cases with agreement, in 42 of them the agreement between DisProt and PDB annotations reached 100%. Upon reviewing the cited literature we found that, in the case of those 42 proteins, DisProt takes as its source of annotations publications that report missing residues from a crystallography experiment, thus effectively incorporating an annotation that is also part of the "indirect" dataset (itself, as mentioned before, made up of data obtained from PDB structures). In 50% of these cases the DisProt entries feature a comment explaining that the annotations are "based solely on missing electron density in the Protein Data Bank". In the other half of the cases no such clarification is made. Amongst these overlapping entries with no clarifying comment, some of them feature references to intrinsic protein disorder in the articles describing the X-Ray experiments (e.g. DP00235, DP00407). For the rest, no mention of intrinsic protein disorder is made (e.g. DP00248, DP00252, DP00422). Since by definition these perfectly agreeing DisProt entries are already present in the X-Ray dataset, we decided to exclude them entirely from all further analysis[1].

In addition to entries with perfect overlap, we additionally removed: entries where the sequence identity between the DisProt sequence and the PDB sequence was below 95%; entries that have been replaced in UniProt and thus can not reliably link the DisProt and PDB annotations; and entries mapped to a UniProt chain and thus not independently annotated. Figure 6.4 provides a detailed view on the number of used and ignored entries. The final dataset, after filtering, contained 566 entries.

The next step was to perform a similar analysis for the case of conflicting anno-

[1] Given potential differences in the numbering of residues between the PDB and DisProt, it is more than likely that entries that show an agreement slightly below 100% will also be based solely on X-Ray experiments. For the sake of simplicity, however, we decided to exclude only those entries showing 100% agreement.

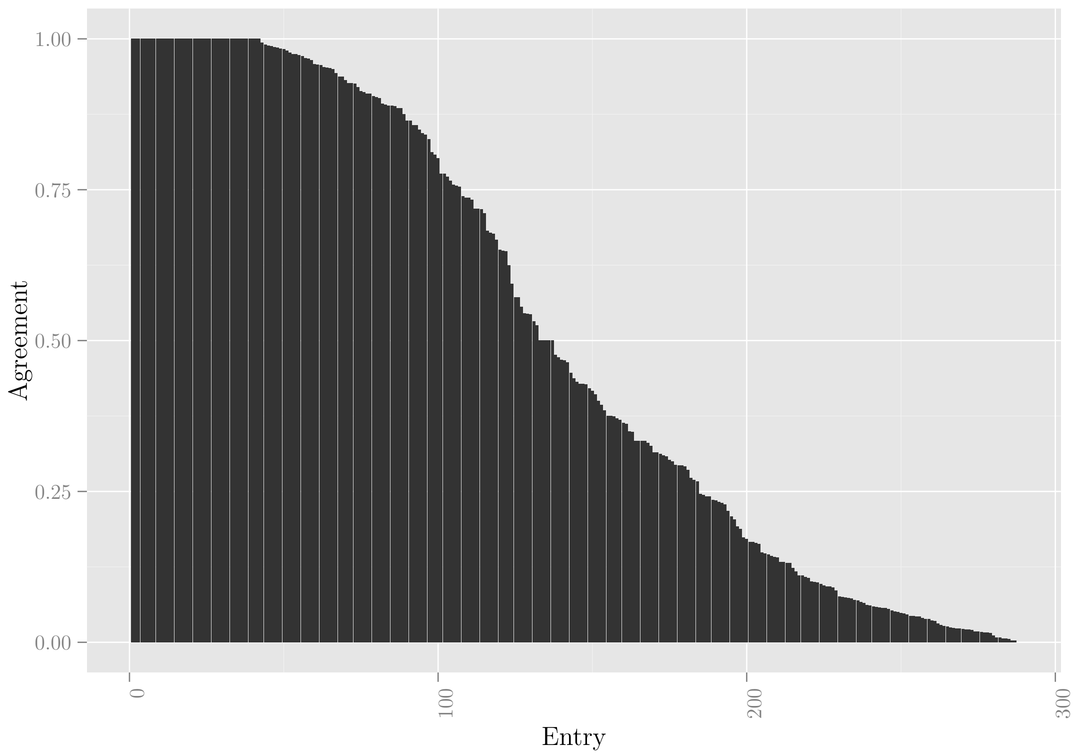


Figure 6.3: Agreement between annotations in DisProt and the PDB. For those entries that feature both DisProt (disorder) and PDB (missing residues) annotations, the figure shows how well these two sources agree. An agreement of 100% means that every single residue considered as disordered by DisProt, is missing from the X-Ray experiment.

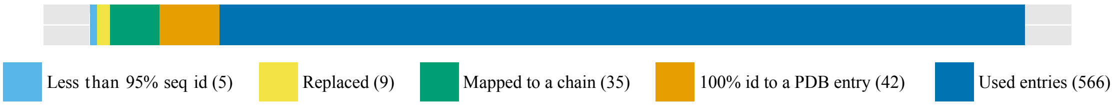


Figure 6.4: Detail of ignored and used entries. Redundant entries between datasets, or entries that can not reliably be mapped between them, were removed from the dataset.

tations between DisProt and the PDB. The results are shown in Figure 6.5. Over 300 entries feature at least some conflict between DisProt and PDB annotations. Under the assumption that interesting folding upon binding cases would occur at sequence regions with lengths that could accommodate functional domains, and in order to keep the number of entries manageable, we decided to include in the following steps only those entries with at least 30% of their sequences in conflict.

6.3 Results and discussion

After reviewing the literature referenced from each DisProt entry (as well as additional publications in case the cited source did not provide enough information), we were able to identify the intrinsic disorder related characteristics for each conflicting region. By grouping the entries into categories and subcategories, we were able to generate a simple classification tree, shown in Figure 6.6.

The simple classification obtained provides a starting point for the analysis

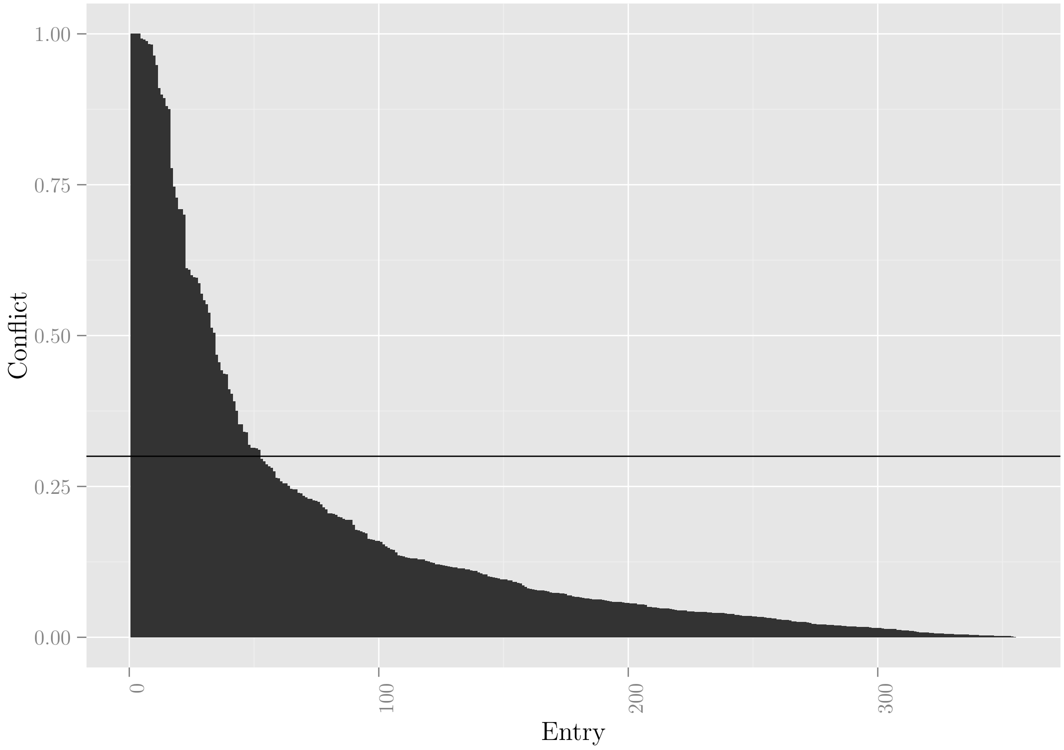


Figure 6.5: Conflict between DisProt and the PDB. For those entries that feature both DisProt (disorder) and PDB (missing residues) annotations, the figure shows how much these sources are in conflict. A conflict of 100% means that every single residue considered as disordered by DisProt, is shown as being part of a well-defined structure in at least one PDB structure. Entries whose conflict surpasses the horizontal line have more than 30% of their sequence's residues in conflict.

of disorder in a more specific manner, steering away from the usual one-size-fits-all approach. The following sections describe the characteristics of the members assigned to each category.

6.3.1 Non-physiological conditions

This category includes those entries that we consider as non-relevant from the point of view of a biological analysis. It contains 8 entries.

Experimental conditions

Entries in this category have been subjected to experiments under conditions that differ from those normally encountered in a living organism. High temperature or pressure, and low pH are the most commonly found situations. As an example, DisProt entry DP00303 documents Myoglobin as being disordered based on an experiment performed at a pressure of 3000 bar. Acknowledging the accuracy of the annotation with regards to what has been reported in the citing article, we would argue that the inclusion of this annotation would work in detriment of any analysis done towards the understanding of intrinsic protein disorder in living organisms.

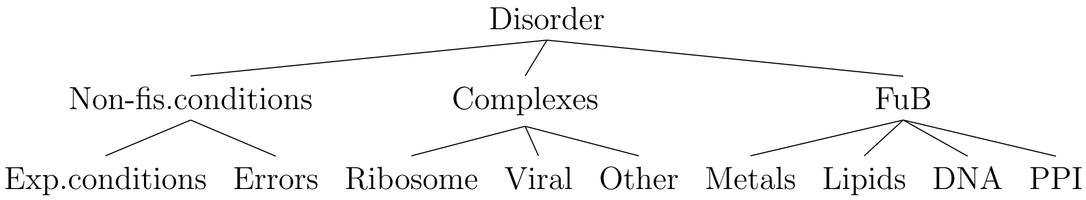


Figure 6.6: Classification of the 52 conflicting entries into a simple category/subcategory scheme, based on the type of disorder reported in literature.

Annotation errors

These entries represent simple errors when transferring the reported annotations from the original publication. An example is DisProt entry DP00607, where Frataxin is reported as being disordered on the c-terminal half. The original report states that Frataxin is disordered in its n-terminus.

6.3.2 Complexes

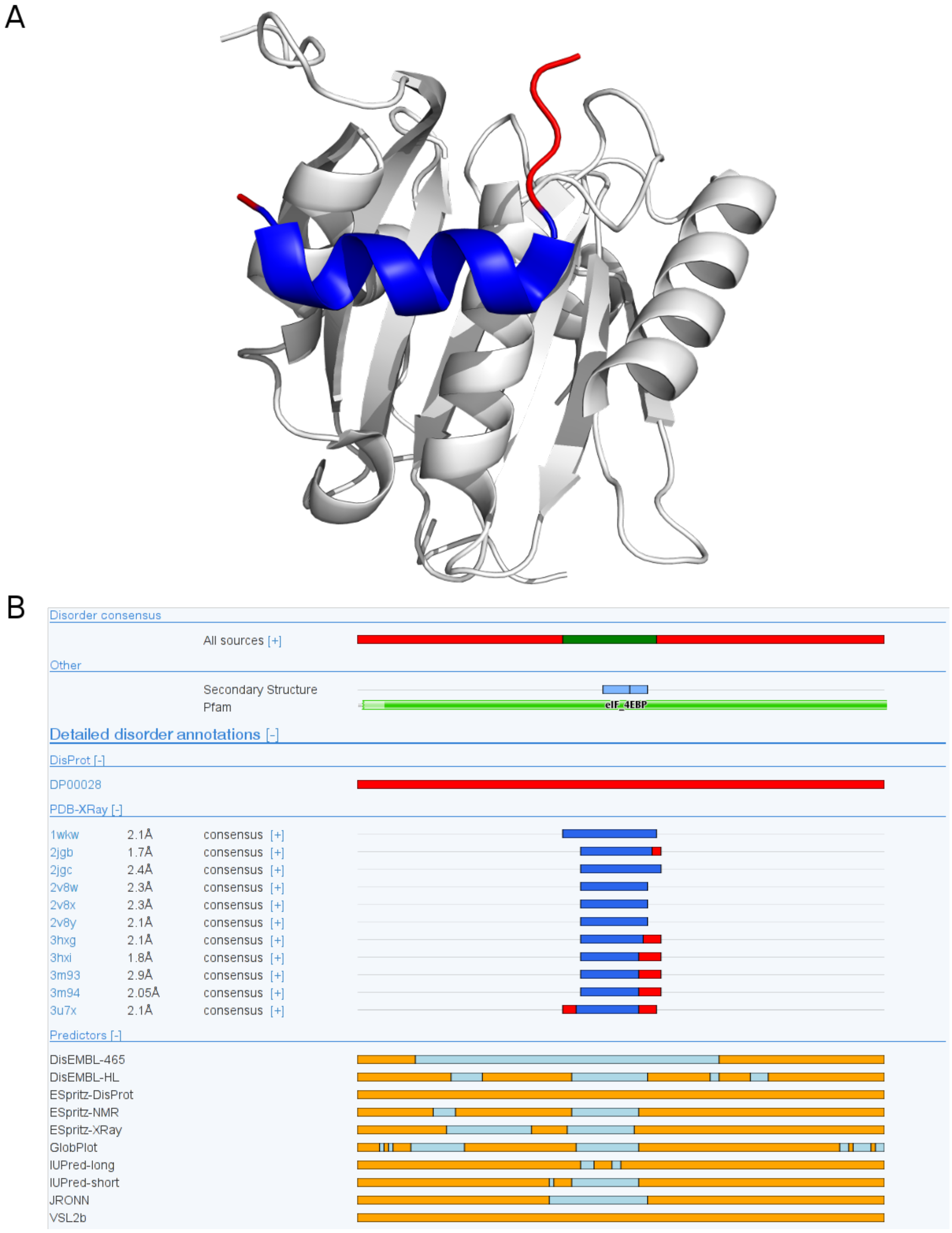
The 15 entries under the "complexes" category are documented to become disordered when they are studied in isolation from the complex to which they belong. Our subset mainly contained examples from the Ribosomal complex and viral complexes, with a few entries corresponding to other complexes. Even if this category would seem to have much in common with the "protein-protein interaction" subcategory under "folding upon binding", we have decided to assign it to its own category based on the fact that its members are part of well-known complexes with many components, are rarely thought of independently.

6.3.3 Folding upon binding

Entries in this section (29 in total) have been documented as existing in an intrinsically disordered state until binding a partner. These partners can range from simple metal ions to complex molecules, like other proteins or even nucleic acids. DisProt entry DP00028 provides an outstanding example of a protein-protein interaction induced folding. PDB structure 2jgb, shown in Figure 6.7, features the otherwise intrinsically disordered 4EBP1 protein having acquired structure upon binding its partner.

6.4 Conclusions

Intrinsic protein disorder annotations have been used extensively for different types of analyses over the course of the last few decades. These efforts, however, usually treat disorder as a single phenomenon. Starting from the hypothesis that intrinsic protein disorder actually encompasses a variety of related phenomena, we implemented a simple workflow for data analysis with the goal of finding subgroups corresponding to different disorder types. We demonstrated that as simple a feature as the conflict of two independent data sources can give out a signal indicating a feature as relevant as folding upon binding. This categorisation of intrinsic protein disorder should be easy to expand in order to cover a larger portion of the available annotations. More specific tools and methods, aware of the existence of the different categories, should in turn allow us to better understand the particulars of each subgroup, and therefore of intrinsic protein disorder.



Disorder consensus

All sources [+]

Other

Secondary Structure

Pfam

Detailed disorder annotations [-]

DisProt [-]

DP00028

PDB-XRay [-]

1wkw	2.1Å	consensus [+]
2jgb	1.7Å	consensus [+]
2jgc	2.4Å	consensus [+]
2v8w	2.3Å	consensus [+]
2v8x	2.3Å	consensus [+]
2v8y	2.1Å	consensus [+]
3hxg	2.1Å	consensus [+]
3hxi	1.8Å	consensus [+]
3m93	2.9Å	consensus [+]
3m94	2.05Å	consensus [+]
3u7x	2.1Å	consensus [+]

Predictors [-]

DisEMBL-465
DisEMBL-HL
ESpritz-DisProt
ESpritz-NMR
ESpritz-XRay
GlobPlot
IUPred-long
IUPred-short
JRONN
VSL2b

Figure 6.7: 4EBP1 structured while bound to its partner. Part A on top shows the PDB structure of 4EBP1 (structured regions in blue, disordered in red) having acquired structure when bound to its partner. Part B shows the corresponding annotations in the MobiDB database (`http://mobidb.bio.unipd.it/entries/Q13541`). Notice how all PDB experiments resolve the central α-helix of 4EBP1, while the rest of the protein is either not observed (red pieces of sequence) or completely absent. DisProt annotates instead the unbound, fully disordered state. The combination of both sources generates a "conflict" region, shown in green in the MobiDB consensus annotation.

7. RUBI: Rapid proteomic scale prediction of lysine ubiquitination and factors influencing predictor performance

This chapter was first published in Walsh I, Di Domenico T, Tosatto SCE (2013) RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. Amino Acids: 1–10.

7.1 Summary

Post-translational modification of protein lysines was recently shown to be a common feature of eukaryotic organisms. The ubiquitin modification is regarded as a versatile regulatory mechanism with many important cellular roles. Large scale datasets are becoming available for H. sapiens ubiquitination. However, using current experimental techniques the vast majority of their sites remain unidentified and in silico tools may offer an alternative. Results: Here, we introduce RUBI a sequence-based ubiquitination predictor designed for rapid application on a genome-scale. RUBI was constructed using an iterative approach. At each iteration important factors which influenced performance and its usability were investigated. The final RUBI model has an AUC of 0.868 on a large cross-validation set and is shown to outperform other available methods on independent sets. Predicted intrinsic disorder is shown to be weakly anti-correlated to ubiquitination for the H. sapiens dataset and improves performance slightly. RUBI predicts the number of ubiquitination sites correctly within three sites for ca. 80% of the tested proteins. The average potentially ubiquitinated proteome fraction is predicted to be at least 25% across a variety of model organisms, including several thousand possible H. sapiens proteins awaiting experimental characterization. RUBI can accurately predict ubiquitination on unseen examples and has a signal across different eukaryotic organisms. The factors which influenced the construction of RUBI could also be tested in other post-translational modification predictors. One of the more interesting factors is the influence of intrinsic protein disorder on ubiquitinated lysines where residues with low disorder probability are preferred.

7.2 Introduction

Post Translational Modifications (PTMs) contribute to the complexity of an organism, bestowing multiple protein functions on a single encoding gene [106]. Lysine ubiquitination is a reversible PTM found in all eukaryotic cells. After translation, a protein can be modified by covalent bonding of ubiquitin, a small and highly conserved regulatory protein. The enzymatic process for ubiquitination involves a three step sequential process between the E1, E2 and E3 enzymes [107]. The bonding can be a single ubiquitin molecule (mono-ubiquitination) or multiple chains (poly-ubiquitination) resulting in a wide variety of cellular processes. One of the earliest functional associations was proteasomal degradation [108]. Although currently the ubiquitin system is regarded as a more versatile regulatory mechanism [109]. For example, Lysine63-linked poly-ubiquitin chain is involved in both DNA repair and endocytosis [110]. Mono-ubiquitination can also modify a protein to perform various functions ranging from membrane transport to transcriptional regulation [111]. In contrast, deregulation of ubiquitin is implicated in cancer [112] and neurogenerative disorders [113]. Given these important functions, targeting the multi-functional role of ubiquitination and its pathway can be of massive therapeutic benefit [114, 115]. In vitro tools are difficult to develop for ubiquitination because the modification is large (ca. 8k Da) and the modified protein has a rapid turnover [116]. Recently, more robust and accurate techniques, such as global mass spectrometry, are now able to process thousands of sites [117]. However, these experimental techniques often vary, capturing different protein properties and experimental detection is hampered by the diversity of the type of ubiquitin chain [118]. Computational tools are thus still needed to fill the gap. In order to guide in vitro experiments, these tools should be fast, with high specificity (minimal false positive rate) and significant sensitivity (true positive rate). Efficient and highly specific predictors are also needed for hypothesis testing. For example, ubiquitination crosstalks with other PTMs resulting in a sophisticated coding scheme for different protein functions [119], hence combinations of ubiquitination and other PTM in silico tools may shed light on this phenomenon. Computational prediction of phosphorylation is well studied and may be used to describe the problem of ubiquitination data shortage by analogy. Phosphorylation prediction methods have been initially constructed in a general [120] and later enzyme-specific manner [121]. However, data deficiency becomes an issue when applying an enzyme-specific approach. Moreover, if the data is split further based on individual organisms (assuming sites are different across species) the data becomes even more deficient, making highly specific prediction tools difficult to implement. This usually results in two computational prediction modes, either enzyme-specific but organism independent or enzyme-independent but organism specific tools. In this work we focus on an enzyme- independent but human specific tool. Computational prediction tools are only recently emerging. UbiPred [122] uses a Support Vector Machine (SVM) approach to predict ubiquitination sites on 105 mainly S. cerevisiae proteins. Another method, UbPred [123], based on the random forest algorithm was trained on a dataset of motifs from S. cerevisiae. The algorithm is fast allowing for large scale analysis and was used to show enrichment in various molecular functions and a preference for proteins with very short half-lives. Two more recent methods [124, 125] highlight the use of a sophisticated encoding scheme and feature selection respectively. Recently, data and predictors have become available for mammalian sites. An updated version of [124] was retrained for H. sapiens sites [126]. UbiProber [127] incorporates both general and species- specific ubiquitination sites using key motif positions and amino acid residue features. It was constructed on three species, H. sapiens, M. musculus and S. cerevisiae, showing interesting performance in particular for cross species predictions. All of these methods were motif based, meaning patterns of N amino acids surrounding each lysine were used for learning. A recent independent assessment of PTM predictors [128] has shown 9 out of 11 predictors behave worse than random on unseen data. In this paper, we try to determine factors which may influence the performance, not as a criticism of other methods but simply as appropriate factors for the model being constructed. We have developed a novel method, RUBI, trained on a large dataset of over 10,000 experimentally determined H. sapiens ubiquitination sites [129]. RUBI (Rapid UBIquitination detection) is a fast predictor aimed for proteomic-scale applications with high specificity and significant sensitivity. We show that it accurately mimics the results of high-throughput mass spectrometry techniques on unseen data.

7.3 Methods

7.3.1 Datasets

Lysine ubiquitination is a binary classification problem. However, while the positive set can be defined easily from high-throughput datasets, negative lysines become a thorny issue [130]. Assignment of negative cases can only be tentative, as new experimental evidence may reveal them to be ubiquitinated. For this reason they are often called background lysines. Performance on PTM classification depends on the amount of redundancy in the underlying data and the availability of high quality annotated data [121]. Here 11,054 positive H. sapiens motifs were taken directly from high resolution mass spectrometry data on 4,273 proteins [129], where for each site position in sequence and local context of length 13 was given. The experimental technique uses immuno-enrichment by anti-di-Glycine antibody and it was this technique which we try to reproduce. The learning set was constructed from this data, while the independent set was constructed from another database, their construction is described in the following and summarized in Table 1. The most clear observation is the massive imbalance of positives to background lysines when using full sequences (e.g. ratio 1:16 for Seq40). Seq40: considers full human protein sequences taken from [129]. Each sequence was reduced to a maximum pairwise identity of 40% with CD-HIT [131]. For this data learning proceeded on the entire sequence (see section 2.2). When testing performance, positive lysines were experimentally validated ones while the background was defined as those lysines not found ubiquitinated in the same study. Similar positive and background extraction was done in [125, 127].

Motif40: considers motifs taken from [129]). Motifs surrounding lysines of length 13 were extracted directly from Seq40. The diversity at the motif level will be analyzed in the results as it is an important factor. For this data learning proceeds on the site motifs (see section 2.2). The ratio of positives to negatives is identical to Seq40. Independent benchmark: This was constructed from Phosphosite plus, an interactive database of manually curated PTMs [132]. Upon model construction, it contained 39,037 ubiquitination sites mainly for M. musculus and H. sapiens. CD-HIT was used to remove pairs of sequences with more than 40% pairwise sequence identity to each other and to the training data. After this, only 22 out of 2563 positive sites (sites of 13 residues surrounding lysines) had $geq9$ residues in common with the training positives. Phosphosite plus annotates ubiquitination in three sub-classes, derived from different experimental protocols. Low-throughput (LTP) data is taken from the literature. High-throughput mass spectrometry data can be distinguished as either taken from multiple sources in the literature (PUB) or unpublished and generated at Cell Signaling Technology, Inc. (CST). Our main independent performance (generalization) will be evaluated on PUB. However, some interesting observations about CST and LTP will also be shown.

Dataset	Proteins	Residues	Lysines	Positives	Background
Motif40	3,705	2,014,272	154,944	9,237	145,686
Seq40	3,705	2,460,023	154,944	9,237	145,686
PUB	1,264	900,370	52,766	2,563	50,204
CST	2,103	1,687,456	108,022	3,768	104,254
LTP	91	70,103	4,500	201	4,299

Table 7.1: Distribution of residues and lysines in the datasets. Sets on white background are the learning sets while shaded in grey are the independent sets. Number of proteins, residues and lysines in the datasets. Positives and background refer to the lysine classification. The main independent dataset is published high-throughput mass spectrometry data from the literature (PUB).

