Computational Investigations of Backbone Dynamics in Intrinsically Disordered Proteins

Tomasz Kościółek

Bioinformatics Group
Department of Computer Science
University College London

A thesis submitted to University College London for the degree of Doctor of Philosophy

October 2015

I, Tomasz Kościółek, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

ABSTRACT

Intrinsically disordered proteins (IDPs), due to their dynamic nature, play important roles in molecular recognition, signalling, regulation, or binding of nucleic acids.

IDPs have been extensively studied computationally in terms of binary disorder/order classification. This approach has proven to be fruitful and enabled researchers to estimate the amount of disorder in prokaryotic and eukaryotic genomes. Other computational methods – molecular dynamics, or other simulation techniques, require a starting structure. However, there are no approaches permitting insight into the behaviour of disordered ensembles from sequence alone. Such a method would facilitate the study of proteins of unknown structures, help to obtain a better classification of the disordered regions, and the design disorder-to-order transitions.

In this work, I develop FRAGFOLD-IDP, a method to address this issue. Using a fragment-based structure prediction approach – FRAGFOLD, I generate the ensembles of IDPs and show that the features extracted from them correspond well with the backbone dynamics of NMR ensembles deposited in the PDB.

FRAGFOLD-IDP predictions significantly improve over a naïve approach and help to get a better insight into the dynamics of the disordered ensembles. The results also show it is not necessary to predict the correct fold of the protein to reliably assign per-residue fluctuations to the sequence in question. This suggests that disorder is a local property and it does not depend on the protein fold.

Next, I validate FRAGFOLD-IDP on the disorder classification task and show that the method performs comparably to machine learning-based approaches designed specifically for this task.

I also found that FRAGFOLD-IDP produces results on par with DynaMine, a machine learning approach to predict the NMR order parameters and that the results of both methods are not correlated. Thus, I constructed a consensus neural network predictor, which takes the results of FRAGFOLD-IDP, DynaMine and physicochemical features to predict per-residue fluctuations, improving upon both input methods.

TABLE OF CONTENTS

1 INTRODUCTION 12

1.1. Ordered and intrinsically disordered proteins 12

1.1.1. Historic perspective 12

1.1.2. Definition of intrinsic disorder in proteins 15

1.1.3. Abundance of IDPs 17

1.2. IDPs functions and associations with diseases 19

1.2.1. IDP functions 19

1.2.2. IDPs and diseases 23

1.3. Experimental characterization of protein disorder 25

1.3.1. Biophysical and biochemical methods 25

1.3.2. Spectroscopic techniques 26

1.3.3. X-ray crystallography 26

1.3.4. NMR characterization of protein structures and disorder 27

1.3.4.1. What does an NMR spectrum tell us? 27

1.3.4.2. Generating models – restrained REMD 29

1.3.4.3. Generating models – ensemble selection 30

1.3.4.4. Assessment of NMR models 30

1.3.4.5. NMR of IDPs 31

1.3.4.6. Disorder classification from NMR structures 33

1.3.4.7. Future 34

1.4. Databases of structural information on intrinsic disorder 35

1.4.1. DisProt 35

1.4.2. mobiDB 37

1.4.3. D^{2}P^2 38

1.4.4. pE-DB (PED) 39

1.4.5. IDEAL 40

1.4.6. Summary 41

1.5. Computational predictions of intrinsic disorder 44

1.5.1. Sequence-based disorder prediction methods 44

1.5.1.1. Disorder as a classification problem 44

1.5.1.2. Approaches to disorder predictions 45

1.5.1.3. Physics-based predictors 45

1.5.1.4. Machine learning-based predictors 47

1.5.1.5. Meta predictors ... 49
1.5.2. Computational simulations of intrinsic disorder ... 50
1.5.2.1. MD-based disorder simulations ... 50
1.5.2.2. Coarse-grained disorder simulations ... 51
1.5.2.3. CABSflex ... 54
1.5.2.4. DynaMine ... 55
1.6. Aims and outline of thesis ... 57