7.3.2 Machine learning

Two different machine learning approaches were used to develop RUBI, SVMs [133] and bi-directional recurrent neural networks (BRNNs) [134]. SVMs were tested with four kernel functions: linear (1 parameter), polynomial (3 parameters), radial basis (2 parameters) and sigmoid (2 parameters). A grid search was used to determine the best SVM parameters for each kernel and the C parameter. We split the data into 10 random folds. Each split consisted of 80% for training parameters, 10% for validation and 10% for testing on unseen data. The validation set was used as an over-fitting flag since generalization on unseen data can be measured. Three values were used to flag over-fitting: low C parameter, large margin (i.e. low $\|w\|$ in SVM formulation, see [133] and sensitivity at 5% false positive rate. The former two are commonly known to affect generalization and the latter gives a measure of the generalization with a low false positive count. The input vectors were the encoded sequence motifs of length M centered on lysines (M=13 here). Assuming the central lysine is redundant information, the input vector for the

SVM was $[X(i-6),\dots,X(i-1),X(i+1),\dots,X(i+6)] \in \mathbb{N}^{252}$ where $X() \in \mathbb{N}^{21}$ and i was the location of the central lysine. Each of the 21 components in X() was sorted alphabetically by the amino acid symbol. One hot encoding was employed, i.e. a component was set to 1 for the corresponding amino acid symbol (e.g 1^{st} position set to 1 for alanine and so on) and the rest set to 0. When the motif was extracted from the N and C termini (i.e. $i < 7$ and $i > N-7$, respectively), the 21st position was set to one. After extensive testing the radial basis kernel outperformed the other alternatives (data not shown), only results for this will be described throughout the paper. BRNNs can be likened to an ensemble of three neural networks, learning the N-terminal sequence context, the sequence and the C-terminal sequence context respectively [134]. This important local context surrounding the lysine was learned and stored as hidden layers using two specialized neural networks. It is important to note that BRNNs capture local context and are unable to capture the entire sequence. Where regular neural networks and SVMs use a sliding window of predetermined size, BRNNs learn this context information through the recursive dynamics of the network. This reduces the number of parameters and extracts information implicitly from the surrounding local context. An identical BRNN was implemented as described in [77] [43], with one hot encoding of the amino acids as above.

7.3.3 Models

Several method variants were tested to develop RUBI, as summarized in Table 2. The main distinction was between motif based SVMs or full sequence based BRNNs with additional input. The background dataset was constructed with all residues present in the set of full sequences excluding known ubiquitination sites. This possibly included some lysines which can be ubiquitinated but for which experimental evidence is missing, i.e. experimental false negatives. The sequence information was then encoded either as a local motif or the full sequence. For the full sequence, learning proceeded on each residue, including non-lysines. For the motifs learning proceeded on the 13 motif fragment only. When calculating performance on unseen data only lysines were considered. Two additional information sources were tested: multiple sequence alignments and predicted intrinsic disorder. Multiple sequence alignments were calculated using three rounds of PSI-BLAST [42] with options -b 3000 -e 0.001 -h 1e-10 on the UniRef90 database. The frequency of each amino acid and gap frequencies in the multiple sequence alignment at a given position along the sequence was used instead of one-hot encoding. This has been previously shown to be useful for secondary structure prediction [135]. For intrinsic disorder, the probability of disorder at a sequence position was derived from ESpritz [77] and encoded as a new component to the input vector. Model construction was an iterative process where experiments such as site distribution, machine learning technology (SVM and BRNN) and alignments were tried with poor performers discarded. Finally, with the best surviving model, particular attention was paid to protein disorder and its relationship to ubiquitination. RUBI performance for each model was estimated using 10-fold cross validation. Both SVM and BRNN produce a prediction score and the threshold producing optimal decisions often depends on the ratio of positive to negative examples. For all models, the decision threshold was determined at low false positive rate on the training set. This decision threshold can thus be considered an extra learning parameter.

Acronym	Sequence	Additional
SVM-M40	Motif only	
BRNN-S40	All residues	
BRNN-S40+ali	All residues	Alignment
Rubi: best+disorder	All residues	Disorder

Table 7.2: Method definitions used in the paper.

7.3.4 Comparison to other methods and availability

In order to compare RUBI to the state of the art, we tried to download and/or request from the authors executables for other published methods. Some methods [124, 125] are available only as a server for single input sequences and had to be discarded. UbiPred [122] and UbPred [123] could be installed locally and will be used for comparison. Two recently published methods, UbiProber and hCKSAAP_UbSite, provided their training and testing data [126, 127]. The UbiProber and hCKSAAP_UbSite comparisons were important because it allowed us to compare our method to recent state-of-the-art human predictors. Four common accuracy measures are used in analogy to our previous work on disorder [77] sensitivity (sens), specificity (spec), Matthews correlation coefficient (MCC) and AUC (Area Under the sensitivity vs false positive rate Curve). RUBI is available both as a Linux executable for download and a web server able to process thousands of sequences at a time from the homepage. All datasets used throughout the paper can be found together with server, executable and online documentation from URL: http://protein.bio.unipd.it/rubi/.

7.4 Results

The goal for developing RUBI was to find a model with high sensitivity at high specificity. High specificity (or low false positive rate) is vital for confident determination of ubiquitination sites and vital for any algorithm aiding in vivo experiments [130]. The first concern was overestimation of performance due to similarity between fragments surrounding each lysine. To this end, we started by analyzing the distribution of lysine local context in the data and proceed iteratively.

7.4.1 Site diversity

Given that there may be similar sites in our data we examined a local context of length 13 surrounding each lysine in Seq40. Between positive and background 10% of the positive sites had high similarity i.e. $geq9$ residues in common with the negative sites (see Supplementary Figure A.3.1). From a pattern matching point of view, local context of the positive and negative lysines intersect considerably, making the discrimination problem difficult. In other words the classification is not easily separable and moreover non-linear algorithms should be used in order to separate data. Within each set of positive and background lysines, there could be pairs of common sites. This can result in algorithm learning more about these sites at the expense of others. Moreover, similar sites will overestimate performance because the testing fold may contain similar data in the learning fold of the cross validation. The positive set was quite diverse with only 1.1% sharing similarity with another site $geq9$ residues (see Supplementary Figure A.3.2 left). The negative set was also diverse with 4.5% having $geq9$ residues common but this reduced dramatically after 9 residues in common (see Supplementary Figure S2 right). In order to have good generalization, it is vital for any learning algorithm to assess the diversity of the local context as opposed to the full sequence diversity. We conclude that the Seq40 set was indeed diverse to a sufficient level and decided to leave it intact in order to maximize data size.

7.4.2 BRNN improves over standard motif based approaches

Standard motif based approaches must define a sequence window a priori (13 residues in this work). However, the BRNN architecture allows learning a dynamic window which potentially captures a greater local context and reduces the likelihood of over-fitting due to a decrease in parameters. It effectively takes the full sequence as input and tries to capture local context dynamically. These reasons, with perhaps others, allow for an increased performance of the BRNN over the SVM trained on the same data (see Table 3). The MCC almost doubled from 0.152 to 0.295 and AUC increased by 12% points when comparing SVM-M40 and BRNN-S40. It was also important to note that the ratio of positive to background lysines was approximately 1:16. However, the results changed very little when balancing the ratio (AUC for SVM-M40: 0.740, 0.724, 0.734, 0.729 and 0.740 for ratios 1:1, 1:4, 1:8, 1:12 and 1:16 respectively). We conclude that training with the BRNN and its full sequence representation to be best. Next we examine if conservation information in the form of multiple sequence alignments improved performance.

Method	MCC	Spec	Sens	AUC
SVM-M40	0.152	0.951	0.199	0.740
BRNN-S40	0.295	0.947	0.375	0.860
BRNN-S40+ali	0.133	0.947	0.187	0.707

Table 7.3: Cross validation results. SVM and BRNN and alignment cross validation performance on the Seq40 set. 5% FPR thresholds found on training folds since it is a parameter.

7.4.3 Conservation

Table 3 shows a simple conservation encoding degraded performance by over 11% points in AUC with a similar trend for MCC (BRNN-S40 vs. BRNN-S40+ali).

Ubiquitination was previously found to be unconserved across different species [124], suggesting that using many species (UniRef90) for our alignment sequence database might need to be revised. Perhaps more sensitive alignments based only on eukaryotic sequences should be considered in the future. It was the goal of this work to construct a fast proteomic scale predictor with good performance, and due to no improvements using simple PSI-BLAST alignments, the basic amino acid encoding was retained. At this stage training with the BRNN and a simple amino acid encoding was found to be the best technique. Next we examined the phenomenon of intrinsic disorder and ubiquitination for our H. sapiens dataset.

7.4.4 Intrinsic disorder and ubiquitination

There have been different views whether intrinsically disordered regions/proteins are involved in ubiquitination, which may be due to the datasets available for each analysis. For example, observations in S. cerevisiae based data suggested ubiquitination to be correlated with disorder [123, 125]. Recently, it was noted that S. cerevisiae disordered proteins are highly ubiquitinated after heat-shock treatment [136]. Other work based on larger scale analysis report a weak correlation between structure and ubiquitination [129, 137]. All these views need to be investigated in detail by collecting many more datasets but for simplicity we restricted our analysis to our available data. Disordered regions can be predicted from sequence with good accuracy [90]. Our accurate and fast method ESpritz [77] was used to predict if each amino acid in Seq40 was disordered. Table 4 proves that disorder as a feature produced a statistically significant gain of 1.6% in MCC and almost 1% point in AUC (RUBI-5%FPR vs BRNN-S40 in Table 3; students -test p-value<0.01). In fact, as a baseline predictor it was observed that disorder probability alone can be used to find ubiquitination with an AUC of 0.584, (see Supplementary Table A.3.1). The final predictor (hereafter termed RUBI) was trained using a BRNN with a simple amino acid encoding plus disorder probability. Table 4 also shows the behavior of RUBI at two strict false positive rates in the cross validation. Although disorder probability improves RUBI's performance we still have not answered the question whether it correlated with structure or intrinsic disorder. A simple statistical test was carried out on the independent set (PUB) as it contains multiple experimental sources. The Wilcoxon rank sum test was calculated for both disordered and structured lysines on the 1,264 PUB proteins. The mean disorder probabilities of ubiquitinated lysines were compared with those of all lysines (i.e. the control set). Figure 7.1 confirms a statistical significance (p-value << 0.001) for the structure-ubiquitination relationship. There were 405 ubquitination sites predicted in disordered regions and 3,799 sites in structured regions. When a lysine is predicted in a disordered region its probability of disorder when ubiquitinated is significantly lower (0.442 vs 0.329). When a lysine is predicted in a structured region the probability of disorder is slightly but significantly lower when ubiquitinated (0.031 vs 0.025). In conclusion, our measurements showed that ubiquitination sites had a preference for structured regions on this data source.

Method	MCC	Spec	Sens	AUC
RUBI-5%FPR	0.311	0.949	0.389	0.868
RUBI-1%FPR	0.211	0.990	0.127	0.868

Table 7.4: Cross validation performance using disorder feature. Experiment with disorder probability for each residue. 1% and 5% FPR thresholds found on training folds since it is a parameter.

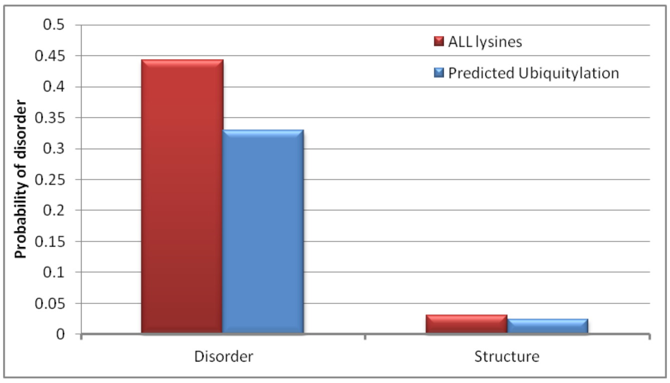


Figure 7.1: Disorder probability for all and ubiquitinated lysines. Mean disorder probability for residues classified in disorder or structure by ESpritz x-ray (5%FPR decision). Predictions performed on the PUB data with ESpritz x-ray disorder probabilities. In total there were 405 ubquitination sites predicted in disordered regions and 3,799 sites in structured regions. Both differences are statistically significant (p-value << 0.001, unpaired Wilcoxon rank sum test).

7.4.5 Ubiquitination content

The main concern in the literature is the determination of the actual lysine sites but little attention is paid to how many sites are contained per protein. In order to test the accuracy of RUBI beyond the mere recognition of individual sites and to check for systematic errors, we investigated the distribution of predicted vs. experimental ubiquitination sites in the dataset. The distribution of experimental ubiquitination sites on the Seq40 dataset indicates that proteins with up to 5 experimental ubiquitination sites account for 91% of the dataset, with those with a single site accounting for 49% alone and the rest decaying exponentially (see Supplementary Figure A.3.3). The results in the cross validation, shown in Figure 7.2, indicate that RUBI predicts the number of ubiquitination sites correctly especially for proteins with fewer sites, ranging from ca. 25% (for mono-ubiquitination) down to ca. 10% (for $geq5$ sites). Allowing up to three over- or underpredictions yields an accuracy of over 80%. These results suggest that RUBI performs well across a wide range of proteins and the number of predicted ubiquitination sites roughly corresponds to the experimental sites.

7.4.6 Independent testing

RUBI was retrained on all 3,705 sequences in Seq40 and tested on an independent set. In addition, UbiPred [122] and UbPred [123] were compared to our model. Table 5 shows the performance of UbPred, UbiPred and our final model on the full sequences in the PUB independent set. In order to demonstrate the different levels of confidence, thresholds at 5% and 1% false positive rate were used. The AUC for RUBI is 0.745, proving that the method behaves well on unseen data.

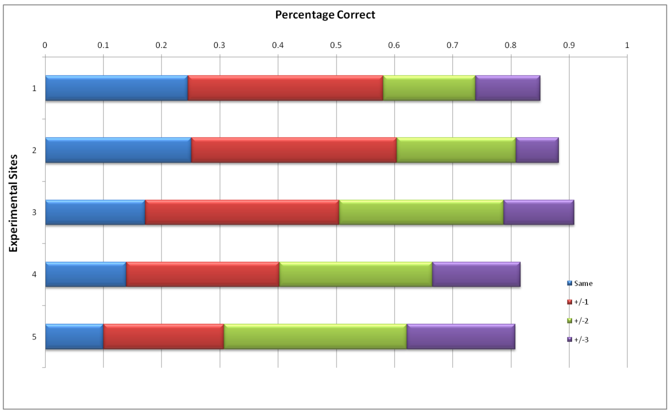


Figure 7.2: Ubiquitination content performance. Percentage of correctly predicted ubiquitination sites (x-axis) is shown with 1 through 5 experimentally ubiquitinated sites (y-axis). The colored bars correspond to the same number of predicted and experimental sites and a difference of up to three predictions more or less than the experimental sites.

The specificity and sensitivity remain high with only 10.9% false positives (specificity 0.891) producing approximately 31% true positive rate (sensitivity 0.306). At a higher confidence level (RUBI-1FPR%), as expected specificity was increased at the expense of sensitivity. On first inspection there was a good signal, albeit with a decrease compared to the cross validation results in Table 4, which may be explained by different high-throughput experimental conditions (see next section). Despite this, RUBI still generalized well on different high-throughput mass spectrometry techniques. Both UbPred and UbiPred performed poorly on our dataset. This was to be expected as performance was tested on human but both were trained on yeast data. While working on RUBI, two new predictors appeared: UbiProber [127] and hCKSAAP_UbSite [126]. UbiProber and hCKSAAP_UbSite reported good performance when detecting H. sapiens sites, thus allowing us to compare fairly. In Table 6 RUBI was directly compared with the independent sets of both together with UbPred and UbiPred. For H. sapiens based predictions, RUBI holds its performance, outperforming all predictors except UbiProber but still remaining comparable (UbiProber vs RUBI: 0.782 vs 0.758). Table 6 also demonstrates RUBI has a good signal on other species such as M. musculus (AUC 0.616) and on the more distant species S. cerevisiae with AUC 0.750. The high S. cerevisiae result is particularly interesting because our model was trained on H. sapiens only, showing the possibility for cross species prediction. Ideally, to avoid intersection of training and independent sets RUBI should be retrained on the exact training and independent split proposed by each method. To ensure this, RUBI was retrained on the H. sapiens sets from UbiProber and hCKSAAP_UbSite. Table 7 shows that RUBI outperforms all others with respect to AUC (UbiProber vs RUBI: 0.782 vs 0.818 and hCKSAAP_UbSite vs RUBI: 0.757 vs 0.820) for H. sapiens detection. Table 7 is encouraging for future versions of RUBI which can be easily retrained and updated with the latest experimental data.

Method	MCC	Spec	Sens	AUC
RUBI-5FPR%	0.131	0.891	0.306	0.745
RUBI-1FPR%	0.053	0.990	0.035	0.745
UbPred	0.024	0.830	0.211	0.599
UbiPred	0.062	0.653	0.537	0.592

Table 7.5: Performance on the PUB independent set. 5% FPR thresholds found on Seq40. UbPred (medium confidence, 5% FPR) motifs of length 29. UbiPred default 0.5 score decision, full sequences.

7.4.7 Experimental annotation is vital

It is commonly held that ubiquitination sites differ by species but not much is known about differences in the experimental technique used to detect them. The Phosphosite plus database [132] offers three styles of annotation allowing fluctuations in RUBI's performance to be measured if the annotation strategy is different. The three sets PUB, CST and LTP (see section 2.1) are considered separately as ROC curves in Figure 7.3. Clearly there was a substantial difference in the techniques used for experimental annotation, with AUCs of 0.745, 0.605 and 0.491 for PUB, CST and LTP respectively. We can postulate that the sequence surrounding the lysine were different in each category. It may even be argued that rather than species variation, differences were due to the experimental technique used on each species. However, more experiments are needed to verify this point. Our model closely resembles the public high-throughput mass spectrometry data in the literature (see blue curve in Figure 7.3, Table 5 for the PUB set performance and in Table 6 and 7 for further proof). This was to be expected as it was the most similar category to the training data. It is interesting that the LTP category was predicted poorly, suggesting our model can only mimic high-throughput experiments. In addition, CST annotations were predicted substantially worse than the published literature annotations (PUB). It is important to note that we are by no means stating any experimental technique is incorrect, but merely trying to find the technique our model captures best.

7.4.8 Proteome-scale predictions

In order to make RUBI viable for high-throughput usage on a proteomic scale, it was benchmarked for speed on a standard Linux server. The CPU time (single core) per protein was found to be 0.42 seconds on average or ca. 7 minutes per thousand proteins (see Supplementary Figure A.3.4), making it for a very efficient predictor which can be easily applied on many genomes. We therefore applied Rubi at 1% FPR to predict the distribution of high-confidence ubiquitination sites across a variety of genomes. The results for a representative set of model organisms is shown in Figure 7.4. The average fraction of proteins with at least one ubiquitination site was found to be at least 25% for all organisms, except Leishmania infantum. The fraction was very constant among higher eukaryotes, but raised and fluctuated among simpler organisms such as fungi and parasites. While the exact proportions will need further investigation, it is interesting to note that RUBI predicts many possible novel high-confidence ubiquitination targets in

UbiProber data				hCKSAAP_UbSite data	
Method	H.sapiens	M.musculus	S.cerevisiae	Method	H.sapiens
RUBI	0.758	0.616	0.750	RUBI	0.888
UbiProber H.sapiens*	0.782	0.838	0.899	hCKSAAP_UbSite*	0.757
UbiPred*	0.586	0.462	0.404	UbiPred*	0.560
UbPred*	0.596	0.644	0.736	UbPred*	0.497

Table 7.6: AUC Testing performance on previously published data in UbiProber and hCKSAAP_UbSite. RUBI model was trained on our original training set but tested on sets from UbiProber and hCKSAAP_ubisite. "*" means AUC taken directly from the corresponding publication. Vertically are the models and horizontally are the testing sets separated by species if available.

Method	UbiProber splits	Method	hCKSAAP_UbSite
RUBI retrained	0.818	RUBI Retrained	0.820
UbiProber H. Sapiens*	0.782	hCKSAAP_UbSite*	0.757
UbiPred*	0.586	UbiPred*	0.560
UbPred*	0.596	UbPred*	0.497

Table 7.7: AUC Testing performance on previously published H. sapiens data in UbiProber and hCKSAAP_UbSite. RUBI was retrained and tested on human sets from UbiProber and hCKSAAP_ubisite. "*" means AUC taken directly from corresponding publication. Vertically are the models and horizontally are the testing sets separated by species if available.

the H. sapiens genome.

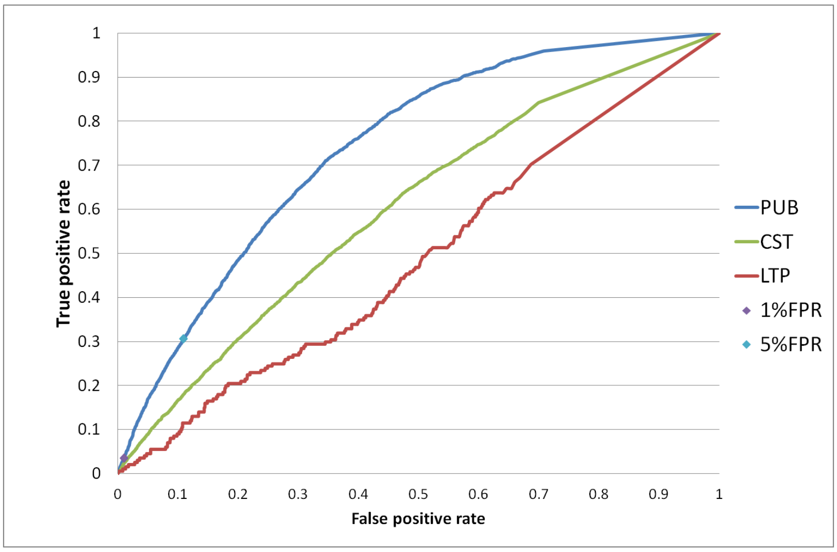


Figure 7.3: Receiver Operating Characteristic (ROC) on different annotation styles. ROC curves are shown for different subsets of the independent dataset. Low throughput (LTP) data is shown in red, while public high-throughput data (PUB) is shown in blue and unpublished Cell Signalling, Inc. data (CST) in green. Purple and cyan diamonds show the 1% FPR and 5% FPR on the PUB curve respectively.

7.5 Conclusion

In this paper we have presented RUBI, a novel method for ubiquitination site prediction from sequence. The method was trained on one of the largest currently available experimental H. sapiens data and was shown to be robust and accurate across a wide range of conditions. The final predictor was constructed in an iterative manner and some factors influencing its performance were illustrated. The factors which boosted RUBI's performance included: (i) the sequence representation imposed by the machine learning algorithm and (ii) intrinsic disorder. Other performance factors, such as local lysine sequence distribution and addition of conservation from sequence alignments, were also analyzed. The former must be calculated in order to ensure site diversity in learning while the latter degraded performance. Each factor could potentially aid in the development of other post-translational modification predictors. In addition to evaluating RUBI on individual sites we also attempted to measure if it could detect the amount of ubiquitination per protein. We believe this is the first such measurement. The best model found in this work can be retrained on different datasets in a matter of days. As higher quality data becomes available, the RUBI server will undergo systematic updates. Protein structure (non disordered regions) was found to correlate with ubiquitination for the datasets used in this work. Intrinsic disorder analysis separating ubiquitination into different data sources (e.g. S. cerevisiae vs H.sapiens or different experimental techniques) might produce different correlations. We plan to integrate RUBI in our database of disorder annotations for proteins MobiDB [92] thus allowing these correlations to be calculated easily. RUBI has a good

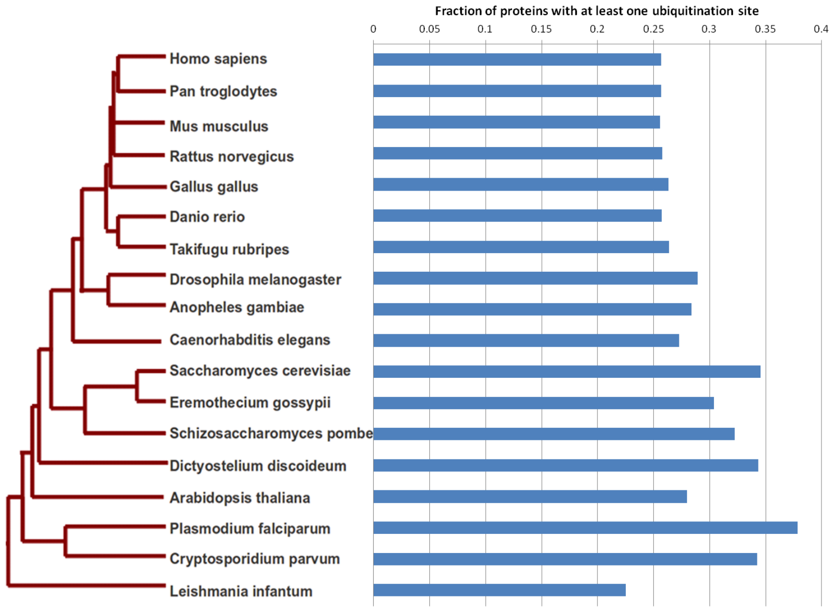


Figure 7.4: Fraction of predicted ubiquitinated proteins per model organism. The phylogenetic tree is shown together with the ratio of proteins with at least one ubiquitination site predicted by Rubi at 1% FPR for each genome

generalization ability and a signal across different eukaryotic organisms. It is also fast enough to enable the first genome-wide comparison of ubiquitination sites, which suggests the existence of thousands of possible ubiquitination sites which awaiting experimental validation.

Part III

Tandem Repeated Proteins

8. RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures

This chapter was first published in Walsh I, Sirocco FG, Minervini G, Di Domenico T, Ferrari C, et al. (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. Bioinformatics 28: 3257–3264.

8.1 Summary

Repeat proteins form a distinct class of structures where folding is greatly simplified. Several classes have been defined, with solenoid repeats of periodicity between ca. 5 and 40 being the most challenging to detect. Such proteins evolve quickly and their periodicity may be rapidly hidden at sequence level. From a structural point of view, finding solenoids may be complicated by the presence of insertions or multiple domains. To the best of our knowledge, no automated methods are available to characterize solenoid repeats from structure. Here we introduce RAPHAEL, a novel method for the detection of solenoids in protein structures. It reliably solves three problems of increasing difficulty: (1) recognition of solenoid domains, (2) determination of their periodicity and (3) assignment of insertions. RAPHAEL uses a geometric approach mimicking manual classification, producing several numeric parameters that are optimized for maximum performance. The resulting method is very accurate, with 89.5% of solenoid proteins and 97.2% of non-solenoid proteins correctly classified. RAPHAEL periodicities have a Spearman correlation coefficient of 0.877 against the manually established ones. A baseline algorithm for insertion detection in identified solenoids has a Q2 value of 79.8%, suggesting room for further improvement. RAPHAEL finds 1931 highly confident repeat structures not previously annotated as solenoids in the Protein Data Bank records.

8.2 Introduction

Protein repeats contain tandem arrays of smaller structural motifs where, unlike most globular domains, folding is reduced to simple coiling and long-range interactions are greatly reduced [138–140]. Repetitive proteins evolve quicker due to the intrinsically error-prone process connected with the formation of repeating sequences [141]. Fourteen percent of all known protein sequences are strictly periodic and it was hypothesized that repeating sequences occur more frequently in eukaryotic proteins [142]. Repeating sequences were estimated to occur in around one in three human proteins [139, 143].

Classification of repeating proteins is usually achieved in terms of repeat unit length [139, 144]. The length of the repeating unit can be as small as one or two residues for different types of crystallites of unlimited size. At the other extreme are repeating units of entire domains (beads on a string) with a typical repeating unit of over 50 residues. The middle ground comprises solenoid repeats with units of 5–40 residues. These are elongated structures containing α-helices and/or β-strands with a large distance between the N and C termini [140]. There has been increasing interest in solenoid proteins over the years, especially their relevance in health [145, 146] and for protein engineering applications [147, 148]. Solenoid proteins have also been shown to fold sequentially, one unit at a time, suggesting that the sequence contains all necessary information to determine the local fold [149]. Understanding solenoid function and evolution passes through their classification from sequence and structural information, which are two different problems. Solenoid sequences evolve quickly while maintaining their fold, thereby hampering detection [138]. Several sequence-based methods predicting tandem repeats from self-alignments have been developed over the years, including RADAR [150], TRUST [151] and HHrepID [152]. Our previous work REPETITA uses a fast Fourier transform to specifically detect solenoids [52]. In all cases there is still room for improvement, with the best methods still missing out many solenoids, especially with insertions. Generally speaking, solenoid repeats tend to be easy to spot through visual inspection in a molecular viewer. However, the manual search of hundreds or thousands of structures to determine if they are solenoid repeats or not is extremely time-consuming and inefficient. Moreover, the definition of repeat length, i.e. repeating blocks containing similar residue numbers, and detection of breaks in the periodicity require objective measures.

Available structural databases such as Protein Data Bank (PDB) [39] and CATH [153] store solenoid structures but do not provide feasible means for extracting them. Tools for discriminating protein repeat structures from globular proteins are rare in the literature. DAVROS [154, 155] is perhaps the first method developed for this purpose. Unfortunately, it is no longer maintained. ProSTRIP [156] is designed to find all similar structural repeats. It requires the selection of the repeat length and alignments from a set of alternatives, making it impractical for large-scale analysis. The Propeat database was designed by extracting recurring protein sub-structures, including internal repeats, but most of the structures contain only two repeating units [157]. A similar self-alignment approach is used by Swelfe to detect internal repeats in structures [158]. When developing

REPETITA, we had to manually derive a dataset of 105 solenoids [52]. Other sequence repeat prediction methods had similar problems in defining the dataset, e.g. in HHrepID the authors resort to structural self-alignment due to the lack of available tools for unbiased detection of solenoid repeats from structure [152].

This study aims to detect solenoid repeat structures using distance and periodic features extracted from the structural coordinates. The algorithm is efficient, has high discrimination power, can determine the repeat unit length and can find insertions that break the periodicity temporarily. The consequences of the algorithm are vast and here we tackle the large-scale extraction of repeats from CATH and the PDB. In addition to the novel data produced, a server is available (URL: http://protein.bio.unipd.it/raphael/) which can determine how periodic a structure is, the repeat length, periodicity and insertion plots.

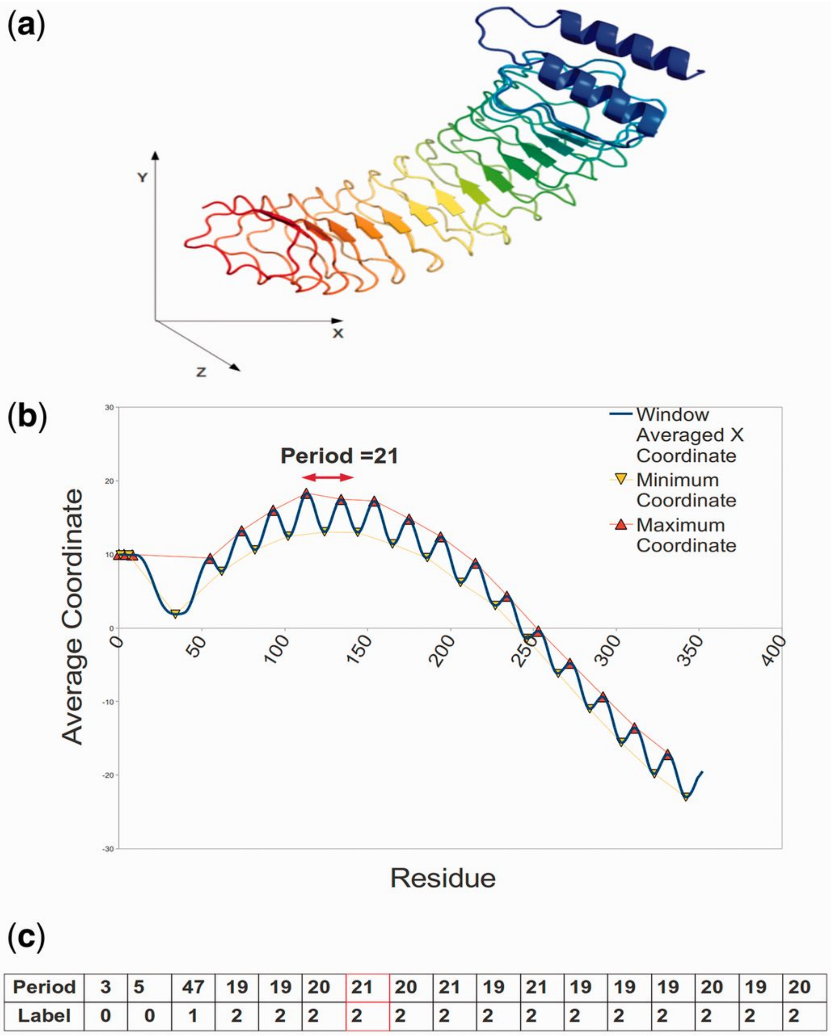
8.3 Methods

Periodicity and distance measures are both important factors when considering a particular protein visually. The aim of our algorithm is to mimic the intuitive definition used by a manual curator, extracting these two factors from the three-dimensional coordinates of the structure. A set of parameters and filters are then derived to capture the essence of periodic spatial patterns. It should be noted that while signal processing methods such as fast Fourier transform can be used for repeat proteins, our previous experience suggests that they do not excel on biological data with intermittent insertions [52].

8.3.1 Periodicity

For each C-alpha coordinate (i.e. x, y and z), a profile/wave is generated, filtering by averaging the profile twice over a window for each coordinate profile. The first pass window size is 6 and the second pass window size is 3. Figure 8.1b shows an example of a coordinate profile derived from C-alpha coordinates. In order to avoid bias due to the initial orientation of the structure, the protein is anchored at a reference point by random translation and rotation. Anchoring is performed 200 times in order to build stable periodicity values, thus producing 3 x 200 profiles (i.e. 200 for each coordinate profile). A period is defined as the distance between consecutive local maxima on the profile curve (consecutive minima are also considered); Figure 8.1b. In order to score the periodicity, two observations are made: (1) frequent adjacent periods, termed window score, indicate solenoid proteins and (2) frequent periods separated by rarely occurring periods, termed bridge score, indicate solenoid proteins.

Let $\Theta_i = max_i + 1 - max_i \forall i = 1, \dots, M-1$ be the period calculated between adjacent local maxima on the coordinate profile (similarly for minima) where M is the total number of local maxima. A labeled sequence is constructed from the sequence of periods $\Theta_i, \dots, \Theta_{M-1}$ where $\Theta_i \in \mathbb{N}$. A period Θ_i is labeled with $k \in \mathbb{N}$ where k is the position of the first occurrence, i.e. $\Theta_i \in [\Theta_k - T, \Theta_k + T]$, when scanning the periods from N- to C-terminus. T is the acceptable difference in



Period	3	5	47	19	19	20	21	20	21	19	21	19	19	19	20	19	20
Label	0	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Figure 8.1: Tagging the periods for the x-coordinate of Leucine-rich effector protein YopM-a from Yersinia pestis (PDB code 1JL5). (a) The structure is shown colored from N-terminus (blue) to C-terminus (red). (b) The period as calculated from two consecutive local maxima on the averaged x-coordinate profile. (c) The period sequence for the profile from (b) with the tagged label sequence below it. Notice how similar periods are assigned to the same tag. As this is clearly a solenoid protein, there are many identical tag labels adjacent to each other.

residues between two periods that allows assignment of the same label. Otherwise a new label is attached. This labeling procedure results in a sequence of labels $L_i, \ldots, L_{M-1}$ representing periodicities found in the structure, which is the only information supplied to the window and bridge scoring functions described below. We found that $T = 5$ produces optimal results, see Figure 8.1c for an example of period sequence and the corresponding label sequence.

8.3.2 Functions

Let $C(L_i)$ be the number of times L_i appears in the label sequence. The window score is defined as:

$$W(L_i, L_j) = \begin{cases} 2C(L_i) & \text{if } |i-j| = 1 \text{ and } L_i = L_j \\ 0 & \text{otherwise} \end{cases} \tag{8.1}$$

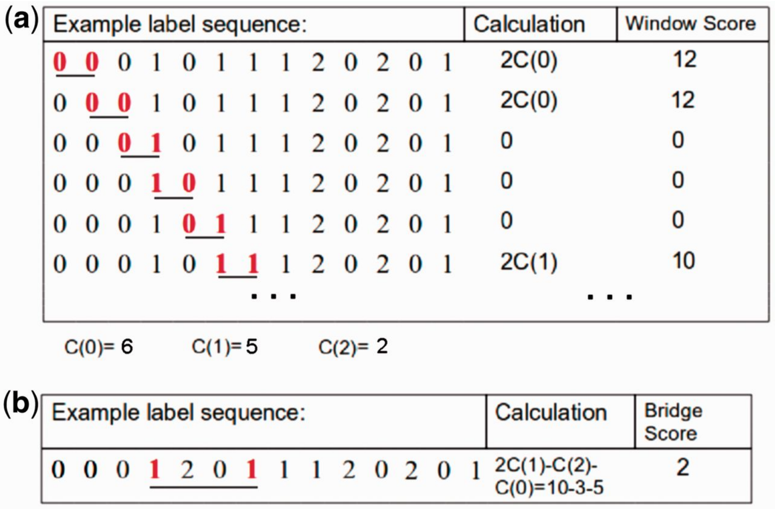
The window score is positive only for identical adjacent labels (i.e. $|i-j|=1$), see Figure 8.2a. Assuming we have two identical labels separated by an insertion of other labels, the bridge score penalizes the periods between them as follows:

$$B(L_i, L_j) = \begin{cases} 2C(L_i) - \sum_{j>k}^{j} C(L_j) & \text{if } L_i = L_j \\ 0 & \text{otherwise} \end{cases} \tag{8.2}$$

Figure 8.2b shows an example of the bridge score labeled sequence. The total periodic score for one coordinate and one random rotation and translation is:

$$Totalscore = \frac{pW^* + (1-p)B^*}{N} \tag{8.3}$$

where W^* and B^* are the final window and bridge scores (respectively) when processing the entire labeled sequence and N is the sequence length. Using a linear grid search on the training set, $P = 0.49$ was found to be the optimal balance parameter. The total score for the entire protein is the average of the three coordinate profiles and the 200 random rotations and translations.



(a)

Example label sequence:	Calculation	Window Score
0 0 0 1 0 1 1 1 2 0 2 0 1	2C(0)	12
0 **0 0** 1 0 1 1 1 2 0 2 0 1	2C(0)	12
0 0 **0 1** 0 1 1 1 2 0 2 0 1	0	0
0 0 0 **1 0** 1 1 1 2 0 2 0 1	0	0
0 0 0 1 **0 1** 1 1 2 0 2 0 1	0	0
0 0 0 1 0 **1 1** 1 2 0 2 0 1	2C(1)	10
. . .		. . .

C(0)= 6 C(1)=5 C(2)= 2

(b)

Example label sequence:	Calculation	Bridge Score
0 0 0 **1** 2 0 **1** 1 1 2 0 2 0 1	2C(1)-C(2)-C(0)=10-3-5	2

Figure 8.2: Example for the window and bridge scores. The positions being considered are shown bold faced in red and underlined. (a) The window score considers identical neighboring labels toward the total score. (b) The bridge score looks for identical labels separated by an insertion, here $i = 4$ and $j = 7$. See text for details

8.3.3 Parameters and optimization

The variance among all the periods found within a structure should intuitively be another important factor for discriminating solenoids. Let $P = \{\Theta^x_{1j}, \Theta^y_{1j}, \Theta^z_{1j}, \ldots, \Theta^x_{Rj}, \Theta^y_{Rj}, \Theta^z_{Rj}\}$ be the set of periods for residue j for all R rotations and translations along each coordinate frame x, y and z. On this set, let F_{kj} be the frequency of period k found in P for residue j. We define the period matrix PM to be a 2D matrix of dimension $60 * N$ with elements $F_{kj}, \forall k = 0, \ldots, 60$ and $j = 0, \ldots, N{-}1$, where N is the length of the protein and j is the index over residues. The cutoff was chosen to be the maximum allowed period since repeating units rarely exceed 60 residues for solenoid structures. Figure 8.3 shows the period matrix for a typical solenoid and non-solenoid protein. In order to measure the variation of periodicity within the entire protein, the standard deviation over all residues is calculated as:

$$SD = \sum_{j=0}^{N-1} \sum_{k=0}^{60} (F_j^{avg} - F_{kj}) \tag{8.4}$$

where F_j^{avg} is average frequency of column j in the period matrix. To complete the periodic information, we use the average period. Before calculating the average, set P is filtered by removing all outliers such that each period must be part of the interval $[P_{avg} - \sigma(P)/2, P_{avg} + \sigma(P)/2]$, where P_{avg} and $\sigma(P)$ are the average and standard deviation of all periods in P. This value is used to determine the solenoid periodicity length (termed P^* throughout the remaining sections).

Some observations about distance may be made through visual inspection of solenoid proteins: (1) solenoids, are usually elongated, (2) contacting residues in solenoids should have low sequence separation relative to globular proteins and (3) there should be regularity in sequence among the contacting residues (conversely, there should be large variance for non-solenoids). Two residues are in contact if the distance between the C-alpha coordinates of both residues is less than a pre-defined threshold. To measure the distance in 3D space between the N- and C-terminus, the following distance is used:

$$MD = min\left[d\left(i, j\right)\right] \quad \forall i \leq 40, j \geq N - 40 \tag{8.5}$$

where $d(i, j)$ is the distance between C-alpha atoms of residue i and j and N is the length of the protein. MD calculates the minimum distance between the first 40 residues and the last 40 residues. This value should give a good measure of protein elongation. Next, the number of contacting residues at a sequence separation ¿55 are calculated as follows:

$$NC = \frac{\sum_{i=0}^{N-1} \sum_{i-55>j>i+55}^{N-1} C_{ij}}{N} \tag{8.6}$$

where $C_{ij} = 1$ if the distance between i and j is ¡6 Å, a value chosen because it closely resembles the hydrogen bond distance. The sequence separation cutoff at 55 was chosen since solenoid unit length rarely exceeds this value for solenoids and contacts between repeating units can therefore be counted by NC. In contrast, long-range contacts are often present in globular protein structures [149, 159].

Finally, the regularity of contacting residues in the sequence is measured by the variance of the residue-wise contact order (RWCO) [160], which for residue i is defined as:

$$RWCO_i = \frac{1}{N} \sum_{i-3>j>i+3}^{N-1} |i - j| \, C_{ij} \tag{8.7}$$

where $C_{ij} = 1$ if the distance between i and j is ¡15 Å. This cutoff was chosen to relax the distance strength and thus allow a sufficient count at all sequence separations. $RWCOi$ is the sum of sequence separations between the ith residue

$i = 0, \ldots, N-1$ and all contacting residues. The variance of this property will give a measure of how regular the sequence separation is for contacting residues. Let $RWCO^{avg}$ be the average and $\sigma(RWCO)$ be the standard deviation of $RWCO$. The final value used for discrimination of solenoids is the standard deviation of the set defined by:

$$\{RWCO : RWCO_i \in [RWCO^{avg} - 0.6\sigma\left(RWCO\right), RWCO^{avg} + 0.6\sigma\left(RWCO\right)]\} \tag{8.8}$$

This gives a measure of the variance of the sequence separation between the contacts while ignoring extreme outliers.

The previously described periodic and distance features were combined using a support vector machine (SVM). The SVM C parameter was set to 0.02 and a simple linear kernel was used. The SVM produces a real number score with positive values indicating predicted solenoids and negative values indicating non-solenoids. The more positive the SVM score, the more solenoid the protein should be.

8.3.4 Finding insertions

A simple baseline method is used to discriminate non-periodic residues or insertions in a structure from the core solenoid repeat. The main source of data is the variation of distances between residue j and $j \pm P^*$ where P^* is the calculated period. For each residue j, we define the minimum periodic distance toward the N and C termini:

$$\begin{array}{l} PD_j^N = min\left[d\left(j, j - P^* \pm \Delta\right)\right] \\ PD_j^C = min\left[d\left(j, j + P^* \pm \Delta\right)\right] \end{array} \quad \forall\Delta = 1, \ldots, w \tag{8.9}$$

$d(.,.)$ is the Euclidean distance between C-alpha atoms on residue pairs. PD_N^j and PD_C^j are used as a double-pointed probe on the structure at residue j. First, it is important to determine the representative distance of a given period since proteins with the similar period do not necessarily repeat at the same distance. Given a protein of length L, the raw set of periodic distances $D = \{PD_1^C, PD_1^C, \ldots, PD_1^C, PD_1^C, \}$is reduced to the subset $Df \subset D$ using the following conditions:

$$\begin{array}{c} PD_j^{N/C} < T \\ PD_j^{N/C} \in \left[D^{avg} - \sigma\left(D\right)/2, D^{avg} + \sigma\left(D\right)/2\right] \end{array} \tag{8.10}$$

where D_{avg} and $\sigma(D)$ are the average and standard deviation of D, respectively. These conditions ensure the removal of extreme outliers and large non-meaningful distances (i.e. non chemical bonds). Let mDf denote the median of the set Df. It is in fact the variation from mDf, which will measure the potential for non-periodicity. This variance profile is defined as follows:

$$VP = \sum_{j=1}^{L} \begin{cases} 1 & \text{if } PD_j^{N/C} < mDf \pm \lambda \\ 0 & \text{otherwise} \end{cases} \tag{8.11}$$

when calculating distances boundary conditions, $|j + P^* \pm \Delta| \leq L$ and $|j + P^* \pm \Delta| \geq 1$, were implemented. The parameters w, T and λ were determined using a grid search on the training folds of the leave one out procedure. Values for the parameters were found to range $w \in [9, 10]$, $\lambda \in [1.5, 2.0]$ and $T \in [12, 15]$ depending on the training fold. Intuitively, the idea is to capture the maximum deviation of each residue from the median periodicity. This is a simple algorithm, which may be further improved with more parameters and machine learning but should nevertheless provide a valid baseline for detecting insertions and repeat boundaries. Throughout this article, we will refer to insertions as non-repeated residues surrounded by solenoid repeats. Only the final experiment is shown in this article, with results for the two partial optimizations shown in the Supplementary Material A.4. All thresholds were found by maximizing Q2 on the training sets.

8.3.5 Datasets

The training and test sets are based on publicly available data from the REPETITA article [52]. Briefly put, an initial set of 32 solenoid repeat proteins was taken from a previous review [140] and expanded using TESE [161] to find more protein domains in CATH [153] belonging to the same solenoid folds as the initial set. Choosing representatives with at most 35% pairwise sequence identity (i.e. CATH 'S' level) yielded a set of 105 solenoid domains. The set of non-solenoid protein domains was generated with TESE by randomly choosing X-ray structures with different topologies and no detectable sequence similarity (i.e. CATH 'T' level) for a total of 247 domains. The sets of solenoid and non-solenoid protein domains were randomly split into training and test sets, with the constraint that solenoid structures of low similarity fall in the same partition. It is worth mentioning that closed repeating structures such as beta-barrels or propellers are not included in the set and our algorithm does not consider these toroidal structures, but may still find their periodic signal.

In addition to the training and test sets, RAPHAEL was also benchmarked on CATH and PDB. The 'S' and 'O' level classifications, with a maximum sequence identity of 35% and 60% were downloaded from the CATH website for the current version (v3.4). The PDB was downloaded as of July 1, 2011. DNA, RNA and protein chains with length ¡30 amino acids were removed. Each structure was separated into chains and reduced to 40% sequence identity using CD-HIT [131] with options -c 0.4 –n 2, creating a diverse set of 16 226 unique chains.

8.3.6 Performance measures

Throughout this article, TP, FP, TN and FN are used for true positives, false positives, true negatives and false negatives, respectively. Sensitivity and precision values are calculated for both periodic (P, positive class) and non-periodic residues/structures (N, negative class). The following measures are used: sensitivity $(P) = TP/(TP + FN)$, precision $(P) = TP/(TP + FP)$, sensitivity $(N) = TN/(TN + FP)$ and precision $(N) = TN/(TN + FN)$. Accuracy is used as synonymous to sensitivity and Q2 is the fraction of correctly predicted residues, i.e. $(TP + TN)/(TP + FP + TN + FN)$. The receiver operator characteristic (ROC) curve describing the overall performance at variable thresholds is plotted as TP rate versus FP rate.

To compare RAPHAEL to existing methods, we chose the structure-based method Swelfe [158] and three sequence-based methods: REPETITA [52], TRUST [151] and RADAR [150]. Since Swelfe returns several alternative predictions, the best was considered in order to overestimate rather than underestimate its performance. The results for the sequence-based methods are taken from our previous publication [52]. The comparison should be considered a baseline only, given that all of these tools (except REPETITA) are not explicitly designed for solenoid detection.

8.4 Results

8.4.1 Solenoid identification

In order to identify possible solenoids, RAPHAEL transforms the coordinates of the protein structure into a period matrix. An example for the transformation of a solenoid and a clearly non-repetitive structure can be seen in Figure 8.3. The solenoid structure produces a compressed signal of higher intensity, which can be used for detection. Several parameters were derived to take advantage of this information (see Section 2). The performance at discriminating solenoids with the combined SVM score on the training set is shown in Table 8.1, while the individual parameters are reported in Supplementary Table A.4.1. Although the window function is the most discriminating feature, the SVM combination improves performance by ca. 4% for solenoids and ca. 7% for non-solenoids, suggesting that different information is captured. Due to the limited number of training data and to be more statistically robust, we also tested the performance of a leave one out cross-validation. Here, training is performed with N-1 protein chains and testing with the remaining chain, while counting the results for all the testing examples (n = 351). This produces results somewhere between the training and test sets, with only 7 false solenoids and 11 false non-solenoids for the entire dataset. Table 8.1 also shows how a stricter SVM threshold of 1.0 produces just 1 false solenoid, at the expense of losing 14 solenoids, thereby increasing positive precision to 98.9% compared with 93.1% for an SVM score of 0. In other words, an SVM threshold of 1.0 corresponds to very confident solenoid assignments.

	TP	FP	TN	FN	Solenoids	Non-solenoids
Training	49	2	117	1	98.0	98.3
Testing	48	6	122	7	87.3	95.3
Leave one out ¿ 0	94	7	240	11	89.5	97.2
Leave one out ¿ 1	91	1	246	14	86.7	99.6

Table 8.1: Accuracy on the training set and test set combining all six features through an SVM. Results are shown for the method optimized on the training set (first two rows) and on the leave one out split (last two row), respectively. The latter are further reported at an SVM threshold of 0 and 1.

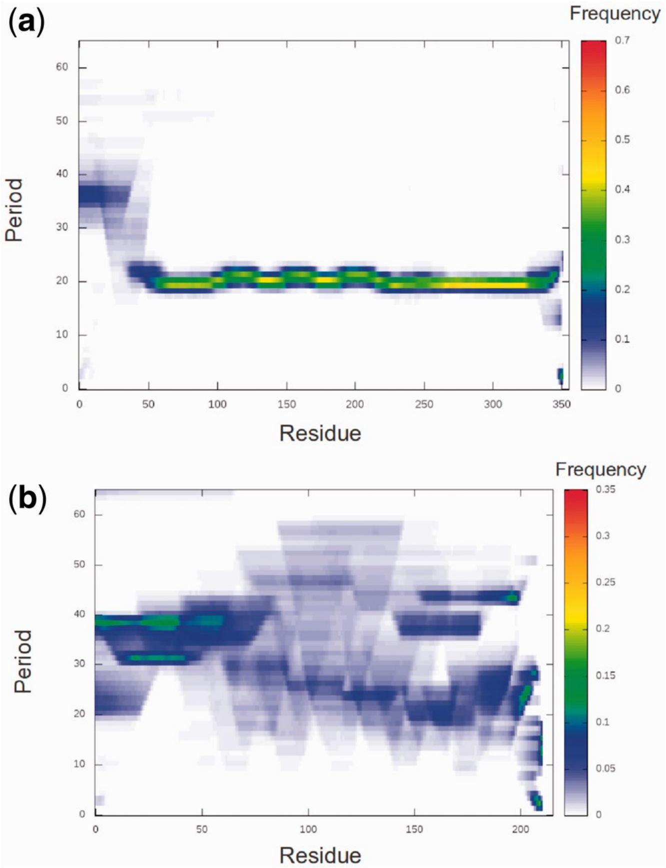


Figure 8.3: The period matrix for (a) solenoid protein YopM-a leucine-rich effector protein from Yersinia pestis (PDB code 1JL5, as in Fig. 8.1) and (b) sulfhydryl protease from the latex of the papaya fruit (PDB code 9PAP). Notice the variation of period frequency for 9PAP while 1JL5 periodicity appears regular.