2 FRAGFOLD-IDP *DE NOVO* PREDICTIONS OF INTRINSICALLY DISORDERED PROTEIN ENSEMBLES ... 59
2.1. Background ... 59
2.1.1. Purpose and challenges ... 59
2.1.2. Other approaches used to predict the properties of IDP ensembles ... 60
2.1.2.1. DynaMine – summary of the results ... 60
2.1.2.2. Study of IDPs using Rosetta ... 62
2.1.3. FRAGFOLD folding engine ... 64
2.2. Dataset ... 67
2.2.1. Construction ... 67
2.2.2. Characterization ... 68
2.2.3. Benchmark set ... 69
2.3. Methods ... 71
2.3.1. FRAGFOLD-IDP workflow ... 71
2.3.2. Ensemble generation ... 72
2.3.3. Ensemble extraction ... 73
2.3.3.1. Structural clustering ... 73
2.3.3.2. Hierarchical clustering ... 73
2.3.3.3. SPICKER ... 74
2.3.3.4. MaxCluster ... 74
2.3.3.5. PFClust ... 75
2.3.4. Ensemble comparison ... 76
2.3.4.1. Structural superposition ... 76
2.3.4.2. Per-residue RMSD profile ... 81
2.3.4.3. Disorder match metric – Spearman’s rank correlation ... 81
2.3.4.4. Structural comparison – TM-score ... 82
2.3.5. Borderline results extraction ... 84
2.3.5.1. Baseline method – naïve disorder assignment ... 84
2.3.5.2. Random ensemble generation ... 85
2.3.6. Relationship between NMR order parameter (S^2) and per-residue RMSD ... 85

2.4. FRAGFOLD-IDP optimisation 88
2.4.1. FRAGFOLD folding parameters 88
2.4.2. Ensemble extraction method 92
2.4.2.1. Clustering algorithm 93
2.4.2.2. Cluster selection criteria 95
2.4.3. Structural superposition method 98
2.4.4. Disorder match metric 101
2.4.5. Optimal ensemble size 101
2.4.6. Final FRAGFOLD-IDP workflow 102
2.5. Full dataset results 104
2.5.1. Interpretation of R_S values 105
2.5.2. Overall performance 108
2.5.2.1. Outliers 111
2.5.3. FRAGFOLD-IDP and naïve predictions 114
2.5.4. Comparison of final FRAGFOLD-IDP ensembles with random and best clusters 117
2.5.5. Impact of protein class on FRAGFOLD-IDP performance 121
2.5.6. Impact of disorder content on FRAGFOLD-IDP performance 124
2.5.7. Structure and backbone dynamics predictions 126
2.5.8. Comparison with other approaches to predict IDP ensembles 132
2.5.8.1. B-factor predictions 132
2.5.8.2. Disorder/order predictors 134
2.5.8.3. Overall comparison 134
2.5.8.4. FRAGFOLD-IFP and DynaMine 135
2.6. Summary 138

3 DISORDER/ORDER CLASSIFICATION WITH FRAGFOLD-IDP 141

3.1. Background 141
3.1.1. Problem formulation 143
3.2. Methods 144
3.2.1. Dataset 144
3.2.2. NMR ensemble disorder boundaries 144
3.2.3. FRAGFOLD-IDP as a disorder/order classifier 145
3.2.4. Sequence-based disorder predictors 146
3.3. Results 149
3.3.1. Impact of RMSD thresholding on disorder/order classification 149
3.3.2. Disorder/order classification on FRAGFOLD-IDP dataset 149
3.3.3. Disorder/order classification of long disordered regions 154

3.3.4. Relationship between disorder/order classification and backbone dynamics predictions 157

3.4. Summary 158

4 CONSENSUS MACHINE LEARNING-BASED PREDICTIONS OF PROTEIN BACKBONE DYNAMICS 160

4.1. Background 160

4.1.1. Summary of FRAGFOLD-IDP and DynaMine results 160

4.1.2. Meta predictors in other bioinformatics approaches 160

4.1.3. Consensus backbone dynamics predictor 162

4.2. Methods 163

4.2.1. Consensus predictor input features 164

4.2.2. Consensus predictor architecture 166

4.2.3. Training procedure 167

4.3. Results 168

4.3.1. Network training 168

4.3.2. Consensus predictor results 170

4.4. Summary 175

5 DISCUSSION AND PERSPECTIVES 177

5.1. Protein backbone dynamics predictions 177

5.2. FRAGFOLD-IDP as a disorder/order classifier 178

5.3. Efficiency of FRAGFOLD-IDP method in predicting disordered ensembles 179

5.4. Biological significance of the current study 181

5.5. Disorder design 182

5.6. Future developments of FRAGFOLD-IDP 185

5.7. Limitations of studying disorder 187

5.8. Concluding remarks 188

APPENDICES 189

A. List of abbreviations 189

B. Software information 190

REFERENCES 192

LIST OF FIGURES

Figure 1. An example of a protein with an intrinsically disordered region 13

Figure 2. An overview of experimental techniques used to study intrinsically disordered proteins. 25