A full ROC curve for the leave one out cross-validation is shown in Figure 8.4, also comparing with Swelfe and three sequence-based methods. Swelfe is not specifically designed for solenoids, but rather tries to detect internal repeats in proteins. It should also be emphasized that solenoid detection from sequence is more difficult and hence such methods can be expected to perform less well. The difference in ROC curve is nevertheless remarkable, with RAPHAEL detecting three times more solenoids than the other methods at low FP rates and the most difficult solenoid at an FP rate of ca. 20%. Table 8.2 shows the distribution of correct and incorrect classifications for leave one out training split in terms of CATH class. Interestingly, it is the alpha-beta class which produces the most errors on solenoids (i.e. 7), suggesting that it may be somewhat more difficult to find solenoids when they have an alpha-beta mix. The datasets do not take into account Class 4 (few secondary structures) as either negative or positive examples.

8.4.2 Periodicity estimation

Once the presence of a solenoid has been established, it is important to define its periodicity, i.e. the length of the repeating unit. Supplementary Figure A.4.1 shows a comparison of the periods determined from the period matrix (see Section 2) to a manual derivation from our previous work [52]. The relationship is

Class	TP	FP	TN	FN	Solenoids	Non-solenoids
Mainly α	40	0	59	2	100.0	96.7
Mainly β	31	0	16	9	100.0	64.0
Mixed α–β	23	7	165	0	76.7	100.0

Table 8.2: Precision as a function of CATH class. Results calculated on the leave one out split. Precision results on solenoids and non-solenoids.

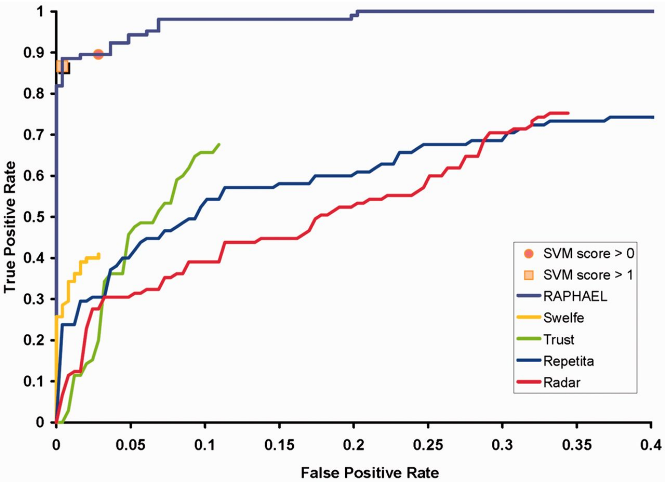


Figure 8.4: ROC curve on the combined training and test set. RAPHAEL trained using the leave one out split is compared with four other methods. The curve ends when a method does not produce further output, i.e. believes to have found all solenoids. Two SVM score thresholds are shown at 0 (orange circle) and 1 (yellow square), respectively.

clearly linear, with an overall Spearman correlation coefficient of 0.877 indicating a strong relationship between RAPHAEL and the manually extracted repeat lengths. Upon inspection, the small number of outliers exhibit period matrices which are highly variable and contain insertions and/or deletions. As expected, it is difficult to determine the repeat length when insertions or deletions are present in the structure. Looking in more detail at the difficulty level of the solenoids, the hard (i.e. solenoids containing many insertions) cases have a Spearman correlation coefficient of 0.753 compared with 0.934 for the easy ones (i.e. solenoids with few or no insertions). Figure 8.5 shows a comparison of RAPHAEL to Swelfe and three sequence-based methods in terms of detecting the correct periodicity. Since the exact period in solenoids with insertions can be somewhat arbitrary, we allow two distinct levels of correctness. In analogy to our previous work [52], we consider one residue around the manually curated periodicity correct for all predictions. For sequence-based methods, we also consider half or double the structural repeat as correct within tolerance. As structure-based methods (RAPHAEL and Swelfe) may be sensitive to insertions, we allow five residues around the exact period as correct within tolerance. The effect of the window size on RAPHAEL predictions is shown in Supplementary Figure A.4.2. As can be seen in Figure 8.5, RAPHAEL and the more accurate sequence-based methods have similar performances in recognizing correct periods. This is somewhat unexpected, but likely due to correct classification of solenoids without insertions where a clear sequence signal corresponds to the structural unit.

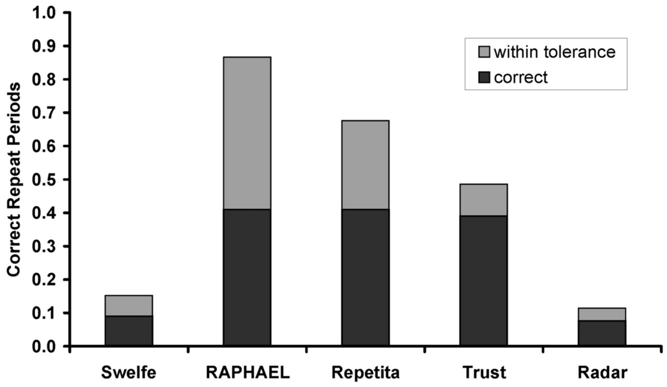


Figure 8.5: Detection of repeat periodicity for RAPHAEL and four other methods. See main text for details on the thresholds used to define the two levels of correctness.

8.4.3 Insertions

Given the performance in detecting solenoid proteins, the next question becomes whether the method is able to detect insertions for these proteins. To test this, every residue in each solenoid structure was annotated as either repeated or not. The Q2, sensitivity and precision measures for the dataset are shown in Table 8.3. It should be emphasized that we are proposing a simple baseline algorithm with a few caveats. First of all, several solenoid structures are rather degenerate, prompting a somewhat arbitrary distinction between approximately repeated and inserted residues. Second, RAPHAEL tends to find clear insertions but finds it difficult to determine less obvious cases, as clear insertions disrupt the regular spatial pattern at the basis of our algorithm. Hence, smaller insertions can be underpredicted, whereas longer insertions are found but often reported as more disruptive than necessary. An example can be seen in Figure 8.6. To the best of our knowledge, this is the first time that an automatic classification of structural repeat insertions is attempted in the literature. It certainly also expands our view on the previously released REPETITA dataset [52].

Measure	All	Easy	Hard
Q2	79.8	83.4	74.1
Sensitivity (P)	95.5	95.7	95.1
Precision (P)	79.5	84.9	69.6
Sensitivity (N)	44.2	40.3	47.4
Precision (N)	81.2	72.7	88.5

Table 8.3: Performance of simple insertion finding algorithm on leave one out cross-validation. The Q2, sensitivity and precision measures are shown after leave one out optimization for maximum Q2.

8.4.4 Large-scale extraction of periodicity data

In order to test RAPHAEL, we decided to process large sets such as the PDB and CATH to generate datasets for future use. For this large-scale search, we trained the SVM on the combined datasets (105 solenoids and 247 non-solenoid domains).

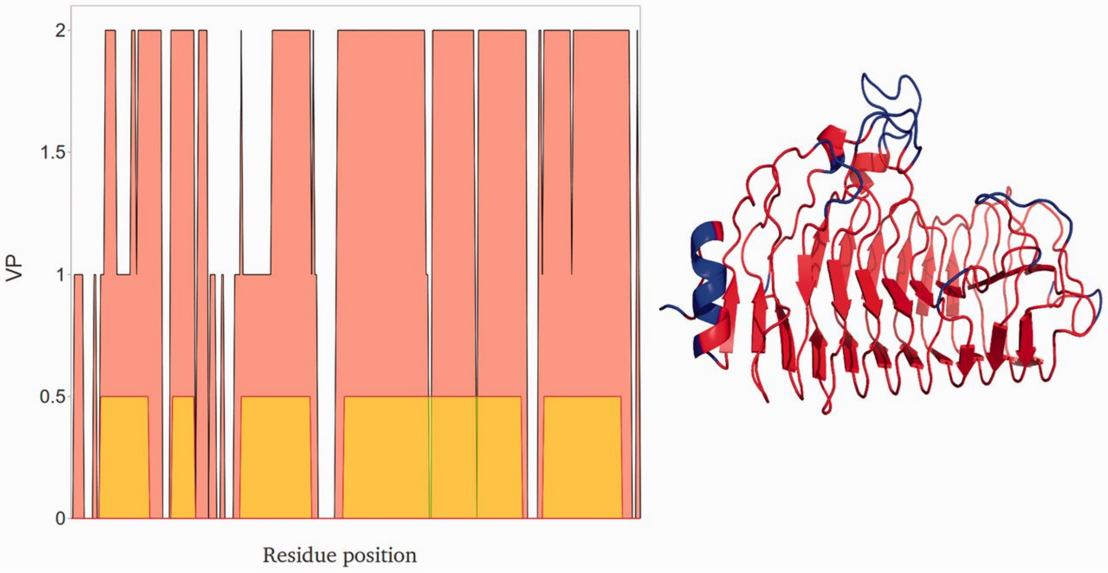


Figure 8.6: Example of insertions found for a β-solenoid. The variance plot (left) shows the score VP used to determine the location of insertions for endopolygalacturonase (PDB code 1HG8 chain A). The yellow area indicates the true positions of the periodic residues (periodicity at 0.5, insertion 0).The same structure (right) is colored in red for residues assumed to be repeated and in blue for insertions. Notice how the algorithm identifies the core solenoid domain, while mispredicting some C-terminal residues.

RAPHAEL was used to detect solenoids on the entire CATH database at the S(35) and O(60) levels corresponding, respectively, to 35% and 60% maximum sequence identity. Supplementary Figure A.4.3 shows the SVM score for all domains at S(35). Choosing this identity cutoff guarantees that solenoid domains are diverse at least at the sequence level but it can also be assumed to be true at the structural level. In total the algorithm considered 748 domains to be solenoids at this sequence identity cutoff (Table 8.4 and Supplementary Figure A.4.4). Obviously, the higher the score, the more expressed the periodicity should become. Upon visual inspection, the better solenoids are represented by an SVM score ¿1 (221 domains; see inset in Supplementary Figure A.4.4). In order to find more solenoid domains which may be useful, we also processed CATH with no sequence pair sharing 60% sequence identity. Using this less stringent cutoff, the algorithm detected 1156 CATH domains, with the distribution of SVM scores shown in Supplementary Figure A.4.5. A list of CATH domains ranked by the solenoid score produced by the SVM can be found on the RAPHAEL website.

	S(35)		O(60)	
Class	%	n	%	n
Mainly α	7.3	141	6.6	200
Mainly β	15.1	301	15.2	492
Mixed $\alpha - \beta$	6.4	302	5.6	456
Few sec. struct.	5.2	4	7.3	8

Table 8.4: Solenoid frequency in CATH. The frequency (%) and absolute number (n) of solenoids found in each CATH class are shown for the S and O levels at 35% and 60% maximum sequence identity, respectively.

Of course the extracted CATH domains will intersect with the set used for

	Structures	SVM > 0	SVM > 1
CATH S(35)	11,330	748	221
CATH O(60)	15,778	1156	308
PDB 40	16,226	1131	551
PDB full	74,020	5419	2,478

Table 8.5: Number of solenoids found in CATH and the PDB. The number of structures found with an SVM score >0 and 1 is shown for the CATH S and O levels (i.e. 35% and 60% maximum sequence identity) as well as for the PDB dataset made non-redundant at 40% sequence identity and the full PDB.

algorithm construction. Using a 50% sequence identity cutoff, we identify 696 proteins on S(35) and 1089 on O(60) which are not homologous to our training data. At the more stringent SVM score ¿1, the number of newly mined domains is 172 for S(35) and 245 for O(60). This has to be compared with the currently available list of 105 solenoid repeats [52].

In addition to CATH domains, we also processed PDB chains with RAPHAEL, finding 1131 chains to be considered solenoids at 40% maximum sequence identity. A more confident set of 551 solenoid chains with SVM score ¿1 was also generated. These numbers increase to 5419 and 2478 for the full PDB (Table 8.5). It is interesting to note how the PDB analysis contains a comparatively higher number of confidently predicted solenoid structures than CATH. This might suggest the existence of solenoid structures outside the already known CATH superfamilies, although further analysis will have to be carried out to verify this hypothesis.

To validate the results and verify the extent to which RAPHAEL detects previously unknown solenoid proteins, we have calculated the overlap of our predictions with PDB entries of proteins having the 'repeat' (or 'repeats') keyword in their respective header records. The results are drawn as a Venn diagram in Figure 8.7. It should be noted that PDB entries with the 'repeat' keyword contain proteins that are not true solenoid repeats, e.g. the repeated spectrin or fibronectin domains. Nevertheless, RAPHAEL overlaps well with the PDB annotation but provides an even greater amount of novel automatic annotations. These can be useful for the automatic annotation of proteins by structural genomics consortia or the PDB itself.

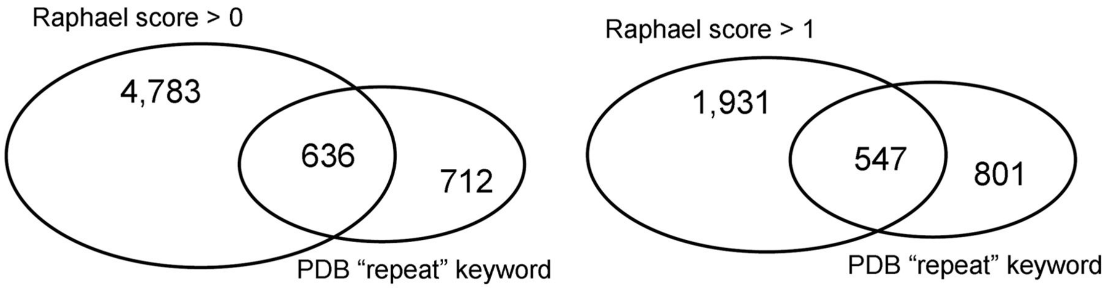


Figure 8.7: Venn diagram of RAPHAEL predictions and PDB repeat annotations. The predictions are shown at the SVM score cutoffs of 0 (left) and 1 (right) on the entire PDB. The PDB headers were scanned for the 'repeat' keyword.

8.5 Conclusions

In this article, we have presented a novel method, RAPHAEL, for the accurate determination of solenoid repeats from PDB structures. The method quantifies repeat structures by mimicking visual interpretation by experts through various parameters. Combination in a SVM provides exceptionally accurate predictions, as tested on a previously published dataset. To the best of our knowledge, we show for the first time that our method is also able to broadly recognize insertions and repeat boundaries. Scanning the entire CATH and PDB databases provides hundreds or thousands of additional solenoid repeats, with automatic annotation for repeat regions. RAPHAEL was implemented in a new web server-based application for automatic repeat protein recognition. Due to the importance of repeat proteins in both design and human diseases, we plan to use this method for systematic large-scale analysis of protein structures, in order to improve our understanding of these peculiar proteins and their impact on organism evolution.

9. RepeatsDB: a database of tandem repeat protein structures

This chapter was first published in Di Domenico T, Potenza E, Walsh I, Gonzalo Parra R, Giollo M, et al. (2013) RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Research 42: D352–D357.

9.1 Summary

RepeatsDB (http://repeatsdb.bio.unipd.it/) is a database of annotated tandem repeat protein structures. Tandem repeats pose a difficult problem for the analysis of protein structures, as the underlying sequence can be highly degenerate. Several repeat types haven been studied over the years, but their annotation was done in a case-by-case basis, thus making large-scale analysis difficult. We developed RepeatsDB to fill this gap. Using state-of-the-art repeat detection methods and manual curation, we systematically annotated the Protein Data Bank, predicting 10 745 repeat structures. In all, 2797 structures were classified according to a recently proposed classification schema, which was expanded to accommodate new findings. In addition, detailed annotations were performed in a subset of 321 proteins. These annotations feature information on start and end positions for the repeat regions and units. RepeatsDB is an ongoing effort to systematically classify and annotate structural protein repeats in a consistent way. It provides users with the possibility to access and download high-quality datasets either interactively or programmatically through web services.

9.2 Introduction

A large portion of proteins contain repetitive motifs, which are generated by internal duplications and frequently correspond to structural and functional units of proteins. Many repetitions in protein sequences can be identified by using different approaches [143, 152, 162, 163]. A more difficult problem for identification is however posed by repeats in protein structure, which can be highly degenerate [141, 164]. In fact, it is possible for a protein to maintain a repetitive structure even in the presence of massive amounts of point mutations [138]. Several repeat families have been studied so far due to their relevance in different biological processes such as health [165], neurodevelopment [145] and protein engineering [147, 148, 166], to name just a few.

Repeats have been previously divided into five broad classes, primarily as a function of repeat length [139, 142]. At the lower end of the repeat length spectrum, i.e. less than five residues, very short repeats can either form insoluble aggregates (crystallites, class I) or long and winding helices of fibrous structures like collagen and α-helical coiled-coils (class II). At the other end of the spectrum, repeats containing ><50 residues appear to fold mostly as domains forming beads-on-a-string structures (class V). In between, for unit lengths of 5–40 residues, the known repeats can form either open elongated solenoids (class III) or closed toroids (class IV). Due to their fundamental functional importance, classes III and IV contain the most studied types of tandem repeat proteins. Solenoid folds appear to follow the distribution of repeat lengths rather closely, from all-beta (e.g. anti-freeze proteins) [167] to mixed alpha/beta (e.g. leucine-rich repeats) [168, 169] to all-alpha structures (e.g. Armadillo and HEAT repeats) [170–172]. They are characterized by some of the largest known autonomously folding domains, with 500 or more residues forming a single structure [140]. Rapid addition or deletion of repeat units even between close homologs is of particular note for solenoid structures [173]. Toroids on the other hand are restricted in overall size by their closed circular nature. Known toroid structures include the highly versatile TIM barrel and large outer membrane beta-barrels [174]. Perhaps a more interesting fold is the beta propeller (e.g. WD repeats), which can accommodate variable numbers of repeat units while maintaining a closed circular structure [175, 176].

An open question regarding repeat proteins is the existence of other common structures that may have gone undetected. After all, the most common way to detect repeat families so far was to manually annotate the sequence family first and only afterwards visually recognize their structural repetitiveness. Such an approach is obviously difficult when dealing with the entire Protein Data Bank (PDB) [177], especially considering the many uncharacterized protein structures deposited by the main structural genomics consortia [178]. The systematic description of repeat structures becomes a question of using automated methods to detect them in protein structures. This field is relatively new, with only few available methods. One of the first attempts was made by the Thornton group [155], but is unfortunately no longer available. Some methods [52, 150–152, 158, 179] were developed to detect internal symmetries in proteins, but these may be difficult to adapt to the systematic classification of repeats. Recently, our group has developed RAPHAEL [180] in an attempt to fill the gap for repeat detection from structure. Widely used structural classifications such as CATH [181] and SCOP [182] also do not explicitly annotate repeats in protein structures, although it may be possible to leverage individual annotations to find similar repeats. Some databases exist for the detection of repeats from sequence [183, 184], but usually these are limited to short tandem repeats and do not take into account divergent repeats, such as solenoids or toroids. The main domain sequence databases such as Pfam [95] and SMART [185] do not excel at the annotation of these repeat types either, as coverage is rather low and many repeat units go undetected. For Pfam most of the largest clusters of human sequence regions not covered were recently found to be repeats [186]. To the best of our knowledge, no database or classification is currently available for repeat structures. This is the motivation for our present work, and we introduce RepeatsDB as a way to fill this gap. The database was developed to provide a central resource for the systematic annotation and classification of repeats. Given the fact that the structure-based search and classification of repeat proteins is more complete than on the basis of sequences or key words, our database will allow more accurate assignment of proteins with repeats to the corresponding families. For example, it will be used to suggest a better subdivision of alpha-solenoid proteins where at present the boundaries between the structures with Armadillo, HEAT, TPR and other repeat types are frequently blurred.

9.3 Database description

9.3.1 Data curation

The initial dataset for RepeatsDB was extracted from the PDB [187]. Repeat candidates were identified from the reduced PDB dataset with RAPHAEL [180], which uses a geometric approach imitating the work of a human curator (score cutoff ≥ 1). The resulting dataset consisted of >10,000 repeat candidates, stored in the database as 'predicted' entries, which underwent a classification and curation process.

The dataset of predicted repeats was manually curated using a two-level annotation system. The first manual annotation level ('manually classified') classifies an entry into structural repeat class and subclass. This classification is based on previous work [139], where five classes of repeat structures are proposed, which are then further divided into subclasses. Class assignment is based mainly on repeat unit length and subclass assignment on secondary and tertiary structure features. The second manual annotation level ('detailed') consists in providing information about the start and end positions of the repeat units, repeat regions and/or insertions. We define a repeat unit as the smallest structural building block that is repeated to form a repeat region. A repeat region is a group of at least three repeat units. Inclusion of proteins with two repeat units would significantly complicate classification because many typical globular domains have this type of architecture. Insertions are non-repeated segments of structure that occur either inside a repeat unit or between two of them. These are particularly interesting because they break the repeat symmetry, and represent a challenge both for automatic detection and for the analysis of repeat structures [180].

Several curators annotated each protein undergoing manual classification by consensus. For first-level annotations, at least 75% of the curators had to agree in order for a protein to be included, otherwise it would be excluded and placed on a reserve list for future annotation. The rationale for this choice is that ambiguous cases are generally difficult to classify but may occasionally represent a novel repeat class. For second-level annotations, the threshold for consensus was at least 65% agreement (typically two of three curators). In case of discrepancy, an expert would arbitrate the final annotation based on the alternative proposals. Proteins with detailed annotations were also used to search for similar sequences in proteins from the PDB. Any PDB chain with at least 40% sequence identity and a coverage of at least 80% of the classified protein, belonging to the initial list of predicted entries, is added to the 'classified by similarity' annotation level. The similarity thresholds were selected to exclude possible false-positives (data not shown).

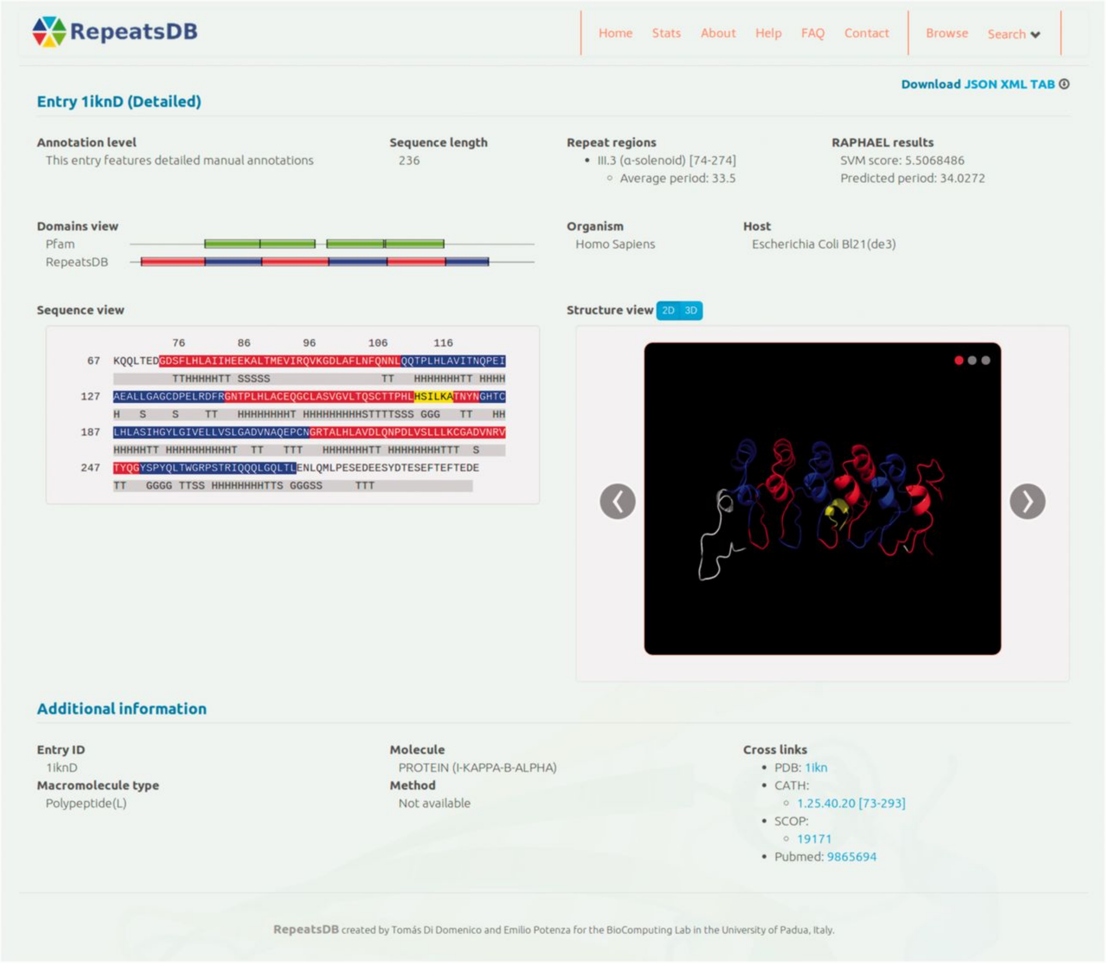
9.3.2 Implementation

RepeatsDB was designed with a multi-tier architecture, using separate modules for data management, data processing and presentation functions. To simplify development and maintenance, all tiers handle the common JSON (JavaScript Object Notation) format, thereby eliminating the need for data conversion. The MongoDB database engine is used for data storage and Node.js as middleware between data and presentation. RepeatsDB exposes its resources through RESTful web services, by using the Restify library for Node.js. The Angular.js framework and Bootstrap library were selected to provide the overall look-and-feel. Angular.js to Bootstrap integration is available through the angular-ui project. A customized version of the BioJS [188] sequence component is used as sequence visualizer. Additional information is added to entries by querying the PDB web services at the structure and chain level. At the structure level, annotations like organism and experimental method used when resolving the structure are provided. At the chain level, secondary structure and links to other databases, among others. RepeatsDB offers users both graphical web interface access and RESTful web services from URL: http://repeatsdb.bio.unipd.it/.

9.4 Using RepeatsDB

The user interface presents an intuitive tree-based browsing mechanism, where the root of the tree is the full database, second-level nodes repeat classes and third-level nodes subclasses. When clicking on a node, the user is presented with the list of RepeatsDB entries corresponding to the selected category. Each row of the list shows basic information about the entry, like its entry ID, title and organism. All annotated chains corresponding to an entry are displayed in a single page. The user interface presents a structure and sequence visualization widget (Figure 9.1). The user may choose to visualize the structure in four static images, or by using the 3D visualizer. If the entry features detailed annotations, the repeat regions, units and/or insertions are displayed using a combination of colours. The sequence visualization widget displays the sequence and secondary structure corresponding to the structure. It displays the same colour coding as the structure visualization widget, associating repeat annotations in the structure and sequence views. Additional information at the structure and chain levels is also provided.

The RepeatsDB search toolbar, available on top of every page, allows to search for entries either by database IDs or UniProt text query. The database ID search allows comma-separated PDB or UniProt IDs. The UniProt text search query uses the full UniProt search engine, see online documentation. RESTful web services are directly accessible through HTTP URLs. All data available on RepeatsDB



Entry 1iknD (Detailed)

Annotation level
This entry features detailed manual annotations

Sequence length
236

Repeat regions
- III.3 (α-solenoid) [74-274]
  - Average period: 33.5

RAPHAEL results
SVM score: 5.5068486
Predicted period: 34.0272

Domains view
Pfam
RepeatsDB

Organism
Homo Sapiens

Host
Escherichia Coli Bl21(de3)

Sequence view

```
                76        86        96        106       116
 67  KQQLTEDGDSFLHLAIIHEEKALTMEVIRQVKGDLAFLNFQNNLQQTPLHLAVITNQPEI
               TTHHHHHTT SSSSS                    TT   HHHHHHHTT HHHH
127  AEALLGAGCDPELRDFRGNTPLHLACEQGCLASVGVLTQSCTTPHLHSILKATNYNGHTC
     H   S    S    TT    HHHHHHHHT HHHHHHHHHSTTTTSSS GGG    TT    HH
187  LHLASIHGYLGIVELLVSLGADVNAQEPCNGRTALHLAVDLQNPDLVSLLLKCGADVNRV
     HHHHHTT HHHHHHHHHHT  TT   TTT   HHHHHHHTT HHHHHHHHTTT  S
247  TYQGYSPYQLTWGRPSTRIQQQLGQLTLENLQMLPESEDEESYDTESEFTEFTEDE
     TT   GGGG TTSS HHHHHHHHTTS GGGSS     TTT
```

Structure view

Additional information

Entry ID
1iknD

Macromolecule type
Polypeptide(L)

Molecule
PROTEIN (I-KAPPA-B-ALPHA)

Method
Not available

Cross links
- PDB: 1ikn
- CATH:
  - 1.25.40.20 [73-293]
- SCOP:
  - 19171
- Pubmed: 9865694

RepeatsDB created by Tomás Di Domenico and Emilio Potenza for the BioComputing Lab in the University of Padua, Italy.

Figure 9.1: Screenshot of a sample RepeatsDB entry results page (PDB entry 1ikn). The sequence viewer and the structure viewer are shown in the middle of the page, towards the left and the right, respectively. Additional annotations at the structure and chain level are displayed, including links to other databases (above) and classifications (below).

are also available for programmatic access. Please refer to the ‘Help’ section of the website for details on using the RepeatsDB web services. Datasets can be downloaded in JSON, XML or text format using the browse function or RESTful web services.

9.4.1 Statistics

Analysis of the full PDB dataset yielded 10 745 repeats predicted by RAPHAEL, of which 2797 were finally classified into the RepeatsDB schema. Table 9.1 shows the distribution between classes and subclasses. The bulk of the annotations (<90%) consist of entries belonging to classes III and IV. No effort was made to balance the distribution of entries between classes in this initial release. As coverage increases in the future, we expect the balance to approximate the real distribution more closely, although it may be necessary to fine-tune RAPHAEL. Of the classified entries, 321 representatives of the entire dataset were annotated in detail with information about the start and end of repeat regions, repeat units and/or insertions (Table 9.1). It is interesting to note the different distribution of insertions between classes. Apparently, some classes such as β-solenoid (class III.1) or TIM barrels (class IV.1) have stronger propensity to accommodate insertions.

9.5 Conclusions and future work

RepeatsDB's goal is to provide the community with a resource for high-quality tandem repeat protein structure annotations. The user can either interactively analyse his proteins of interest via the user interface, or create and download datasets for offline use. Far from being a static classification process, the annotation effort for the initial RepeatsDB dataset alone already motivated the extension of the original classification schema [139]. Some of the curated structures, while clearly representing structural repeats, did not belong to any of the pre-defined subclasses. To allow them to be classified, subclasses IV.5 (α/β prism) and IV.6 (α-barrel) were added to the initial schema [139]. Class V also underwent a re-classification according to the secondary structure content of the single domain repeats ('beads') to allow a broader classification range beyond individual repeat families, as the list of possible beads-on-a-string folds may be considerably larger than currently appreciated. The 'other' subclass was also added to allow collection of repeats that do not fit into the current classification scheme. RepeatsDB provides the community with a previously unavailable opportunity to easily create datasets of tandem repeat proteins. The detailed annotation subset further presents a unique opportunity to better understand the nature of tandem repeat proteins.

Beyond its initial release, RepeatsDB is a continuous effort to expand, revise and improve tandem protein repeat annotations. Predictions for new PDB structures are simple and fully automated, allowing regular database updates every 3 months. Manual curation of new entries for inclusion is also ongoing, aiming at regular and steady updates. Options to involve the community into the annotation process through crowd-sourcing tools are currently being analysed. A main goal for future versions is the extension of the annotation of repeats at the sequence level, starting from annotation for intrinsically disordered regions from MobiDB [92]. We anticipate that RepeatsDB should prove valuable towards the understanding of the sequence–structure relationship in tandem repeat proteins and their evolutionary relationship.

Subclass	Name	Detailed	Classified (manually)	Classified (by similarity)	Predicted
I.1	Poly-alanine β structure	0	0	0	0
II.1	Collagen triple-helix	0	5	0	0
II.2	α helical coiled coil	23	38	69	0
III.1	β-solenoid	43	113	21	0
III.2	α/β solenoid	21	43	27	0
III.3	α-solenoid	48	246	631	0
III.4	Trimer of β spirals	7	0	13	0
III.5	Single layer anti-parallel β	4	3	0	0
IV.1	TIM-barrel	84	118	626	0
IV.2	β-barrel	8	1	8	0
IV.3	β-trefoil	20	0	29	0
IV.4	β-propeller	40	182	227	0
IV.5	α/β prism	0	17	0	0
IV.6	α-barrel	6	0	0	0
V.1	α-beads	2	1	0	0
V.2	β-beads	29	12	71	0
V.3	α/β-beads	3	3	1	0
V.other	Unknown subclass	3	0	4	0
UA	Unassigned	0	0	0	7948
	Total	321	749	1727	7948

Table 9.1: Statistics for RepeatsDB. The subclass name is shown together with the number of entries on each of the four annotation levels. Note that 'Unassigned' entries are automatically predicted by RAPHAEL and therefore not assigned to a specific class.

Part IV

Network-based tools for the analysis of protein data

10. RING: networking interacting residues, evolutionary information and energetics in protein structures

This chapter was first published in Martin AJM, Vidotto M, Boscariol F, Di Domenico T, Walsh I, et al. (2011) RING: networking interacting residues, evolutionary information and energetics in protein structures. Bioinformatics 27: 2003–2005.

10.1 Summary

Residue interaction networks (RINs) have been used in the literature to describe the protein 3D structure as a graph where nodes represent residues and edges physico–chemical interactions, e.g. hydrogen bonds or van-der-Waals contacts. Topological network parameters can be calculated over RINs and have been correlated with various aspects of protein structure and function. Here we present a novel web server, RING, to construct physico–chemically valid RINs interactively from PDB files for subsequent visualization in the Cytoscape platform. The additional structure-based parameters secondary structure, solvent accessibility and experimental uncertainty can be combined with information regarding residue conservation, mutual information and residue-based energy scoring functions. Different visualization styles are provided to facilitate visualization and standard plugins can be used to calculate topological parameters in Cytoscape. A sample use case analyzing the active site of glutathione peroxidase is presented.

10.2 Introduction

The last 15 years have seen the advent of network representation to tackle the complexity inherent in many problems in biology. By focusing on the network properties of a system it is often possible to gain a new level of insight into apparently unpredictable systems [189, 190]. The most studied example in biology is protein interaction networks, where nodes represent proteins and connections between nodes (physical) interactions. Tools have been developed to visualize and analyze such networks, with Cytoscape [191] probably being the most widely accepted standard platform due to its open structure and extendibility. More recently, there has been a growing interest in representing protein structures as so-called residue interaction networks (RINs) [192]. RINs consider single amino acids in the protein structure as nodes and connections as physico–chemical interactions, such as covalent bonds and non-covalent contacts (e.g. hydrogen bonds). The intuitive idea is to analyze a structure with the same approach as a protein interaction network in order to investigate whether the same rules apply. Several papers have already used RINs to analyze protein stability and folding, allosteric communication, enzyme catalysis or mutation effect prediction [193–199]. A more complete bibliography is available online.

Here, we present RING as a novel tool to generate RINs for use in Cytoscape. The tool was conceived to yield a simple, intuitive representation that is physico–chemically meaningful and integrates different types of structure-based information with evolutionary information and energy scoring functions.

10.3 Program overview

RING requires as input a valid PDB identifier or user specified PDB file and provides two user interfaces. A simple user interface with meaningful default values for most parameters is intended for the inexperienced user. A more complex interface where the user can specify exactly how the RIN should be defined is also provided. Three alternative types of network definitions are available, with closest atoms as default. Interactions are defined distinguishing disulfides, salt bridges, hydrogen bonds and aromatic interactions from generic van-der-Waals contacts. Structural features are generated for each node and include secondary structure, solvent accessibility and experimental uncertainty for X-ray structures (i.e. B factor and occupancy). Protein sequence conservation and mutual information [193, 200], determined from PSI-BLAST profiles, and conformational energy preferences determined with FRST and TAP score [201, 202] are to the best of our knowledge unique features of RING. The protocol uses standard programs to derive the data and is described on the website. The output consists in an archive file containing all the necessary network definition files for Cytoscape. Once generated, the RIN can be easily visualized in Cytoscape. Visualization filters are provided in RING to highlight different structural features through the Cytoscape VizMapper feature. Topological network properties can be analyzed with readily available plugins, e.g. NetworkAnalyzer [203]. Online help pages are provided for the web server together with an extensive documentation on the implemented file format and tutorials.

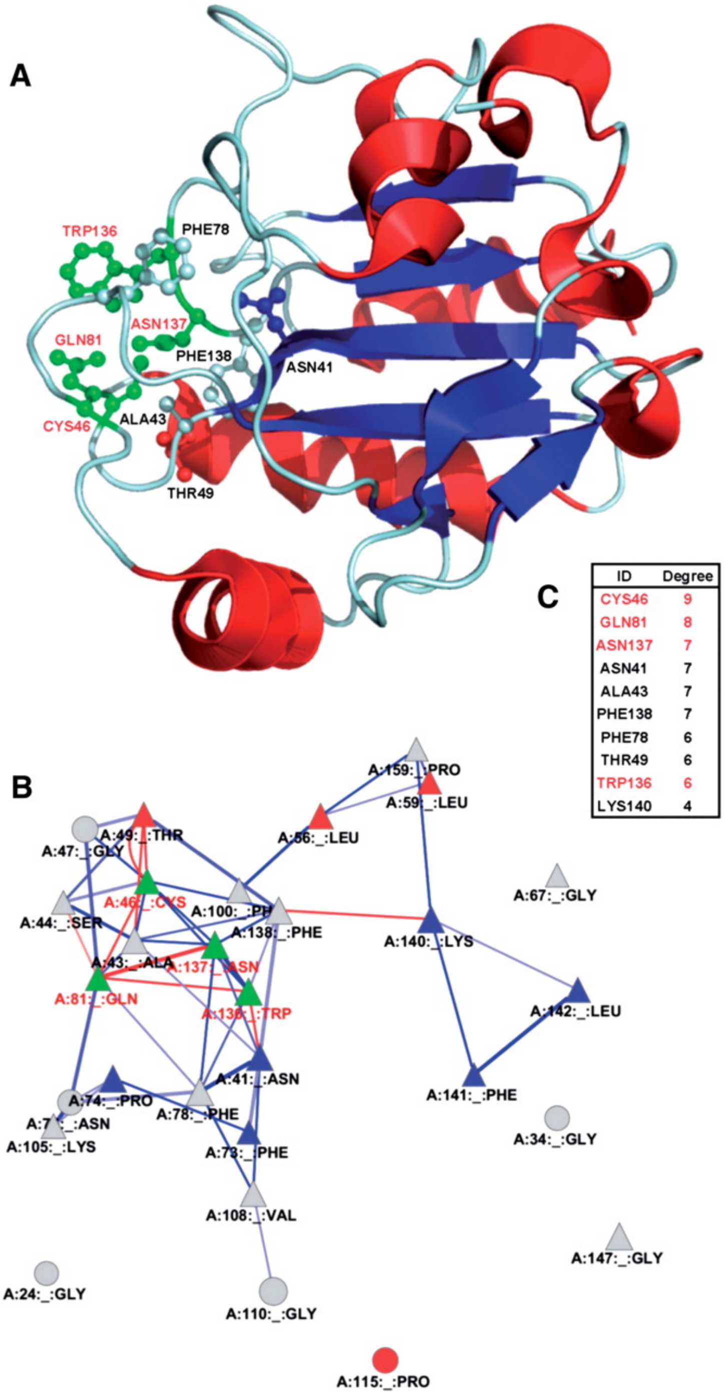
A special feature of RING is to generate meaningful sub-networks. It has been suggested that RIN complexity can be focused on essential interactions by limiting analysis to buried residues [196]. RING allows the user to limit the generated network to buried residues, conserved residues or both. Intuitively, limiting RIN analysis to a network of conserved residues will help to focus on the essential interactions for a protein active site as shown in the following case study.

10.4 Case study

The glutathione peroxidase (GPx) enzymes are an evolutionarily conserved family catalyzing the reduction of hydroperoxides to alcohols and the concomitant oxidation of thiols to disulfides [204]. Long thought to function by means of a catalytic triad, a detailed structural analysis of sequence conservation has recently suggested a fourth residue to be catalytically active which was experimentally confirmed [205].

Figure 10.1 shows the GPx structure and corresponding RING sub-network of conserved residues. The GPx active site can be easily identified as the nodes with the highest number of connections in the RIN. The combination of evolutionary conservation and basic network topology thus provides an easy way to replicate the work that led to the recent characterization of the GPx active site [205]. Moreover, mutual information provided by RING can serve to investigate cases of co-evolution between residues. An online tutorial explains the necessary steps to reproduce the steps to generate the RIN visualization. Several further example are available as part of the online tutorial.

In summary, RING is a novel web tool for use with Cytoscape designed for the visualization and analysis of protein structures in terms of physico–chemical interactions, evolutionary information and energetics while taking advantage of the powerful network paradigm. We anticipate this novel combination to facilitate the functional annotation of proteins.



ID	Degree
CYS46	9
GLN81	8
ASN137	7
ASN41	7
ALA43	7
PHE138	7
PHE78	6
THR49	6
TRP136	6
LYS140	4

Figure 10.1: Protein structure of human glutathione peroxidase 4 (PDB identifier 2obi). The 3D structure is shown as cartoons (A) colored by secondary structure. The active site is shown in balls and sticks and labeled. Lighter labels are used for the catalytic tetrad. The corresponding RING residue interaction network in Cytoscape is shown for positions with at least 80% sequence conservation (B) with circles representing exposed and diamonds buried residues. Nodes are colored by secondary structure. Edges are colored according to interaction type and width proportional to mutual information. Notice how the most highly connected nodes in the sub-network correspond to the active site and its immediate surroundings. The top 10 most connected nodes of the network are shown in inset (C).

11. PANADA: Protein Association Network Annotation, Determination and Analysis

This chapter was first published in Martin AJM, Walsh I, Di Domenico T, Mičetić I, Tosatto SCE (2013) PANADA: Protein Association Network Annotation, Determination and Analysis. PLoS ONE 8: e78383.

11.1 Summary

Increasingly large numbers of proteins require methods for functional annotation. This is typically based on pairwise inference from the homology of either protein sequence or structure. Recently, similarity networks have been presented to leverage both the ability to visualize relationships between proteins and assess the transferability of functional inference. Here we present PANADA, a novel toolkit for the visualization and analysis of protein similarity networks in Cytoscape. Networks can be constructed based on pairwise sequence or structural alignments either on a set of proteins or, alternatively, by database search from a single sequence. The Panada web server, executable for download and examples and extensive help files are available at URL: `http://protein.bio.unipd.it/panada/`.

11.2 Introduction

The main protein sequence databases contain tens of millions of entries with many more sequences becoming continuously available due to the numerous genome sequencing efforts [206]. Currently, most known proteins lack any functional annotation [207] and very little is known about the vast majority. There are many ongoing projects trying to reduce the gap between known proteins and their functional annotation either computationally [208] or experimentally [209]. Recently there has also been the first Critical Assessment of Function Annotation (CAFA) experiment to assess the performance of function prediction methods [210]. Most computational approaches rely on pairwise similarity to known proteins to suggest functional annotations derived by homology to annotated database entries [211, 212]. Current methods still lack tools for the visualization of their results, in order to aid in their interpretation, analysis and to aid experts with curation.

Precomputed pairwise comparisons with functional and structural annotations are available for instance in SIMAP [209], but one must build a similarity network by hand from this database. The Phytoscape framework [213] is available to build similarity networks, but it must be installed locally and offers a limited way to simplify large networks to be used in Cytoscape.

Protein sequence and structure similarity networks are bi-dimensional graphs where proteins are nodes with edges between them representing the pairwise similarity between the nodes they connect [214]. Such networks are increasingly being used for functional and structural protein annotation [215, 216]. They have also been used to detect errors in function annotation [217] and to study the evolution of multi-domain proteins [218]. Similarity networks complement phylogenetic trees and multiple sequence alignments, two more traditional approaches generally used to study and infer information derived from comparisons of protein sequences. The advantage of similarity networks is to leverage the human visual analytic skills to identify interesting patterns, e.g. of protein function or phylogenetic distribution, among a large protein set.

Here we describe PANADA, an automatic toolkit to visualize and study sequence and structure similarities between proteins to infer function by homology to other known proteins for use with the Cytoscape platform [191].

11.3 Implementation

PANADA is available as both a web server and a Linux executable for download. The toolkit has been designed to be flexible, allowing the user to consider either protein sequences or structures. In similarity networks, nodes are protein sequences or structures. Edges represent associations between nodes, with a weight for the degree of similarity between nodes. PANADA operates either with input from an entire group or a single protein. Analysis of a group of sequences or structures is used to establish relationships among them. When a single protein is provided, the server first performs a search for close sequences or structures in publicly available databases. This can be especially useful to suggest functional annotations of uncharacterized proteins or to study relationships among different proteins belonging to the same family. Either way, proteins in the set are compared to each other in a pairwise manner. The overall workflow of PANADA is shown in Fig. 11.1.

Sequence similarity in PANADA is computed using BLASTALL [42], reporting pairwise local alignments measuring the percentage of sequence identity, E-value, bit score and alignment length. When generating structural similarity networks, PANADA compares protein structures using either MUSTANG [219] or TMalign [220]. TMalign computes root mean square distance (RMSD) or its scaled version TMscore, after a residue-to-residue alignment based on structural similarity using dynamic programming between two Cα traces. MUSTANG computes RMSD based solely on structural correspondence after Cα trace superimposition. TM-score, with values ranging between [0,...,1], is more sensitive to the topology of the protein structures being compared and less affected by local variations than

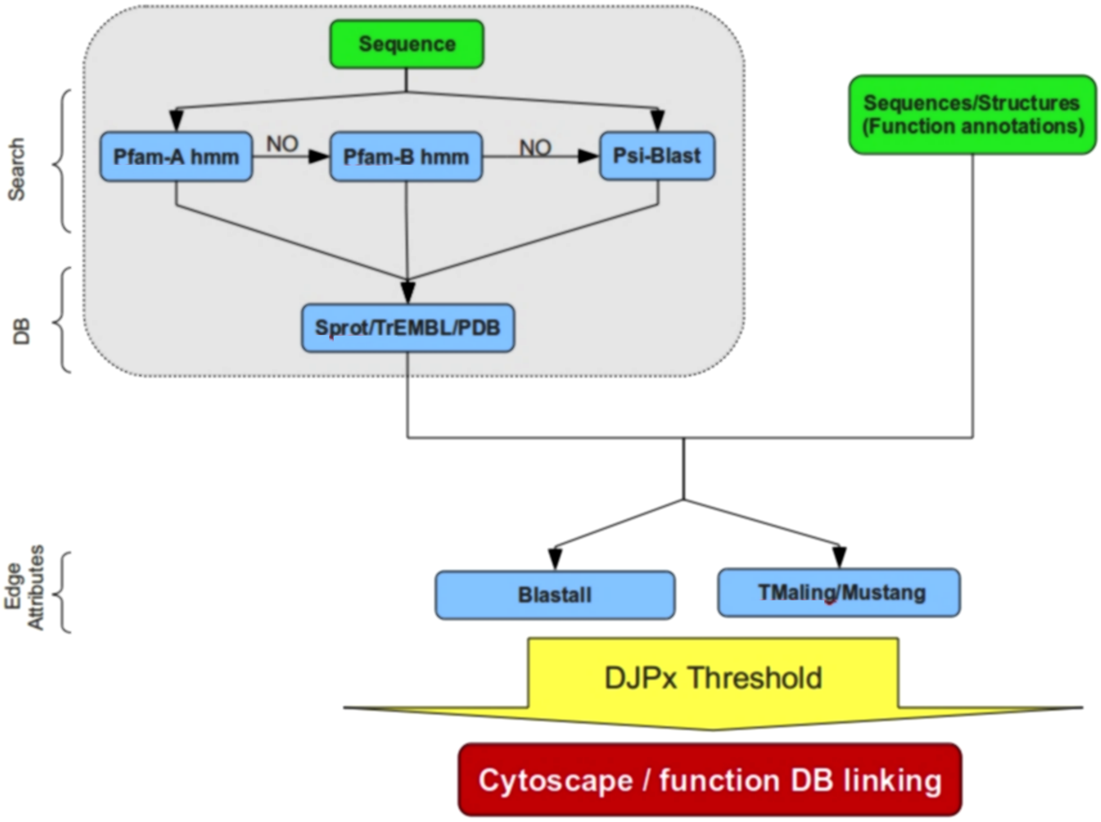


Figure 11.1: An overview of alternative steps performed by PANADA is shown. Depending on the input, a set of proteins, sequences or structures, or a single sequence may be submitted. In the latter case PANADA uses Pfam and/or BLAST to find homologs. Functional annotations may be provided by the user if a set of sequences or structures is uploaded. In all cases the network is generated for a given similarity threshold and with a maximum number of edges per node (DJPx algorithm). The output can then be used in Cytoscape for visual analysis.

RMSD. In general, TMscores below 0.17 mean the two compared proteins are structurally unrelated while proteins with scores above 0.5 share the same overall topology [221]. Due to the asymmetry of the comparison methods, each protein pair is compared in both directions, i.e. A vs. B and B vs. A, and the better value used.

PANADA can find related sequences from only a single input protein using Pfam protein family HMMs [219] and PSI-BLAST [42]. PDB [39] and the SwissProt and TrEMBL sections of UniProt [206] can be searched. In the initial step when using Pfam HMMs, a search is performed to identify to which Pfam family or families the query sequence belongs to. The search is first performed against Pfam-A, the manually curated set of protein families, and only when no significant match is found extended to Pfam-B. The identified Pfam HMMs are then used to search against the selected protein database for sequences containing the same functional region. PSI-BLAST is used only when the user desires to perform the initial database search using it or when there are no significant matches in Pfam. Proteins identified using PSI-BLAST could share only short stretches of local similarity or be biased due to the contents of protein databases [222]. PSI-BLAST parameters used in PANADA ensure that short regions are disregarded when query sequence and found hits share only one domain but the rest of the sequences are very different (i.e. belong to different domains). Due to the nature of the software used, multidomain proteins may however be problematic and require the user's judgment. Once the similarities between different proteins are computed, a selected measure is used to build the network by normalizing the similarity values in the range [0,..,1] (see online documentation). The highest value represents the shortest distance between the nodes in the network (greatest homology), and 0 the highest distance (lowest similarity). As PANADA produces multiple edges, the choice of measure is left to the user. This can be easily achieved by removing unwanted edges in Cytoscape through filtering.

Additionally, PANADA fetches GO [208] functional annotations for the compared proteins whenever available, i.e. for proteins with UniProt and PDB identifiers, with three different confidence levels. The user may select only experimental annotations, those considered reliable or everything. Reliable annotations include those inferred from electronic annotation (IEA) [208]. GO annotations of a node may be transferred to its neighbors without annotations, as the network explicitly represents similarity between proteins (property transfer by homology). GO terms can also be used to validate the annotations for single proteins or for all nodes in the network. If the same or related GO terms are present, these are more likely to be real. PANADA allows to color nodes according to their respective protein annotations in each of the three GO ontologies (molecular function, biological process and cellular component). Annotations for each protein are associated to their respective GO Slim and the most common GO Slim term for each node is given a hexadecimal ASCII color code that can be used to color the nodes in Cytoscape.

Since fully connected networks do not provide more information than standard pairwise comparison methods, e.g. BLAST search, removing edges in similarity networks increases the information content and enhances their interpretability [223]. PANADA implements two algorithms to reduce the number of connections present in a network. Edges are filtered either by leaving only edges representing high similarity (above a fixed threshold) or keeping the top X weighted edges for each node. The protocol used to keep the X top edges is a very simple modification of Prim's algorithm (DJPx). Prim's algorithm demonstrates that a minimum spanning tree (MSP) can be constructed on a graph (or network) by iteratively growing a tree from the minimum weight (i.e. highest similarity) edges connecting nodes not already attached to the MSP [224]. The DJPx algorithm used in PANADA generalizes Prim's algorithm by considering the top X edges instead of just one edge. Briefly, a single similarity measure is chosen to rank all normalized edges for all nodes from highest to lowest similarity. Starting from the highest similarity edge in the list, an edge is kept only if the nodes it connects do not yet have X edges. The selection is repeated until all nodes have X edges assigned and all remaining unassigned edges are removed. This ensures that the most relevant edges are kept and only low-quality ones are pruned. The two mechanisms to remove edges in the network, threshold and DJPx, can be used separately or combined. When both are combined in the final network, only the top X edges for each protein are kept while ensuring that they represent meaningful associations. When BLASTALL or TMalign are used to generate pairwise comparisons, connections between nodes are also removed if alignment coverage is lower than a predefined threshold.