Figure 3. FRAGFOLD flow diagram 65

Figure 4. Length distribution of proteins in the dataset. 68

Figure 5. Disorder content distribution in the dataset. 69

Figure 6. Schematic workflow used in FRAGFOLD-IDP method. 72

Figure 7. Comparison of least-squares (LS), maximum-likelihood (ML) superposition and principal component analysis (PC1) 77

Figure 8. A toy example of sliding window superposition procedure 79

Figure 9. Comparison of sliding window size versus the fit of predicted ensembles to NMR ensembles 81

Figure 10. Relation between order parameter S^2 and backbone RMSD 86

Figure 11. Differences in FRAGFOLD performance using different sets of potentials........... 91

Figure 12. Box-and-whiskers plot comparing the distribution of R_S values between different sets of FRAGFOLD parameters 92

Figure 13. Comparison of standard clustering algorithms.. 93

Figure 14. Comparison of RMSDclust and PFClust performance on the benchmark set. 95

Figure 15. Comparison of cluster selection criteria using PFClust on the whole dataset...... 97

Figure 16. Disorder profile of stefin A (1DVD) 99

Figure 17. Visualisation of an NMR ensemble and FRAGFOLD predictions of 1DVD (stefin A) 100

Figure 18. Optimal ensemble size determination... 102

Figure 19. Final FRAGFOLD-IDP workflow.. 102

Figure 20. Disorder profile of 2KKJ (nuclear coactivator binding domain of CBP) 107

Figure 21. Distribution of FRAGFOLD-IDP results on the 200 protein dataset. 108

Figure 22. Disorder profile of a poor FRAGFOLD-IDP prediction (1SIY; R_S = 0.21).............. 109

Figure 23. Disorder profile of a medium quality FRAGFOLD-IDP prediction (1P94; R_S = 0.54). 110

Figure 24. Disorder profile of an excellent FRAGFOLD-IDP prediction (2KJV; R_S = 0.82)..... 111

Figure 25. NMR PDB ensemble of 1G6M (Cobrotoxin II)... 113

Figure 26. NMR ensemble of 1K0T (photosystem I Subunit PsaC) 114

Figure 27. Comparison of R_S values between a naive approach and FRAGFOLD-IDP.......... 115

Figure 28. Disorder profile of 1NSH (Rabbit protein S100-A11). 117

Figure 29. Comparison of R_S from the final cluster extracted using PFClust with median R_S extracted from 1,000 randomly generated ensembles. 118

Figure 30. Disorder profile of 1NY4 (30S ribosomal protein S28e from Pyrococcus horikoshii). 119

Figure 31. Comparison of the best cluster from PFClust in terms of R_S with FRAGFOLD-IDP. 120

Figure 32. Comparison of the top cluster from 1,000 randomly generated ensembles with the best cluster in terms of R_S extracted using PFClust. 121

Figure 33. Comparison of full dataset results by CATH class. 122

Figure 34. NMR PDB ensemble of 1TXB (long neurotoxin 2) 123

Figure 35. Relationship between the disorder content in NMR ensembles and the quality of FRAGFOLD-IDP predictions. 125

Figure 36. Impact of structure prediction quality (TM-score) on backbone dynamics predictions (R_S). 127

Figure 37. Disorder profile of 1APS (Horse acylphosphatase-2). 128

Figure 38. Backbone dynamics predictions quality (R_S) from TM-score binning. 130

Figure 39. Median R_S values between FRAGFOLD-IDP and other computational techniques. 135

Figure 40. Per target comparison of FRAGFOLD-IDP and DynaMine results. 136

Figure 41. Disorder/order classification ROC curve. 150

Figure 42. Disorder/order classification ROC curve up to 10% false positive rate. 153

Figure 43. ROC curve of disorder/order classification of long disordered regions (≥ 20 residues). 155

Figure 44. ROC curve of disorder/order classification of long disordered regions up to 10% FPR. 156

Figure 45. Sample neural network architecture 163

Figure 46. A schematic representation of the consensus backbone dynamics predictor. .. 166

Figure 47. Optimisation of the consensus predictor 169

Figure 48. Optimisation of the number of hidden units in the consensus predictor. 170

Figure 49. Comparison of FRAGFOLD-IDP, DynaMine and consensus predictor median R_S values. 171

Figure 50. Example of an excellent consensus predictor result – 1P94 173

LIST OF TABLES

Table 1. Summary of the most common NMR parameters. 28

Table 2. Summary of IDP databases. 35

Table 3. Comparison of mobiDB and D2P2 databases contents. 42

Table 4. Some major computational simulations of IDPs. 53

Table 5. Summary of benchmark set proteins. 70

Table 6. Naïve predictions optimisation. 85

Table 7. Results from the initial studies using FRAGFOLD-IDP on the benchmark set. 90