The PANADA server produces a global output page with links to the downloadable output and Cytoscape network files as well as GO annotations and other relevant statistics. The output page includes normalized distance of direct neighbors to the query protein when using the automatic search option or all found GO terms for each protein when using a selected set of proteins. In both cases, the occurrence of each GO term belonging to proteins in the network is also shown.

11.4 Usage Examples

Figure 11.2 shows the results of a PANADA search of Thioredoxin fold class structures. The dataset contains the same structures as those used in a previous publication [225], with 159 protein chains at less than 60% sequence identity. The network was generated using MUSTANG to compare the 3D structures with default parameters. The results clearly separate the proteins into three main clusters representing the three main biological processes in which Thioredoxin fold class proteins are involved, showing how the overall structure of a protein chain relates to its catalytic function. This approach can be used to assign functional annotations inferred by homology of any query sequence and to determine possible misannotations and uncertainties within biologically related sets of proteins. For selected proteins of known structure, the PANADA analysis may be further combined with a residue interaction network analysis using RING [226] to determine the key structural components. Multiple sequence alignments will also provide complementary information about the proteins in the network and help to identify conserved residues that are likely to be related to protein function.

To demonstrate the use for single proteins with unknown function, Figure 11.3 shows the PANADA network for protein AC4 from Bean golden yellow mosaic virus (UniProt accession number P0DJX3), generated with by default parameters in the automatic search (SwissProt database and full GO annotation). This protein's existence was inferred by homology and although the genome is published [227], it was added to SwissProt on May 1, 2013. The two parts of Fig. 11.3 show the same network using the Cytoscape organic layout and only edges representing sequence identity. In the network, there are 22 different proteins. Eleven nodes have Cellular Component annotations, three "nucleus" (IEA) and eight "host cell plasma membrane" (IEA). Figure 3a shows the network colored according to the nodes Cellular Component annotations. According to the network, it is possible to infer AC4 cellular location GO terms to be "host cell plasma membrane" since the other annotations are in an unconnected cluster from the query protein. The same happens with the Biological Process annotations, see Figure 11.3b. Five proteins have Biological Process GO terms of two types, three with "DNA recombination; DNA repair; regulation of transcription, DNA-dependent; transcription, DNA-dependent" (all IEA); and two with "virus-host interaction" (IEA). For the same reasons as for Cellular Component one can infer AC4 terms as those proteins in the same subnet are likely to perform the same function.

To further explore the effects of parameters on using PANADA, we created

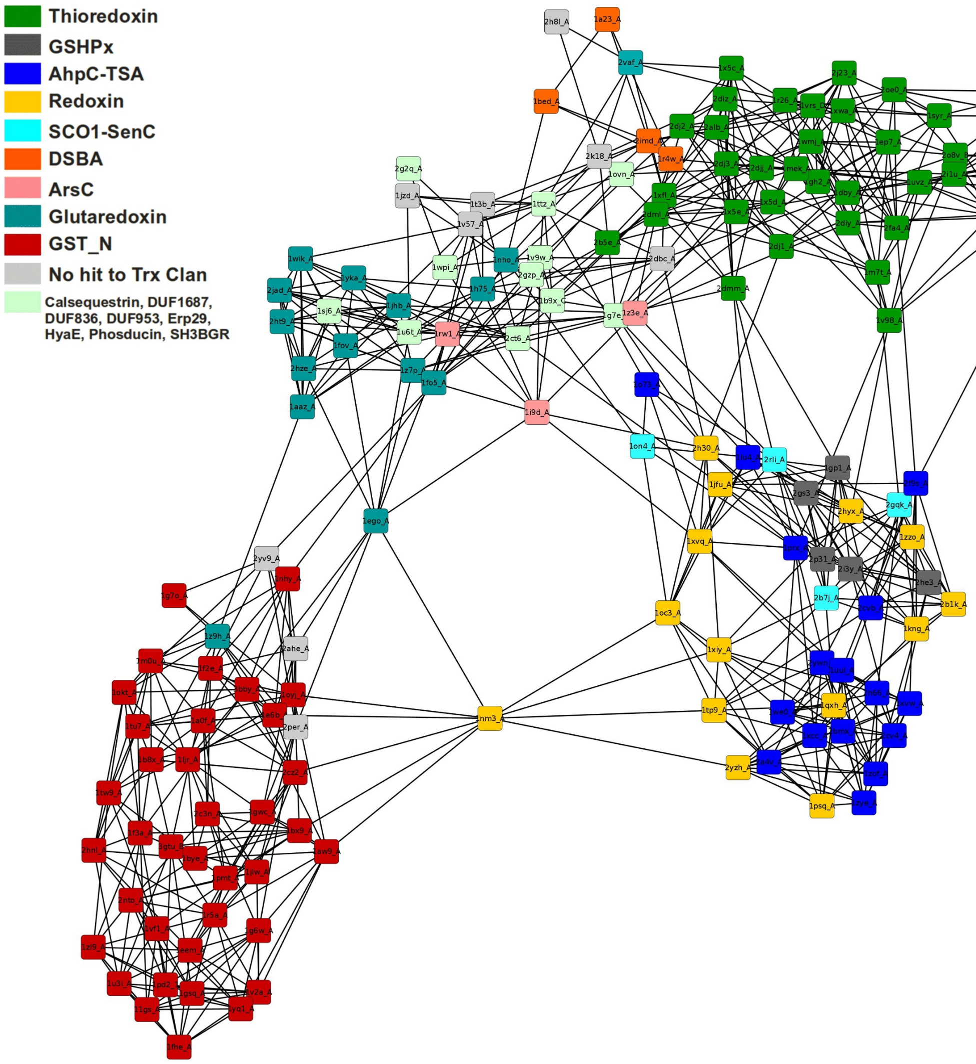


Figure 11.2: The PDB codes of structures from a previous publication [225] are used in PANADA to derive a network representation coloured by functional class. The organic layout was generated in Cytoscape with PANADA default parameters. The correspondence between colour codes and functional groups is shown in the upper left part. Notice how structures with the same functional class form tightly packed cluster separated from each other by a few connecting structures.

several sequence networks using the automatic search for E. coli protein YebC (PDB code 1KON). This protein belongs to Pfam-A family PF01709 and until recently lacked GO terms. Currently it has the following IEA GO terms: Biological Process “regulation of transcription, DNA-dependent”, Molecular Function “DNA binding” and Cellular Component “cytoplasm”. Figure 11.4 shows networks created using alignment coverage of at least 50% and a maximum number of top edges per node (DJPx threshold) of 100, 75, 50 and 25. All other parameters were at default values. The networks contain 2,088 different proteins, of which 571 have at least one GO term associated. 413 proteins have exactly the same annotations as YebC, 71 share the same Molecular Function annotation as YebC, and 89 have several different Molecular Function terms. As can be seen in Figure 11.4, edge reduction by changing the maximun number of nodes simplifies the network. It is interesting to note how the network decomposes into local sub-clusters with

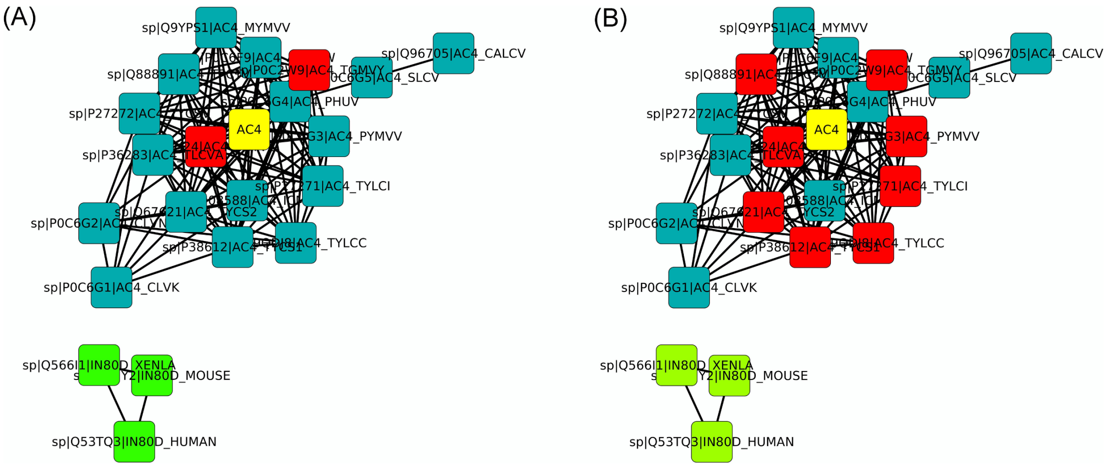


Figure 11.3: Starting from the viral protein AC4 (UniProt accession number P0DJX3), the constructed network is shown in Cytoscape with organic layout. (A) is colored by biological process with red for "virus-host interaction" (IEA) and green for "DNA recombination; DNA repair; regulation of transcription, DNA-dependent" (IEA). (B) is colored by cellular component with red for "host cell plasma membrane" (IEA) and green for "nucleus" (IEA). In both cases, the yellow color is used for the query protein AC4.

decreasing threshold values. Since the closest YebC neighbors in the figure share the same GO terms, they confirm the electronically inferred annotations assigned to YebC.

11.5 Conclusions

PANADA is a new online toolkit that generates protein similarity networks to be used with Cytoscape. PANADA allows the user to either automatically search similar sequences or to generate a network with a set of selected proteins. The similarity networks can be used for the visual analysis of similarity relationships among sequences or to asses functional annotation inferred from homology. PANADA complements other more traditional tools such as phylogenetic trees and multiple sequence alignments, making use of the user's visual skills to identify patterns that allow the inference of novel properties. The main advantages consist in the automatic search and annotation of proteins with GO terms from the database and the ability to choose two different approaches to prune the network topology. This produces networks that only contain edges for those pairwise comparisons that represent the highest similarities above a given threshold. Different utilities in Cytoscape, such as filters and the NetworkAnalyzer tools, add to the usefulness of PANADA providing the means for interpretation and analysis of similarity networks. PANADA automatically produces coloring based on the GO annotations of the proteins in the similarity network. Users can also define their own coloring scheme or their own annotations for each protein present in a network adding versatility to this toolkit. We anticipate PANADA to be of use for the visual analysis of protein function through similarity networks.

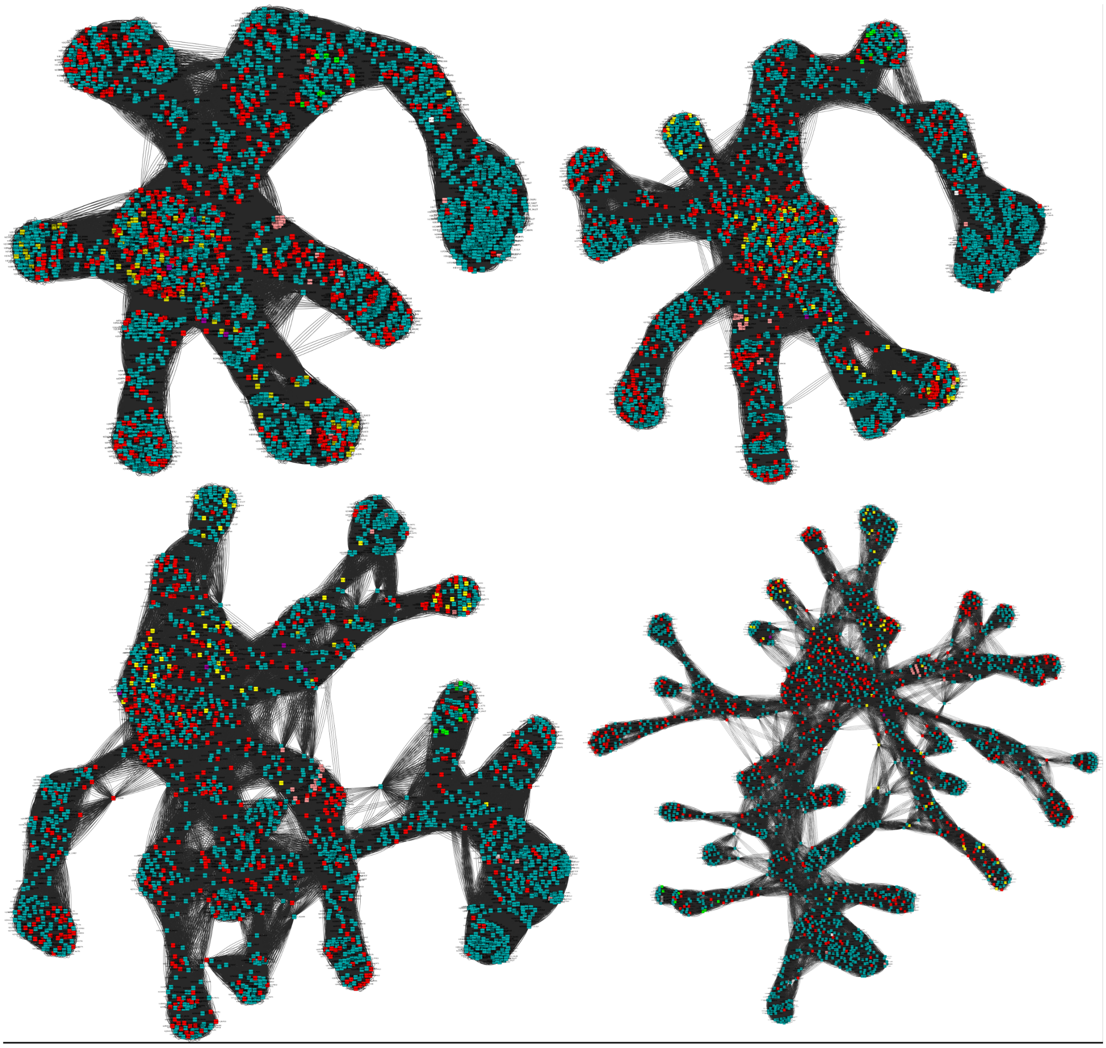


Figure 11.4: An automatic search for E. coli protein YebC (PDB code 1KON) represented with Cytoscape organic layout and different maximum number of top edges per node (DJPx). Related proteins are found in UniProt database using the Pfam-A family PF01709. Edges are shown for pairwise sequence identity greater than 40% and alignment coverage at least 50%. From left to right and top to bottom, the networks shown the top 100 edges per node, 75, 50 and 25. In all cases, nodes sharing the Biological Process GO terms electronically assigned (IEA) to the query protein are colored in red.

Part V

Conclusions

12. Conclusions

Intrinsic protein disorder is a challenging field of research. Because of its novelty, it constitutes a highly competitive topic. Because of its controversial nature, obtained results can be readily challenged when presented to the community. And because of its complexity, it demands an open mind by not settling for a definitive explanation.

My involvement in the development of intrinsic protein disorder predictors (see Chapter 2 and Chapter 3) during the early stages of my PhD, and my becoming familiar with the then recently developed MOBI web server [59], situated me on a good position to start my research into intrinsic protein disorder. Both of the predictors have proven to be valuable assets for the community, being highly accessed and still representing, at the time of writing, the state-of-the-art in the field. They score among the most accurate methods ever since their original release.

The complexity of the field and the difficulties to obtain definite answers despite the variety of data we were handling, motivated me to expand what was initially to be a database of NMR disorder annotations[1] into MobiDB: a comprehensive database of intrinsic protein disorder annotations [92]. Initially covering all proteins contained in the SwissProt database (roughly half a million at the time), it constituted the first published database integrating annotations of protein disorder. Interest by the community drove us to continue its development, which one and a half years after the original release date resulted in a second major release of MobiDB. This new version covers the full UniProt database (about 50 million sequences), includes additional annotations and a renovated user interface. The recent cross-linking of MobiDB from what is arguably the most relevant resource of protein sequence annotations, the UniProt database, has helped increase the visibility of the resource and constitutes a validation of its scientific relevance. We hope the latest updates and this higher visibility will also increase its usefulness for the community.

Several years of work on the topic have led me to hold a rather critical view on intrinsic protein disorder. In fact, I believe that it is quite possible that disorder does not exist when a protein is executing its function. The possibility of many of the cases classified as intrinsic protein disorder being actually artifacts of our methods for resolving proteins [99] seems also very appealing to my mind. I would argue that some of us may have been caught on a sort of "disorder paradigm"

[1] The original plans to develop a database of NMR disorder annotations, obtained from the MOBI web server, were what gave origin to the name "MobiDB". Even though the database was expanded to cover all types of disorder annotations, we decided to keep the name unchanged.

paralysis. This does not necessarily mean, however, that the disorder signal we detect when a protein is in isolation or under certain experimental conditions is not relevant. Hub proteins and their promiscuous binding to many partners have been demonstrated to play key roles in cell signalling and regulation. It is highly likely that, while waiting for a certain partner, these proteins may exist for a certain amount of time in a disordered form. I believe that disorder is quite likely a state that most proteins, if not all, undergo at least partially at a certain stage of their cycles, and that the frequency and/or extensiveness of this state is most likely linked to the diversity of their function.

It is my feeling that we may be reaching a point where we will have exhausted the amount of novel information we can extract from currently available intrinsic protein disorder annotations. There is a great need for more experimental data, in particular the kind that can be obtained in conditions as close as possible as those found in a living organism (e.g. in-cell NMR). I am convinced that the availability of this data will allows us to better understand the role that disorder plays in the life of proteins.

The analysis of tandem repeat proteins was done following a similar workflow to that which was applied to the study of intrinsic protein disorder. RAPHAEL (Chapter 8) picked up the flag of the disorder predictors as our data generator, by allowing us to quickly obtain a set of candidate repeats from the Protein Data Bank. Manual curation efforts were then done on this dataset to classify and annotate true repeat proteins. To the best of our knowledge, it is the first attempt to produce a high quality, manually curated set of annotations providing details up to the repeat unit level (i.e. the start and end residue for each repeat unit inside a repeat region). This level of detail, backed by the assurance of the quality of the data associated with a reviewed manual effort, should prove very useful for understanding the particularities of each class of repeat proteins. The curated dataset was made available to the public through the RepeatsDB database (Chapter 9). The resource provides a series of methods of access ranging from simple graphical usage to advanced dataset generation through web services.

Unlike intrinsic protein disorder, which has been actively studied for decades, the field of tandem repeats is still an open area with many questions still to be answered. The fact that tandem repeat proteins have well defined structures whose folding mechanisms are not fully understood yet, makes the topic very interesting from the structural biology point of view. Their known involvement in neurodegenerative diseases make them a highly interesting target for biomedical research. Many of the analysis methods and tools that we have successfully applied to intrinsic protein disorder can easily be adapted to work on tandem repeat proteins. Without a doubt, the next few years will be witness to a great expansion of available data and resources regarding this type of protein structures.

As described above, both in the case of intrinsic protein disorder as well as in that of tandem repeats, we have shown that it is possible to classify the already available data into more biologically specific subsets. Such is the case of the different types of disorder presented in Chapter 6, and of the classification of tandem repeats in Chapter 9. We believe that the development of more specific tools, which either understand the existence of these subgroups or even deal with only a few of them a the time, will be essential to the advancement of these fields. These more specific tools will in turn allow us to obtain more specific, better quality data. Adding experimental data to the equation, we would then be presented with the cycle of development that I believe is the ideal workflow for bioinformatics research, illustrated in Figure 12.1.

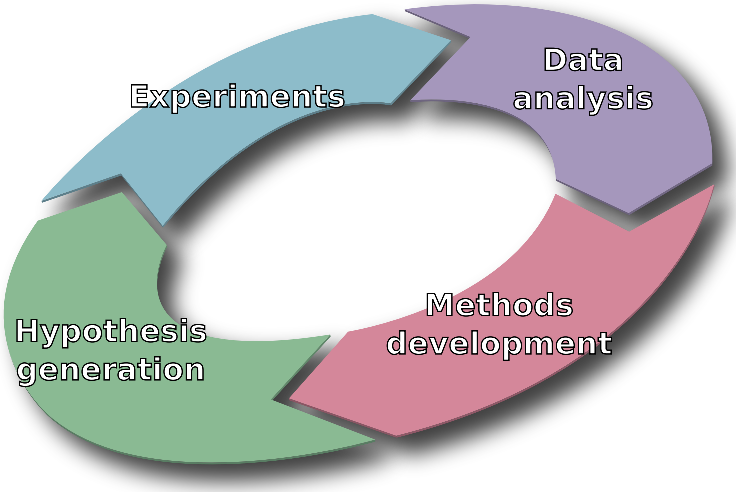


Figure 12.1: Bioinformatics development cycle. The ideal cycle of development for bioinformatics, where experimental work is used to create or improve computational methods, whose output is eventually used as the basis for further experimental work (image modified from the original found at https://commons.wikimedia.org/wiki/File:PDCA_Cycle.svg).

The network analysis tools presented in Chapter 10 and Chapter 11, if quite similar in their respective underlying concepts, are applied to very different biological problems. RING presents the user with a tool to easily represent and analyse a protein structure as a graph. PANADA allows the user to easily construct and visualise protein similarity networks. Our goal when developing these tools was to put an abstraction layer between the complexity of the data and the researcher, which at the same time makes use of state-of-the-art techniques for data processing and visualisation. I believe we have reached that goal. RING's popularity[2] (it is as of date one of the most used resources in our laboratory) shows us that there is indeed room to fill when it comes to the development of tools that simplify the analysis of protein data through computational methods, and that these tools can be quickly embraced by the community. It would be very interesting -and quite straightforward- to apply these network-based methods to further the understanding of the different categories of intrinsically disordered and tandem repeat proteins.

[2] Given PANADA's recent date of publication, I consider it to be too early to judge its impact based on usage statistics.

Bibliography

[1] "The data deluge". In: *Nature Cell Biology* 14.8 (8 Aug. 2012), pp. 775–775. DOI: 10.1038/ncb2558 (see p. 3).

[2] K R Hess et al. "Microarrays: handling the deluge of data and extracting reliable information". In: *Trends in biotechnology* 19.11 (11 Nov. 2001), pp. 463–468 (see p. 3).

[3] Ramiro Logares et al. "Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches". In: *Journal of microbiological methods* 91.1 (1 Oct. 2012), pp. 106–113. DOI: 10.1016/j.mimet.2012.07.017 (see p. 3).

[4] Timothy H. Vines et al. "The Availability of Research Data Declines Rapidly with Article Age". In: *Current Biology* (2013). DOI: 10.1016/j.cub.2013.11.014 (see p. 3).

[5] B Hesper and P Hogeweg. "Bioinformatica: een werkconcept". In: *Kameleon* 1.6 (6 1970), pp. 28–29 (see p. 3).

[6] N M Luscombe, D Greenbaum, and M Gerstein. "What is bioinformatics? A proposed definition and overview of the field". In: *Methods of information in medicine* 40.4 (4 2001), pp. 346–358 (see pp. 3, 4).

[7] Mark Gerstein. *What is Bioinformatics?* 1998. URL: http://bioinfo.mbb.yale.edu/what-is-it/ (see p. 4).

[8] F. Orosz and J. Ovadi. "Proteins without 3D structure: definition, detection and beyond". In: *Bioinformatics* 27 (2011), pp. 1449–1454 (see pp. 5, 38).

[9] Peter E Wright and H.Jane Dyson. "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm". In: *Journal of Molecular Biology* 293 (2 Oct. 22, 1999), pp. 321–331. DOI: 10.1006/jmbi.1999.3110 (see pp. 10, 19, 33, 38).

[10] H. Jane Dyson and Peter E. Wright. "Intrinsically unstructured proteins and their functions". In: *Nature Reviews Molecular Cell Biology* 6 (3 Mar. 2005), pp. 197–208. DOI: 10.1038/nrm1589 (see p. 10).

[11] Peter Tompa and Monika Fuxreiter. "Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions". In: *Trends in Biochemical Sciences* 33 (1 2008), pp. 2–8. DOI: 10.1016/j.tibs.2007.10.003 (see pp. 10, 20, 38).

[12] Peter Tompa. “Intrinsically unstructured proteins”. In: *Trends in Biochemical Sciences* 27 (10 Oct. 1, 2002), pp. 527–533. DOI: 10.1016/S0968-0004(02)02169-2 (see pp. 10, 20, 38).

[13] A. Keith Dunker et al. “Intrinsic Disorder and Protein Function†”. In: *Biochemistry* 41 (21 May 1, 2002), pp. 6573–6582. DOI: 10.1021/bi012159+ (see p. 10).

[14] Michael A. Weiss et al. “Folding transition in the DMA-binding domain of GCN4 on specific binding to DNA”. In: *Nature* 347 (6293 Oct. 11, 1990), pp. 575–578. DOI: 10.1038/347575a0 (see p. 10).

[15] Peter Tompa et al. “Close encounters of the third kind: disordered domains and the interactions of proteins”. In: *BioEssays* 31 (3 2009), pp. 328–335. DOI: 10.1002/bies.200800151 (see pp. 10, 20).

[16] A K Dunker et al. “Intrinsic protein disorder in complete genomes”. In: *Genome informatics. Workshop on Genome Informatics* 11 (2000), pp. 161–171 (see pp. 10, 20).

[17] J.J. Ward et al. “Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life”. In: *Journal of Molecular Biology* 337 (3 Mar. 26, 2004), pp. 635–645. DOI: 10.1016/j.jmb.2004.02.002 (see pp. 10, 14, 20, 29, 33, 34, 38, 47).

[18] Christian Schaefer, Avner Schlessinger, and Burkhard Rost. “Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be”. In: *Bioinformatics* 26 (5 Mar. 1, 2010), pp. 625–631. DOI: 10.1093/bioinformatics/btq012 (see p. 10).

[19] Jessica Siltberg-Liberles. “Evolution of Structurally Disordered Proteins Promotes Neostructuralization”. In: *Molecular Biology and Evolution* 28 (1 2011), pp. 59–62. DOI: 10.1093/molbev/msq291 (see p. 10).

[20] S. Lise and D. T. Jones. “Sequence patterns associated with disordered regions in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 58 (1 2005), pp. 144–150. DOI: 10.1002/prot.20279 (see p. 10).

[21] Michail Yu Lobanov et al. “ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder”. In: *Nucleic Acids Research* 38 (suppl 1 2010), pp. D283–D287. DOI: 10.1093/nar/gkp963 (see p. 10).

[22] Robert B. Russell and Toby J. Gibson. “A careful disorderliness in the proteome: Sites for interaction and targets for future therapies”. In: *FEBS Letters* 582 (8 Apr. 9, 2008), pp. 1271–1275. DOI: 10.1016/j.febslet.2008.02.027 (see p. 10).

[23] Toby J. Gibson. “Cell regulation: determined to signal discrete cooperation”. In: *Trends in Biochemical Sciences* 34 (10 Oct. 2009), pp. 471–482. DOI: 10.1016/j.tibs.2009.06.007 (see p. 10).

[24] Francesca Diella et al. “Understanding eukaryotic linear motifs and their role in cell signaling and regulation”. In: *Frontiers in bioscience: a journal and virtual library* 13 (2008), pp. 6580–6603 (see p. 10).

[25] Monika Fuxreiter, Peter Tompa, and István Simon. "Local structural disorder imparts plasticity on linear motifs". In: *Bioinformatics* 23 (8 Apr. 15, 2007), pp. 950–956. DOI: 10.1093/bioinformatics/btm035 (see p. 10).

[26] Cathryn M. Gould et al. "ELM: the status of the 2010 eukaryotic linear motif resource". In: *Nucleic Acids Research* 38 (suppl 1 2010), pp. D167–D180. DOI: 10.1093/nar/gkp1016 (see pp. 10, 12).

[27] Eugene Melamud and John Moult. "Evaluation of disorder predictions in CASP5". In: *Proteins: Structure, Function, and Bioinformatics* 53 (S6 2003), pp. 561–565. DOI: 10.1002/prot.10533 (see p. 10).

[28] Vladimir N Uversky. "What does it mean to be natively unfolded?" In: *European journal of biochemistry / FEBS* 269 (1 2002), pp. 2–12 (see pp. 10, 20).

[29] Zoran Obradovic et al. "Exploiting heterogeneous sequence properties improves prediction of protein disorder". In: *Proteins* 61 Suppl 7 (2005), pp. 176–182. DOI: 10.1002/prot.20735 (see pp. 10, 20, 38).

[30] Zsuzsanna Dosztányi et al. "The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins". In: *Journal of Molecular Biology* 347 (4 Apr. 8, 2005), pp. 827–839. DOI: 10.1016/j.jmb.2005.01.071 (see pp. 10, 14, 20, 34).

[31] Jianlin Cheng, Michael J. Sweredoski, and Pierre Baldi. "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data". In: *Data Mining and Knowledge Discovery* 11 (3 Nov. 1, 2005), pp. 213–222. DOI: 10.1007/s10618-005-0001-y (see pp. 10, 11, 20).

[32] David T. Jones and Jonathan J. Ward. "Prediction of disordered regions in proteins from position specific score matrices". In: *Proteins: Structure, Function, and Bioinformatics* 53 (S6 2003), pp. 573–578. DOI: 10.1002/prot.10528 (see p. 10).

[33] Rune Linding et al. "Protein disorder prediction: implications for structural proteomics". In: *Structure (London, England: 1993)* 11 (11 Nov. 2003), pp. 1453–1459 (see pp. 10, 20, 39, 40).

[34] Alessandro Vullo et al. "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines". In: *Nucleic Acids Research* 34 (suppl 2 July 1, 2006), W164–W168. DOI: 10.1093/nar/gkl166 (see pp. 10, 11, 20, 25, 30).

[35] Liam J. McGuffin. "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models". In: *Bioinformatics* 24 (16 Aug. 15, 2008), pp. 1798–1804. DOI: 10.1093/bioinformatics/btn326 (see pp. 10, 20).

[36] Marcin J. Mizianty et al. "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources". In: *Bioinformatics* 26 (18 Sept. 15, 2010), pp. i489–i496. DOI: 10.1093/bioinformatics/btq373 (see pp. 10, 20, 21, 25, 26).

[37] Bin Xue et al. “PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids”. In: *Biochimica et biophysica acta* 1804 (4 Apr. 2010), pp. 996–1010. DOI: 10.1016/j.bbapap.2010.01.011 (see pp. 10, 14, 20, 21).

[38] Avner Schlessinger et al. “Improved Disorder Prediction by Combination of Orthogonal Approaches”. In: *PLoS ONE* 4 (2 Feb. 11, 2009), e4433. DOI: 10.1371/journal.pone.0004433 (see pp. 10, 20).

[39] Helen Berman et al. “The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data”. In: *Nucleic Acids Research* 35 (suppl 1 2007), pp. D301–D303. DOI: 10.1093/nar/gkl971 (see pp. 11, 20, 23, 34, 38, 41, 54, 80, 111).

[40] Megan Sickmeier et al. “DisProt: the Database of Disordered Proteins”. In: *Nucleic Acids Research* 35 (suppl 1 2007), pp. D786–D793. DOI: 10.1093/nar/gkl893 (see pp. 11, 20, 23, 34, 38, 41, 54).

[41] Sven Mika and Burkhard Rost. “UniqueProt: creating representative protein sequence sets”. In: *Nucleic Acids Research* 31 (13 July 1, 2003), pp. 3789–3791. DOI: 10.1093/nar/gkg620 (see pp. 11, 13, 24).

[42] Stephen F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic Acids Research* 25 (17 Sept. 1, 1997), pp. 3389–3402. DOI: 10.1093/nar/25.17.3389 (see pp. 11, 20, 23, 34, 67, 110, 111).

[43] Gianluca Pollastri and Aoife McLysaght. “Porter: a new, accurate server for protein secondary structure prediction”. In: *Bioinformatics* 21 (8 Apr. 15, 2005), pp. 1719–1720. DOI: 10.1093/bioinformatics/bti203 (see pp. 11, 12, 21, 67).

[44] Gianluca Pollastri et al. “Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information”. In: *BMC Bioinformatics* 8 (1 June 14, 2007), p. 201. DOI: 10.1186/1471-2105-8-201 (see p. 11).

[45] Pierre Baldi and Gianluca Pollastri. “The Principled Design of Large-scale Recursive Neural Network Architectures–dag-rnns and the Protein Structure Prediction Problem”. In: *J. Mach. Learn. Res.* 4 (Dec. 2003), pp. 575–602. DOI: 10.1162/153244304773936054 (see p. 11).

[46] Orly Noivirt-Brik, Jaime Prilusky, and Joel L. Sussman. “Assessment of disorder predictions in CASP8”. In: *Proteins: Structure, Function, and Bioinformatics* 77 (S9 2009), pp. 210–216. DOI: 10.1002/prot.22586 (see pp. 12, 13, 20, 25, 26, 28).

[47] Peter Sollich and Anders Krogh. *Learning with ensembles: How over-fitting can be useful.* 1996 (see p. 12).

[48] Mario Albrecht et al. “Simple consensus procedures are effective and sufficient in secondary structure prediction”. In: *Protein Engineering* 16 (7 July 1, 2003), pp. 459–462. DOI: 10.1093/protein/gzg063 (see p. 12).

[49] Fernanda L Sirota et al. "Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset". In: *BMC Genomics* 11 (Suppl 1 Feb. 10, 2010), S15. DOI: 10.1186/1471-2164-11-S1-S15 (see pp. 12, 13, 20, 24, 29, 30).

[50] Matthew J. Hemsley et al. "Linear motifs in the C-terminus of D. melanogaster cryptochrome". In: *Biochemical and Biophysical Research Communications* 355 (2 Apr. 6, 2007), pp. 531–537. DOI: 10.1016/j.bbrc.2007.01.189 (see p. 15).

[51] Peter Vanhee et al. "Protein-Peptide Interactions Adopt the Same Structural Motifs as Monomeric Protein Folds". In: *Structure* 17 (8 Aug. 12, 2009), pp. 1128–1136. DOI: 10.1016/j.str.2009.06.013 (see p. 16).

[52] Luca Marsella et al. "REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform". In: *Bioinformatics* 25 (12 June 15, 2009), pp. i289–i295. DOI: 10.1093/bioinformatics/btp232 (see pp. 16, 22, 80, 81, 86–90, 92, 96).

[53] Antonio Trovato, Flavio Seno, and Silvio C. E. Tosatto. "The PASTA server for protein aggregation prediction". In: *Protein Engineering Design and Selection* 20 (10 Oct. 1, 2007), pp. 521–523. DOI: 10.1093/protein/gzm042 (see p. 16).

[54] A. Keith Dunker and Zoran Obradovic. "The protein trinity—linking function and disorder". In: *Nature Biotechnology* 19 (9 Sept. 2001), pp. 805–806. DOI: 10.1038/nbt0901-805 (see pp. 19, 38).

[55] A Keith Dunker et al. "Function and structure of inherently disordered proteins". In: *Current Opinion in Structural Biology* 18 (6 Dec. 2008), pp. 756–764. DOI: 10.1016/j.sbi.2008.10.002 (see p. 20).

[56] Avner Schlessinger et al. "Protein disorder — a breakthrough invention of evolution?" In: *Current Opinion in Structural Biology* 21 (3 June 2011), pp. 412–418. DOI: 10.1016/j.sbi.2011.03.014 (see pp. 20, 21, 26, 29, 30, 34, 38).

[57] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. "Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept". In: *Annual Review of Biophysics* 37 (1 2008), pp. 215–246. DOI: 10.1146/annurev.biophys.37.032807.125924 (see pp. 20, 38).

[58] Zsuzsanna Dosztányi et al. "Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution". In: *Journal of Proteome Research* 5 (11 Nov. 1, 2006), pp. 2985–2995. DOI: 10.1021/pr060171o (see pp. 20, 38).

[59] Alberto J. M. Martin, Ian Walsh, and Silvio C. E. Tosatto. "MOBI: a web server to define and visualize structural mobility in NMR protein ensembles". In: *Bioinformatics* 26 (22 Nov. 15, 2010), pp. 2916–2917. DOI: 10.1093/bioinformatics/btq537 (see pp. 20, 25, 28, 34, 38, 119).

[60] Slobodan Vucetic et al. "Flavors of protein disorder". In: *Proteins: Structure, Function, and Bioinformatics* 52 (4 2003), pp. 573–584. DOI: 10.1002/prot.10437 (see pp. 20, 33).

[61] Vladimir N. Uversky, Joel R. Gillespie, and Anthony L. Fink. "Why are "natively unfolded" proteins unstructured under physiologic conditions?" In: *Proteins: Structure, Function, and Bioinformatics* 41 (3 2000), pp. 415–427. DOI: 10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7 (see p. 20).

[62] Oxana V Galzitskaya, Sergiy O Garbuzynskiy, and Michail Y Lobanov. "Prediction of Amyloidogenic and Disordered Regions in Protein Chains". In: *PLoS Comput Biol* 2 (12 Dec. 29, 2006), e177. DOI: 10.1371/journal.pcbi.0020177 (see p. 20).

[63] Michail Yu Lobanov and Oxana V. Galzitskaya. "The Ising model for prediction of disordered residues from protein sequence alone". In: *Physical Biology* 8 (3 June 1, 2011), p. 035004. DOI: 10.1088/1478-3975/8/3/035004 (see p. 20).

[64] Jaime Prilusky et al. "FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded". In: *Bioinformatics* 21 (16 Aug. 15, 2005), pp. 3435–3438. DOI: 10.1093/bioinformatics/bti537 (see p. 20).

[65] Shuichi Hirose et al. "POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions". In: *Bioinformatics* 23 (16 Aug. 15, 2007), pp. 2046–2053. DOI: 10.1093/bioinformatics/btm302 (see p. 20).

[66] Takashi Ishida and Kengo Kinoshita. "PrDOS: prediction of disordered protein regions from amino acid sequence". In: *Nucleic Acids Research* 35 (suppl 2 July 1, 2007), W460–W464. DOI: 10.1093/nar/gkm363 (see p. 20).

[67] Zheng Rong Yang et al. "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins". In: *Bioinformatics* 21 (16 Aug. 15, 2005), pp. 3369–3376. DOI: 10.1093/bioinformatics/bti534 (see p. 20).

[68] Ian Walsh et al. "CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs". In: *Nucleic Acids Research* 39 (suppl 2 July 1, 2011), W190–W196. DOI: 10.1093/nar/gkr411 (see pp. 20–22, 25, 29, 30, 34, 38).

[69] William R. Atchley et al. "Solving the protein sequence metric problem". In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (18 May 3, 2005), pp. 6395–6400. DOI: 10.1073/pnas.0408677102 (see p. 22).

[70] Shuichi Kawashima, Hiroyuki Ogata, and Minoru Kanehisa. "AAindex: Amino Acid Index Database". In: *Nucleic Acids Research* 27 (1 1999), pp. 368–369. DOI: 10.1093/nar/27.1.368 (see p. 22).

[71] David T Jones and Mark B Swindells. “Getting the most from PSI–BLAST”. In: *Trends in Biochemical Sciences* 27 (3 Mar. 1, 2002), pp. 161–164. DOI: 10.1016/S0968-0004(01)02039-4 (see p. 24).

[72] S. Velankar et al. “E-MSD: an integrated data resource for bioinformatics”. In: *Nucleic Acids Research* 33 (suppl 1 2005), pp. D262–D265. DOI: 10.1093/nar/gki058 (see pp. 24, 34).

[73] Michail Yu. Lobanov et al. “Library of Disordered Patterns in 3D Protein Structures”. In: *PLoS Comput Biol* 6 (10 Oct. 14, 2010), e1000958. DOI: 10.1371/journal.pcbi.1000958 (see p. 26).

[74] A. Keith Dunker et al. “The unfoldomics decade: an update on intrinsically disordered proteins”. In: *BMC Genomics* 9 (Suppl 2 Sept. 16, 2008), S1. DOI: 10.1186/1471-2164-9-S2-S1 (see pp. 33, 38).

[75] Peter Tompa and Alan Fersht. *Structure and Function of Intrinsically Disordered Proteins*. CRC Press, Dec. 12, 2010. 362 pp. (see p. 33).

[76] Jeremy Bellay et al. “Bringing order to protein disorder through comparative genomics and genetic interactions”. In: *Genome Biology* 12 (2 2011), R14. DOI: 10.1186/gb-2011-12-2-r14 (see pp. 34, 35, 42).

[77] Ian Walsh et al. “ESpritz: accurate and fast prediction of protein disorder”. In: *Bioinformatics* 28 (4 Feb. 15, 2012), pp. 503–509. DOI: 10.1093/bioinformatics/btr682 (see pp. 34, 38–40, 67, 68, 70).

[78] The UniProt Consortium. “Ongoing and future developments at the Universal Protein Resource”. In: *Nucleic Acids Research* 39 (suppl 1 2011), pp. D214–D219. DOI: 10.1093/nar/gkq1020 (see pp. 34, 54).

[79] Robert D. Finn et al. “The Pfam protein families database”. In: *Nucleic Acids Research* 38 (suppl 1 2010), pp. D211–D222. DOI: 10.1093/nar/gkp985 (see p. 34).

[80] Wolfgang Kabsch and Christian Sander. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. In: *Biopolymers* 22 (12 1983), pp. 2577–2637. DOI: 10.1002/bip.360221211 (see pp. 34, 41).

[81] Mark Wells et al. “Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain”. In: *Proceedings of the National Academy of Sciences* 105 (15 Apr. 15, 2008), pp. 5762–5767. DOI: 10.1073/pnas.0801353105 (see p. 35).

[82] P. Tompa. “Unstructural biology coming of age”. In: *Curr Opin Struct Biol* 21 (2011), pp. 419–425 (see p. 38).

[83] R. Pancsa and P. Tompa. “Structural disorder in eukaryotes”. In: *PLoS ONE* 7 (2012), e34687 (see pp. 38, 48).

[84] J. H. Fong et al. “Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis”. In: *PLoS Comput Biol* 5 (2009), e1000316 (see p. 38).

[85] V. N. Uversky et al. “Prediction of intrinsic disorder and its use in functional proteomics”. In: *Methods Mol. Biol* 408 (2007), pp. 69–92 (see p. 38).

[86] S. Fukuchi et al. “Intrinsically Disordered proteins with Extensive Annotations and Literature”. In: *Nucleic Acids Res* 40 (2012), pp. D507–511 (see pp. 38, 41).

[87] B. W. Brandt, J. Heringa, and J. A. M. Leunissen. “SEQATOMS: a web tool for identifying missing regions in PDB in sequence context”. In: *Nucleic Acids Res* 36 (2008), W255–259 (see pp. 38, 54).

[88] S. Vucetic et al. “DisProt: a database of protein disorder”. In: *Bioinformatics* 21 (2005), pp. 137–140 (see p. 38).

[89] X. Deng, J. Eickholt, and J. Cheng. “A comprehensive overview of computational protein disorder prediction methods”. In: *Mol Biosyst* 8 (2012), pp. 114–121 (see p. 38).

[90] B. Monastyrskyy et al. “Evaluation of disorder predictions in CASP9”. In: *Proteins* 79 (Suppl 10 2011), pp. 107–118 (see pp. 38, 70).

[91] Z. Dosztanyi et al. “IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content”. In: *Bioinformatics* 21 (2005), pp. 3433–3434 (see pp. 39, 40).

[92] T. Di Domenico et al. “MobiDB: a comprehensive database of intrinsic protein disorder annotations”. In: *Bioinformatics* 28 (2012), pp. 2080–2081 (see pp. 39, 42, 75, 100, 119).

[93] M. Tagari et al. “E-MSD: improving data deposition and structure quality”. In: *Nucleic Acids Res* 34 (2006), pp. D287–290 (see pp. 40, 41).

[94] The UniProt Consortium. “Reorganizing the protein space at the Universal Protein Resource (UniProt)”. In: *Nucleic Acids Res* 40 (2012), pp. D71–75 (see pp. 40, 41).

[95] M. Punta et al. “The Pfam protein families database”. In: *Nucleic Acids Res* 40 (2012), pp. D290–301 (see pp. 41, 96).

[96] A. M. Altenhoff et al. “OMA 2011: orthology inference among 1000 complete genomes”. In: *Nucleic Acids Res* 39 (2011), pp. D289–294 (see pp. 41, 42).

[97] A. L. Cuff et al. “Extending CATH: increasing coverage of the protein structure universe and linking structure with function”. In: *Nucleic Acids Res* 39 (2011), pp. D420–426 (see p. 41).

[98] M. A. Larkin et al. “Clustal W and Clustal X version 2.0”. In: *Bioinformatics* 23 (2007), pp. 2947–2948 (see p. 42).

[99] Joël Janin and Michael J E Sternberg. “Protein flexibility, not disorder, is intrinsic to molecular recognition”. In: *F1000 biology reports* 5 (2013), p. 2. DOI: 10.3410/B5-2 (see pp. 53–55, 119).

[100] Stephen J. Demarest et al. “Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators”. In: *Nature* 415.6871 (2002), pp. 549–553. DOI: 10.1038/415549a (see p. 53).

[101] Dimitra Keramisanou et al. “Disorder-order folding transitions underlie catalysis in the helicase motor of SecA”. In: *Nature Structural & Molecular Biology* 13.7 (2006), pp. 594–602. DOI: 10.1038/nsmb1108 (see p. 53).

[102] Alexey G. Murzin. “Metamorphic Proteins”. In: *Science* 320.5884 (June 27, 2008), pp. 1725–1726. DOI: 10.1126/science.1158868 (see p. 53).

[103] M. Madan Babu, Richard W. Kriwacki, and Rohit V. Pappu. “Versatility from Protein Disorder”. In: *Science* 337.6101 (Sept. 21, 2012), pp. 1460–1461. DOI: 10.1126/science.1228775 (see p. 53).

[104] Márton Münz, Jotun Hein, and Philip C. Biggin. “The Role of Flexibility and Conformational Selection in the Binding Promiscuity of PDZ Domains”. In: *PLoS Comput Biol* 8.11 (Nov. 1, 2012), e1002749. DOI: 10.1371/journal.pcbi.1002749 (see p. 53).

[105] Allan Chris M Ferreon et al. “Modulation of allostery by protein intrinsic disorder”. In: *Nature* 498.7454 (June 20, 2013), pp. 390–394. DOI: 10.1038/nature12294 (see p. 53).

[106] Tony Hunter. “The age of crosstalk: phosphorylation, ubiquitination, and beyond”. In: *Molecular cell* 28 (5 Dec. 14, 2007), pp. 730–738. DOI: 10.1016/j.molcel.2007.11.019 (see p. 64).

[107] Michael H Glickman and Aaron Ciechanover. “The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction”. In: *Physiological reviews* 82 (2 Apr. 2002), pp. 373–428. DOI: 10.1152/physrev.00027.2001 (see p. 64).

[108] V Chau et al. “A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein”. In: *Science (New York, N.Y.)* 243 (4898 Mar. 24, 1989), pp. 1576–1583 (see p. 64).

[109] Lijun Sun and Zhijian J Chen. “The novel functions of ubiquitination in signaling”. In: *Current opinion in cell biology* 16 (2 Apr. 2004), pp. 119–126. DOI: 10.1016/j.ceb.2004.02.005 (see p. 64).

[110] Zhijian J Chen and Lijun J Sun. “Nonproteolytic functions of ubiquitin in cell signaling”. In: *Molecular cell* 33 (3 Feb. 13, 2009), pp. 275–286. DOI: 10.1016/j.molcel.2009.01.014 (see p. 64).

[111] L Hicke. “Protein regulation by monoubiquitin”. In: *Nature reviews. Molecular cell biology* 2 (3 Mar. 2001), pp. 195–201. DOI: 10.1038/35056583 (see p. 64).

[112] Daniela Hoeller, Christina-Maria Hecker, and Ivan Dikic. “Ubiquitin and ubiquitin-like proteins in cancer pathogenesis”. In: *Nature reviews. Cancer* 6 (10 Oct. 2006), pp. 776–788. DOI: 10.1038/nrc1994 (see p. 64).

[113] Baris Bingol and Morgan Sheng. “Deconstruction for Reconstruction: The Role of Proteolysis in Neural Plasticity and Disease”. In: *Neuron* 69 (1 2011), pp. 22–32. DOI: 10.1016/j.neuron.2010.11.006 (see p. 64).

[114] Grzegorz Nalepa, Mark Rolfe, and J Wade Harper. “Drug discovery in the ubiquitin-proteasome system”. In: *Nature reviews. Drug discovery* 5 (7 July 2006), pp. 596–613. DOI: 10.1038/nrd2056 (see p. 64).

[115] Brian R Wong et al. “Drug discovery in the ubiquitin regulatory pathway”. In: *Drug Discovery Today* 8 (16 Aug. 15, 2003), pp. 746–754. DOI: 10.1016/S1359-6446(03)02780-6 (see p. 64).

[116] Junmin Peng et al. “A proteomics approach to understanding protein ubiquitination”. In: *Nature biotechnology* 21 (8 Aug. 2003), pp. 921–926. DOI: 10.1038/nbt849 (see p. 64).

[117] Namrata D Udeshi et al. “Refined preparation and use of anti-diglycine remnant (K-ϵ-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments”. In: *Molecular & cellular proteomics: MCP* 12 (3 Mar. 2013), pp. 825–831. DOI: 10.1074/mcp.O112.027094 (see p. 64).

[118] Fumiyo Ikeda and Ivan Dikic. “Atypical ubiquitin chains: new molecular signals. ’Protein Modifications: Beyond the Usual Suspects’ review series”. In: *EMBO reports* 9 (6 June 2008), pp. 536–542. DOI: 10.1038/embor.2008.93 (see p. 64).

[119] David M Lonard and Bert W O’malley. “Nuclear receptor coregulators: judges, juries, and executioners of cellular regulation”. In: *Molecular cell* 27 (5 Sept. 7, 2007), pp. 691–700. DOI: 10.1016/j.molcel.2007.08.012 (see p. 64).

[120] N Blom, S Gammeltoft, and S Brunak. “Sequence and structure-based prediction of eukaryotic protein phosphorylation sites”. In: *Journal of molecular biology* 294 (5 Dec. 17, 1999), pp. 1351–1362. DOI: 10.1006/jmbi.1999.3310 (see p. 64).

[121] Nikolaj Blom et al. “Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence”. In: *Proteomics* 4 (6 June 2004), pp. 1633–1649. DOI: 10.1002/pmic.200300771 (see pp. 64, 65).

[122] Chun-Wei Tung and Shinn-Ying Ho. “Computational identification of ubiquitylation sites from protein sequences”. In: *BMC bioinformatics* 9 (2008), p. 310. DOI: 10.1186/1471-2105-9-310 (see pp. 64, 68, 71).

[123] Predrag Radivojac et al. “Identification, analysis, and prediction of protein ubiquitination sites”. In: *Proteins* 78 (2 Feb. 1, 2010), pp. 365–380. DOI: 10.1002/prot.22555 (see pp. 64, 68, 70, 71).

[124] Zhen Chen et al. “Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs”. In: *PloS one* 6 (7 2011), e22930. DOI: 10.1371/journal.pone.0022930 (see pp. 65, 68, 70).

[125] Yudong Cai et al. “Prediction of lysine ubiquitination with mRMR feature selection and analysis”. In: *Amino acids* 42 (4 Apr. 2012), pp. 1387–1395. DOI: 10.1007/s00726-011-0835-0 (see pp. 65, 68, 70).

[126] Zhen Chen et al. “hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties”. In: *Biochimica et biophysica acta* 1834 (8 Aug. 2013), pp. 1461–1467. DOI: 10.1016/j.bbapap.2013.04.006 (see pp. 65, 68, 72).

[127] Xiang Chen et al. “Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites”. In: *Bioinformatics (Oxford, England)* 29 (13 July 1, 2013), pp. 1614–1622. DOI: 10.1093/bioinformatics/btt196 (see pp. 65, 68, 72).

[128] Daniel Schwartz. “Prediction of lysine post-translational modifications using bioinformatic tools”. In: *Essays in biochemistry* 52 (2012), pp. 165–177. DOI: 10.1042/bse0520165 (see p. 65).