Table 8. Top clustering method performance on the benchmark set. 94

Table 9. Comparison of mean RS values for different cluster selection criteria. 97

Table 10. Outliers in FRAGFOLD-IDP predictions. 112

Table 11. Correlations between FRAGFOLD-IDP RS value and disorder content by protein CATH class. 126

Table 12. Correlations between TM-score and RS for different cluster selection criteria. 129

Table 13. Order-disorder boundary of NMR ensembles. 145

Table 14. Performance of the disorder/order classification methods and FRAGFOLD-IDP. 151

Table 15. Summary of consensus predictor per-residue features. 165

Table 16. Good and excellent predictions produced by the algorithms. 172

Chapter 1.
INTRODUCTION

1.1. Ordered and intrinsically disordered proteins

1.1.1. *Historic perspective*

Current understanding of protein structure and function remains greatly impacted by the 19th century image created by Emil Fischer's lock-and-key paradigm for enzyme catalysis (Fischer, 1894; Uversky and Dunker, 2013). In late 1950s the seminal work on the structure of myoglobin (Kendrew et al., 1958), followed in the early 1960s with Anfinsen's works on protein sequence-structure relationships (Anfinsen et al., 1961) strengthened the common understanding of proteins as well-ordered biological machines. Although they may additionally possess some degree of conformational flexibility (Koshland's induced fit: Koshland, 1959; or similar mechanisms), they were otherwise held be well defined in terms of their shape, fold and structural details.

It did not take long for the idea of fixed-shape proteins to dissolve. First indirect evidence for a protein (serum albumin) to be functional in an ensemble of conformations was as early as 1950 (Dunker et al., 2001; Karush, 1950). But it was not until 1971, when the first protein (extracellular nuclease from *Staphylococcus aureus*) X-ray structure showing missing electron density region was reported (Arnone et al., 1971; Uversky and Dunker, 2013). Since then, the existence of proteins that have high flexibility or disordered regions, yet remain functional has been acknowledged, but for numerous reasons they were not in the spotlight. The reasons were largely of experimental nature, as these proteins, now commonly referred to as Intrinsically Disordered Proteins (IDPs), do not crystallize easily and biochemical procedures for the preparation of samples were heavily biased towards compact, folded proteins (Dyson and Wright, 2005).

It took until the turn of the millennium, for IDPs to get recognition as an important part of the protein universe, rather than being viewed as incidental peculiarities

(Dunker et al., 2001; Uversky et al., 2000; Wright and Dyson, 1999). Currently, IDPs form the fourth ‘tribe’ within the protein kingdom, alongside globular, fibrillar and transmembrane proteins (Figure 1; Uversky and Dunker, 2010). They are estimated to contribute to a significant fraction of both the eukaryotic and prokaryotic proteomes, having been an important development of evolution (more on the abundance of IDPs in section 1.1.3; Schlessinger et al., 2011; Ward et al., 2004b). Subsequently, it was experimentally confirmed that the behaviour of IDPs observed in isolation corresponds to that in the cellular environment and intrinsically disordered proteins are not merely experimental artefacts (Bodart et al., 2008; Theillet et al., 2014). These realizations led to the rapid increase in the intensity of studies of IDPs, not only because such proteins (or regions) are widely present, but because they yield important biological functions. Disorder is predominant in regulation, signal transduction, one-to-many and low affinity-high specificity binding, kinase activity, or is associated with several disease states, including Parkinson’s disease or polyglutamate-repeat disorders, i.e. Huntington’s disease (described in detail in section 1.2; Dyson and Wright, 2005; Ward et al., 2004b).