[129] Sebastian A Wagner et al. “A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles”. In: *Molecular & cellular proteomics: MCP* 10 (10 Oct. 2011), p. M111.013284. DOI: 10.1074/mcp.M111.013284 (see pp. 65, 66, 70).

[130] Daniel Schwartz, Michael F Chou, and George M Church. “Predicting protein post-translational modifications using meta-analysis of proteome scale data sets”. In: *Molecular & cellular proteomics: MCP* 8 (2 Feb. 2009), pp. 365–379. DOI: 10.1074/mcp.M800332-MCP200 (see pp. 65, 68).

[131] Weizhong Li and Adam Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics* 22 (13 July 1, 2006), pp. 1658–1659. DOI: 10.1093/bioinformatics/btl158 (see pp. 65, 86).

[132] Peter V Hornbeck et al. “PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse”. In: *Nucleic acids research* 40 (Database issue 2012), pp. D261–270. DOI: 10.1093/nar/gkr1122 (see pp. 66, 73).

[133] Corinna Cortes and Vladimir Vapnik. “Support-Vector Networks”. In: *Mach. Learn.* 20 (3 Sept. 1995), pp. 273–297. DOI: 10.1023/a:1022627411411 (see p. 66).

[134] Pierre Baldi et al. “Exploiting the past and the future in protein secondary structure prediction”. In: *Bioinformatics* 15 (11 Nov. 1, 1999), pp. 937–946. DOI: 10.1093/bioinformatics/15.11.937 (see pp. 66, 67).

[135] B Rost and C Sander. “Prediction of protein secondary structure at better than 70% accuracy”. In: *Journal of molecular biology* 232 (2 July 20, 1993), pp. 584–599. DOI: 10.1006/jmbi.1993.1413 (see p. 67).

[136] Alex H M Ng et al. “System-wide analysis reveals intrinsically disordered proteins are prone to ubiquitylation after misfolding stress”. In: *Molecular & cellular proteomics: MCP* 12 (9 Sept. 2013), pp. 2456–2467. DOI: 10.1074/mcp.M112.023416 (see p. 70).

[137] Tzachi Hagai et al. “Intrinsic disorder in ubiquitination substrates”. In: *Journal of molecular biology* 412 (3 Sept. 23, 2011), pp. 319–324. DOI: 10.1016/j.jmb.2011.07.024 (see p. 70).

[138] Miguel A. Andrade, Carolina Perez-Iratxeta, and Chris P. Ponting. “Protein Repeats: Structures, Functions, and Evolution”. In: *Journal of Structural Biology* 134 (2–3 May 2001), pp. 117–131. DOI: 10.1006/jsbi.2001.4392 (see pp. 80, 95).

[139] Andrey V Kajava. “Tandem repeats in proteins: from sequence to structure”. In: *Journal of structural biology* 179 (3 Sept. 2012), pp. 279–288. DOI: 10.1016/j.jsb.2011.08.009 (see pp. 80, 96, 97, 100).

[140] Bostjan Kobe and Andrey V Kajava. “When protein folding is simplified to protein coiling: the continuum of solenoid protein structures”. In: *Trends in Biochemical Sciences* 25 (10 Oct. 1, 2000), pp. 509–515. DOI: 10.1016/S0968-0004(00)01667-4 (see pp. 80, 86, 96).

[141] J Buard and G Vergnaud. “Complex recombination events at the hypermutable minisatellite CEB1 (D2S90)”. In: *The EMBO journal* 13 (13 July 1, 1994), pp. 3203–3210 (see pp. 80, 95).

[142] Edward M. Marcotte et al. “A census of protein repeats”. In: *Journal of Molecular Biology* 293 (1 Oct. 15, 1999), pp. 151–160. DOI: 10.1006/jmbi.1999.3136 (see pp. 80, 96).

[143] Julien Jorda and Andrey V. Kajava. “Protein Homorepeats: Sequences, Structures, Evolution, and Functions”. In: *Advances in Protein Chemistry and Structural Biology*. Ed. by Alexander McPherson. Vol. Volume 79. Advances in Protein Chemistry and Structural Biology. Academic Press, 2010, pp. 59–88 (see pp. 80, 95).

[144] Andrey V. Kajava. “Review: Proteins with Repeated Sequence—Structural Prediction and Modeling”. In: *Journal of Structural Biology* 134 (2–3 May 2001), pp. 132–144. DOI: 10.1006/jsbi.2000.4328 (see p. 80).

[145] Joris de Wit et al. “Role of Leucine-Rich Repeat Proteins in the Development and Function of Neural Circuits”. In: *Annual Review of Cell and Developmental Biology* 27 (1 2011), pp. 697–729. DOI: 10.1146/annurev-cellbio-092910-154111 (see pp. 80, 95).

[146] Andrey V Kajava, John M Squire, and David A D Parry. “Beta-structures in fibrous proteins”. In: *Advances in protein chemistry* 73 (2006), pp. 1–15. DOI: 10.1016/S0065-3233(06)73001-7 (see p. 80).

[147] Ewan RG Main et al. “A recurring theme in protein engineering: the design, stability and folding of repeat proteins”. In: *Current Opinion in Structural Biology* 15 (4 Aug. 2005), pp. 464–471. DOI: 10.1016/j.sbi.2005.07.003 (see pp. 80, 95).

[148] Nikolas Stefan et al. “DARPins Recognizing the Tumor-Associated Antigen EpCAM Selected by Phage and Ribosome Display and Engineered for Multivalency”. In: *Journal of Molecular Biology* 413 (4 Nov. 4, 2011), pp. 826–843. DOI: 10.1016/j.jmb.2011.09.016 (see pp. 80, 95).

[149] Tommi Kajander et al. "A new folding paradigm for repeat proteins". In: *Journal of the American Chemical Society* 127 (29 July 27, 2005), pp. 10188–10190. DOI: 10.1021/ja0524494 (see pp. 80, 84).

[150] Andreas Heger and Liisa Holm. "Rapid automatic detection and alignment of repeats in protein sequences". In: *Proteins: Structure, Function, and Bioinformatics* 41 (2 2000), pp. 224–237. DOI: 10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z (see pp. 80, 87, 96).

[151] Radek Szklarczyk and Jaap Heringa. "Tracking repeats using significance and transitivity". In: *Bioinformatics* 20 (suppl 1 Aug. 4, 2004), pp. i311–i317. DOI: 10.1093/bioinformatics/bth911 (see pp. 80, 87, 96).

[152] A. Biegert and J. Söding. "De novo identification of highly diverged protein repeats by probabilistic consistency". In: *Bioinformatics* 24 (6 Mar. 15, 2008), pp. 807–814. DOI: 10.1093/bioinformatics/btn039 (see pp. 80, 81, 95, 96).

[153] F. M. G. Pearl et al. "The CATH database: an extended protein family resource for structural and functional genomics". In: *Nucleic Acids Research* 31 (1 2003), pp. 452–455. DOI: 10.1093/nar/gkg062 (see pp. 80, 86).

[154] Kevin B. Murray, Denise Gorse, and Janet M. Thornton. "Wavelet transforms for the characterization and detection of repeating motifs". In: *Journal of Molecular Biology* 316 (2 Feb. 15, 2002), pp. 341–363. DOI: 10.1006/jmbi.2001.5332 (see p. 80).

[155] Kevin B. Murray, William R. Taylor, and Janet M. Thornton. "Toward the detection and validation of repeats in protein structure". In: *Proteins: Structure, Function, and Bioinformatics* 57 (2 2004), pp. 365–380. DOI: 10.1002/prot.20202 (see pp. 80, 96).

[156] R. Sabarinathan, Raunak Basu, and K. Sekar. "ProSTRIP: A method to find similar structural repeats in three-dimensional protein structures". In: *Computational Biology and Chemistry* 34 (2 Apr. 2010), pp. 126–130. DOI: 10.1016/j.compbiolchem.2010.03.006 (see p. 80).

[157] Edward S.C. Shih and Ming-Jing Hwang. "Alternative alignments from comparison of protein structures". In: *Proteins: Structure, Function, and Bioinformatics* 56 (3 2004), pp. 519–527. DOI: 10.1002/prot.20124 (see p. 80).

[158] Anne-Laure Abraham, Eduardo P. C. Rocha, and Joël Pothier. "Swelfe: a detector of internal repeats in sequences and structures". In: *Bioinformatics* 24 (13 July 1, 2008), pp. 1536–1537. DOI: 10.1093/bioinformatics/btn234 (see pp. 80, 87, 96).

[159] Ewan RG Main, Sophie E Jackson, and Lynne Regan. "The folding and design of repeat proteins: reaching a consensus". In: *Current Opinion in Structural Biology* 13 (4 Aug. 2003), pp. 482–489. DOI: 10.1016/S0959-440X(03)00105-2 (see p. 84).

[160] Akira R. Kinjo and Ken Nishikawa. “Recoverable one-dimensional encoding of three-dimensional protein structures”. In: *Bioinformatics* 21 (10 May 15, 2005), pp. 2167–2170. DOI: 10.1093/bioinformatics/bti330 (see p. 84).

[161] Francesco Sirocco and Silvio C. E. Tosatto. “TESE: generating specific protein structure test set ensembles”. In: *Bioinformatics* 24 (22 Nov. 15, 2008), pp. 2632–2633. DOI: 10.1093/bioinformatics/btn488 (see p. 86).

[162] John C. Wootton. “Non-globular domains in protein sequences: Automated segmentation using complexity measures”. In: *Computers & Chemistry* 18 (3 Sept. 1994), pp. 269–285. DOI: 10.1016/0097-8485(94)85023-2 (see p. 95).

[163] M. Gribskov, A. D. McLachlan, and D. Eisenberg. “Profile analysis: detection of distantly related proteins”. In: *Proceedings of the National Academy of Sciences* 84 (13 July 1, 1987), pp. 4355–4358 (see p. 95).

[164] Elke Schaper et al. “Repeat or not repeat?–Statistical validation of tandem repeat prediction in genomic sequences”. In: *Nucleic Acids Research* 40 (20 Nov. 2012), pp. 10005–10017. DOI: 10.1093/nar/gks726 (see p. 95).

[165] Andrey V. Kajava and Alasdair C. Steven. “β-Rolls, β-Helices, and Other β-Solenoid Proteins”. In: *Advances in Protein Chemistry*. Ed. by John M. Squire Andrey Kajava and David A. D. Parry. Vol. Volume 73. Fibrous Proteins: Amyloids, Prions and Beta Proteins. Academic Press, 2006, pp. 55–96 (see p. 95).

[166] Yalda Javadi and Laura S Itzhaki. “Tandem-repeat proteins: regularity plus modularity equals design-ability”. In: *Current Opinion in Structural Biology* 23 (4 Aug. 2013), pp. 622–631. DOI: 10.1016/j.sbi.2013.06.011 (see p. 95).

[167] Alex Bateman, Alexey G. Murzin, and Sarah A. Teichmann. “Structure and distribution of pentapeptide repeats in bacteria”. In: *Protein Science* 7 (6 1998), pp. 1477–1480. DOI: 10.1002/pro.5560070625 (see p. 96).

[168] J. Bella et al. “The leucine-rich repeat structure”. In: *Cellular and Molecular Life Sciences* 65 (15 Aug. 1, 2008), pp. 2307–2333. DOI: 10.1007/s00018-008-8019-0 (see p. 96).

[169] Bostjan Kobe and Andrey V Kajava. “The leucine-rich repeat as a protein recognition motif”. In: *Current Opinion in Structural Biology* 11 (6 Dec. 1, 2001), pp. 725–732. DOI: 10.1016/S0959-440X(01)00266-4 (see p. 96).

[170] Rita Tewari et al. “Armadillo-repeat protein functions: questions for little creatures”. In: *Trends in Cell Biology* 20 (8 Aug. 2010), pp. 470–481. DOI: 10.1016/j.tcb.2010.05.003 (see p. 96).

[171] Andrey V. Kajava et al. “New HEAT-like repeat motifs in proteins regulating proteasome structure and function”. In: *Journal of Structural Biology* 146 (3 June 2004), pp. 425–430. DOI: 10.1016/j.jsb.2004.01.013 (see p. 96).

[172] Miguel A Andrade et al. “Comparison of ARM and HEAT protein repeats”. In: *Journal of Molecular Biology* 309 (1 May 25, 2001), pp. 1–18. DOI: 10.1006/jmbi.2001.4624 (see p. 96).

[173] Åsa K Björklund, Diana Ekman, and Arne Elofsson. “Expansion of Protein Domain Repeats”. In: *PLoS Comput Biol* 2 (8 Aug. 25, 2006), e114. DOI: 10.1371/journal.pcbi.0020114 (see p. 96).

[174] M Remmert et al. “Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin”. In: *Molecular biology and evolution* 27 (6 June 2010), pp. 1348–1358. DOI: 10.1093/molbev/msq017 (see p. 96).

[175] Zahra Jawad and Massimo Paoli. “Novel Sequences Propel Familiar Folds”. In: *Structure* 10 (4 Apr. 2002), pp. 447–454. DOI: 10.1016/S0969-2126(02)00750-5 (see p. 96).

[176] Indronil Chaudhuri, Johannes Söding, and Andrei N. Lupas. “Evolution of the β-propeller fold”. In: *Proteins: Structure, Function, and Bioinformatics* 71 (2 2008), pp. 795–803. DOI: 10.1002/prot.21764 (see p. 96).

[177] Helen M. Berman et al. “The future of the protein data bank”. In: *Biopolymers* 99 (3 2013), pp. 218–222. DOI: 10.1002/bip.22132 (see p. 96).

[178] Benoît H. Dessailly et al. “PSI-2: Structural Genomics to Cover Protein Domain Family Space”. In: *Structure* 17 (6 June 10, 2009), pp. 869–881. DOI: 10.1016/j.str.2009.03.015 (see p. 96).

[179] R. Gonzalo Parra et al. “Detecting Repetitions and Periodicities in Proteins by Tiling the Structural Space”. In: *The Journal of Physical Chemistry B* 117 (42 Oct. 24, 2013), pp. 12887–12897. DOI: 10.1021/jp402105j (see p. 96).

[180] Ian Walsh et al. “RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures”. In: *Bioinformatics (Oxford, England)* 28 (24 Dec. 15, 2012), pp. 3257–3264. DOI: 10.1093/bioinformatics/bts550 (see pp. 96, 97).

[181] Ian Sillitoe et al. “New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures”. In: *Nucleic Acids Research* 41 (D1 2013), pp. D490–D498. DOI: 10.1093/nar/gks1211 (see p. 96).

[182] Antonina Andreeva et al. “Data growth and its impact on the SCOP database: new developments”. In: *Nucleic acids research* 36 (Database issue 2008), pp. D419–425. DOI: 10.1093/nar/gkm993 (see p. 96).

[183] Julien Jorda, Thierry Baudrand, and Andrey V. Kajava. “PRDB: Protein Repeat DataBase”. In: *PROTEOMICS* 12 (9 2012), pp. 1333–1336. DOI: 10.1002/pmic.201100534 (see p. 96).

[184] Hong Luo et al. “ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins”. In: *Nucleic acids research* 40 (Database issue 2012), pp. D394–399. DOI: 10.1093/nar/gkr1019 (see p. 96).

[185] Ivica Letunic, Tobias Doerks, and Peer Bork. "SMART 7: recent updates to the protein domain annotation resource". In: *Nucleic acids research* 40 (Database issue 2012), pp. D302–305. DOI: 10.1093/nar/gkr931 (see p. 96).

[186] Jaina Mistry et al. "The challenge of increasing Pfam coverage of the human proteome". In: *Database: The Journal of Biological Databases and Curation* 2013 (Apr. 19, 2013). DOI: 10.1093/database/bat023 (see p. 96).

[187] Peter W Rose et al. "The RCSB Protein Data Bank: new resources for research and education". In: *Nucleic acids research* 41 (Database issue 2013), pp. D475–482. DOI: 10.1093/nar/gks1200 (see p. 97).

[188] John Gómez et al. "BioJS: an open source JavaScript framework for biological data visualization". In: *Bioinformatics (Oxford, England)* 29 (8 Apr. 15, 2013), pp. 1103–1104. DOI: 10.1093/bioinformatics/btt100 (see p. 98).

[189] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". In: *Nature* 406 (6794 July 27, 2000), pp. 378–382. DOI: 10.1038/35019019 (see p. 105).

[190] Albert-László Barabási and Réka Albert. "Emergence of Scaling in Random Networks". In: *Science* 286 (5439 Oct. 15, 1999), pp. 509–512. DOI: 10.1126/science.286.5439.509 (see p. 105).

[191] Paul Shannon et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13 (11 Nov. 1, 2003), pp. 2498–2504. DOI: 10.1101/gr.1239303 (see pp. 106, 110).

[192] Peter Csermely. "Creative elements: network-based predictions of active centres in proteins and cellular and social networks". In: *Trends in Biochemical Sciences* 33 (12 Dec. 2008), pp. 569–576. DOI: 10.1016/j.tibs.2008.09.006 (see p. 106).

[193] Cristina Marino Buslje et al. "Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification". In: *PLoS Comput Biol* 6 (11 Nov. 4, 2010), e1000978. DOI: 10.1371/journal.pcbi.1000978 (see p. 106).

[194] Antonio del Sol et al. "Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages". In: *Genome Biology* 8 (5 May 25, 2007), R92. DOI: 10.1186/gb-2007-8-5-r92 (see p. 106).

[195] Nikolay V. Dokholyan et al. "Topological determinants of protein folding". In: *Proceedings of the National Academy of Sciences* 99 (13 June 25, 2002), pp. 8637–8641. DOI: 10.1073/pnas.122076099 (see p. 106).

[196] Venkataramanan Soundararajan et al. "Atomic Interaction Networks in the Core of Protein Domains and Their Native Folds". In: *PLoS ONE* 5 (2 Feb. 23, 2010), e9391. DOI: 10.1371/journal.pone.0009391 (see p. 106).

[197] Gürol M. Süel et al. “Evolutionarily conserved networks of residues mediate allosteric communication in proteins”. In: *Nature Structural & Molecular Biology* 10 (1 2003), pp. 59–69. DOI: 10.1038/nsb881 (see p. 106).

[198] Liskin Swint-Kruse. “Using Networks To Identify Fine Structural Differences between Functionally Distinct Protein States†”. In: *Biochemistry* 43 (34 Aug. 1, 2004), pp. 10886–10895. DOI: 10.1021/bi049450k (see p. 106).

[199] Michele Vendruscolo et al. “Three key residues form a critical contact network in a protein folding transition state”. In: *Nature* 409 (6820 Feb. 1, 2001), pp. 641–645. DOI: 10.1038/35054591 (see p. 106).

[200] Cristina Marino Buslje et al. “Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information”. In: *Bioinformatics* 25 (9 May 1, 2009), pp. 1125–1131. DOI: 10.1093/bioinformatics/btp135 (see p. 106).

[201] Silvio C.E. Tosatto. “The Victor/FRST Function for Model Quality Estimation”. In: *Journal of Computational Biology* 12 (10 Dec. 2005), pp. 1316–1327. DOI: 10.1089/cmb.2005.12.1316 (see p. 106).

[202] Silvio CE Tosatto and Roberto Battistutta. “TAP score: torsion angle propensity normalization applied to local protein structure evaluation”. In: *BMC Bioinformatics* 8 (1 May 15, 2007), p. 155. DOI: 10.1186/1471-2105-8-155 (see p. 106).

[203] Yassen Assenov et al. “Computing topological parameters of biological networks”. In: *Bioinformatics* 24 (2 2008), pp. 282–284. DOI: 10.1093/bioinformatics/btm554 (see p. 106).

[204] Stefano Toppo et al. “Evolutionary and Structural Insights Into the Multifaceted Glutathione Peroxidase (Gpx) Superfamily”. In: *Antioxidants & Redox Signaling* 10 (9 Sept. 2008), pp. 1501–1514. DOI: 10.1089/ars.2008.2057 (see p. 107).

[205] Silvio C. E. Tosatto et al. “The Catalytic Site of Glutathione Peroxidases”. In: *Antioxidants & Redox Signaling* 10 (9 Sept. 2008), pp. 1515–1526. DOI: 10.1089/ars.2008.2055 (see p. 107).

[206] The UniProt Consortium. “Reorganizing the protein space at the Universal Protein Resource (UniProt)”. In: *Nucleic Acids Research* 40 (D1 Nov. 18, 2011), pp. D71–D75. DOI: 10.1093/nar/gkr981 (see pp. 109, 111).

[207] Paul D. Thomas et al. “On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report”. In: *PLoS Comput Biol* 8 (2 Feb. 16, 2012), e1002386. DOI: 10.1371/journal.pcbi.1002386 (see p. 109).

[208] Emily C. Dimmer et al. “The UniProt-GO Annotation database in 2011”. In: *Nucleic Acids Research* 40 (D1 2012), pp. D565–D570. DOI: 10.1093/nar/gkr1048 (see pp. 109, 112).

[209] Richard J. Roberts et al. "COMBREX: a project to accelerate the functional annotation of prokaryotic genomes". In: *Nucleic Acids Research* 39 (suppl 1 2011), pp. D11–D14. DOI: 10.1093/nar/gkq1168 (see pp. 109, 110).

[210] Predrag Radivojac et al. "A large-scale evaluation of computational protein function prediction". In: *Nature Methods* 10 (3 Mar. 2013), pp. 221–227. DOI: 10.1038/nmeth.2340 (see p. 109).

[211] Stefan Gotz et al. "High-throughput functional annotation and data mining with the Blast2GO suite". In: *Nucleic Acids Research* 36 (10 June 2008), pp. 3420–3435. DOI: 10.1093/nar/gkn176 (see p. 109).

[212] Damiano Piovesan et al. "BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences". In: *Nucleic Acids Research* 39 (suppl 2 July 1, 2011), W197–W202. DOI: 10.1093/nar/gkr292 (see p. 109).

[213] Alan E. Barber and Patricia C. Babbitt. "Pythoscape: a framework for generation of large protein similarity networks". In: *Bioinformatics* 28 (21 Nov. 1, 2012), pp. 2845–2846. DOI: 10.1093/bioinformatics/bts532 (see p. 110).

[214] Holly J. Atkinson et al. "Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies". In: *PLoS ONE* 4 (2 Feb. 3, 2009), e4345. DOI: 10.1371/journal.pone.0004345 (see p. 110).

[215] Ioannis Valavanis, George Spyrou, and Konstantina Nikita. "A similarity network approach for the analysis and comparison of protein sequence / structure sets". In: *Journal of Biomedical Informatics* 43 (2 Apr. 2010), pp. 257–267. DOI: 10.1016/j.jbi.2010.01.005 (see p. 110).

[216] Shoshana D Brown and Patricia C Babbitt. "Inference of functional properties from large-scale analysis of enzyme superfamilies". In: *The Journal of biological chemistry* 287 (1 2012), pp. 35–42. DOI: 10.1074/jbc.R111.283408 (see p. 110).

[217] Alexandra M. Schnoes et al. "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies". In: *PLoS Comput Biol* 5 (12 Dec. 11, 2009), e1000605. DOI: 10.1371/journal.pcbi.1000605 (see p. 110).

[218] Nan Song et al. "Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins". In: *PLoS Comput Biol* 4 (5 May 16, 2008), e1000063. DOI: 10.1371/journal.pcbi.1000063 (see p. 110).

[219] Arun S. Konagurthu et al. "MUSTANG: A multiple structural alignment algorithm". In: *Proteins: Structure, Function, and Bioinformatics* 64 (3 2006), pp. 559–574. DOI: 10.1002/prot.20921 (see pp. 110, 111).

[220] Yang Zhang and Jeffrey Skolnick. "TM-align: a protein structure alignment algorithm based on the TM-score". In: *Nucleic Acids Research* 33 (7 2005), pp. 2302–2309. DOI: 10.1093/nar/gki524 (see p. 110).

[221] Jinrui Xu and Yang Zhang. “How significant is a protein structure similarity with TM-score = 0.5?” In: *Bioinformatics* 26 (7 Apr. 1, 2010), pp. 889–895. DOI: 10.1093/bioinformatics/btq066 (see p. 111).

[222] Burkhard Rost. “Review: Protein Secondary Structure Prediction Continues to Rise”. In: *Journal of Structural Biology* 134 (2–3 May 2001), pp. 204–218. DOI: 10.1006/jsbi.2001.4336 (see p. 111).

[223] Leonard Apeltsin et al. “Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution”. In: *Bioinformatics* 27 (3 Feb. 1, 2011), pp. 326–333. DOI: 10.1093/bioinformatics/btq655 (see p. 112).

[224] R. C. Prim. “Shortest Connection Networks And Some Generalizations”. In: *Bell System Technical Journal* 36 (6 1957), pp. 1389–1401. DOI: 10.1002/j.1538-7305.1957.tb01515.x (see p. 112).

[225] Holly J. Atkinson and Patricia C. Babbitt. “An Atlas of the Thioredoxin Fold Class Reveals the Complexity of Function-Enabling Adaptations”. In: *PLoS Comput Biol* 5 (10 Oct. 23, 2009), e1000541. DOI: 10.1371/journal.pcbi.1000541 (see pp. 113, 114).

[226] Alberto J. M. Martin et al. “RING: networking interacting residues, evolutionary information and energetics in protein structures”. In: *Bioinformatics* 27 (14 July 15, 2011), pp. 2003–2005. DOI: 10.1093/bioinformatics/btr191 (see p. 113).

[227] Tsuto Morinaga et al. “Total Nucleotide Sequences of the Infectious Cloned DNAs of Bean Golden Mosaic Virus”. In: *Microbiology and Immunology* 31 (2 1987), pp. 147–154. DOI: 10.1111/j.1348-0421.1987.tb03078.x (see p. 113).

A. Supplementary material

URLs for supplementary material referenced throughout this work are listed in this section.

A.1 CSpritz

http://nar.oxfordjournals.org/content/39/suppl_2/W190/suppl/DC1

A.2 ESpritz

http://bioinformatics.oxfordjournals.org/content/28/4/503/suppl/DC1

A.3 RUBI

http://link.springer.com/content/esm/art:10.1007/s00726-013-1645-3/file/MediaObjects/726_2013_1645_MOESM1_ESM.doc

A.4 RAPHAEL

http://bioinformatics.oxfordjournals.org/content/28/24/3257/suppl/DC1

\o/