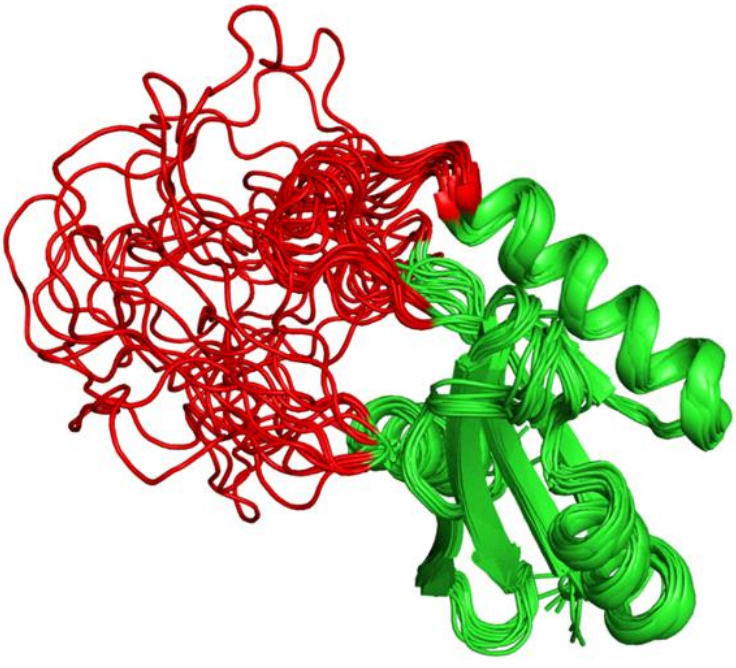


Figure 1. An example of a protein with an intrinsically disordered region. The ordered region of the protein is highlighted in green and the disordered region of the protein is highlighted in red. The figure shows an experimental structure of putative pre-16S rRNA nuclease (PDB id: 1OVQ) from *E. Coli* solved using NMR spectroscopy.

Much is discovered about protein disorder every year and the importance of this phenomenon cannot be underestimated. IDP-related research is now at its exponential growth phase with more than 530 IDP-related papers published in 2014 alone (based on Scopus keyword search). Despite the great interest in IDPs, current knowledge still seems to be far from complete in understanding the structural properties and biological role of these proteins.

Equally importantly, the physical bases and biophysical properties of intrinsic disorder are not yet fully understood. This is of utmost importance for several reasons: to be able to adequately describe this ‘tribe’ of proteins and to be able to relate the physical properties of IDPs to their function and, therefore, to gain better understanding of protein folding in general.

1.1.2. *Definition of intrinsic disorder in proteins*

Unfortunately, many researchers publishing on protein disorder give only vague descriptions of protein disorder which lack scientific stringency. The most common definition used is that disordered proteins are proteins lacking stable tertiary structure under physiological conditions. Other, more specific definitions include:

> "Intrinsically disordered proteins fold after binding, structured proteins do that before" (Uversky and Dunker, 2013)

> "Flexibility in IDPs refers to massive changes in backbone and side chain dihedral angles leading to changes in shape, whilst in folded proteins flexibility corresponds to oscillating motions around equilibrium positions, so that the overall shape is maintained" (*Op. cit.*)

> "We consider as disorder whatever is predicted as such." (Schaefer et al., 2010)

> "We refer to disordered regions as those regions in proteins that, when in isolation (i.e., not bound to other molecules), do not fold into a well-defined 3D structure but rather sample a large portion of their available conformational space." (Schlessinger et al., 2011)

> "These proteins lack a stable equilibrium conformation but exist as dynamic ensembles within which atom positions exhibit extreme temporal fluctuations without specific equilibrium values." (Orosz and Ovadi, 2011)

> "(…) IDPs possess no well-defined 3-D structure but rather adopt an ensemble of conformations in solution, yet they are functional." (Habchi et al., 2014)

It is indeed difficult to provide a single comprehensive definition of protein disorder, as the behaviour of disordered regions depends on the environment or partner molecules, but perhaps the definition rests mostly on the functional role of such proteins or regions. Moreover, serious experimental difficulties exist to observe the dynamics of IDPs, hence verification of structural hypotheses is hindered. Nevertheless, several important physical and biological features arise from the above mentioned definitions and other literature:

(1) The free energy landscape of IDPs has a relatively “flat” bottom to its spectrum in comparison to ordered proteins. The intermediate states occupy and interchange local energy minima.

(2) Intrinsically disordered regions (IDRs) have a fluctuating backbone.

(3) In IDRs aperiodic backbone motions should be observed, in contrast to ordered but flexible regions where oscillating (periodic) motions occur.

(4) Flexibility is intrinsic to a protein as a biopolymer, whilst disorder is intrinsic to protein’s or region’s function. Flexibility describes changes involving few degrees of freedom, while disorder is a state where the changes involve many degrees of freedom or there are no constraints on the degrees of freedom whatsoever (Janin and Sternberg, 2013).

(5) IDRs are functional, but their function is not limited to folding-upon-binding. There are also regions where disorder serves as an entropic chain stabilizing the protein thermodynamically, or as flexible linker between protein domains, or with only partial folding as observed in fuzzy complexes (Fuxreiter, 2012).

It is also important to stress what intrinsically disordered proteins are NOT. At present, it is a consensus to use the term intrinsically disordered proteins (or IDPs), but until recently this class of proteins was also referred to as unstructured, or unfolded proteins, among many other names (Dunker et al., 2014). Both those terms are misleading, because unstructured suggests a permanent lack of structure (what is not the case in IDPs) and also lack of transient structures. For example, some IDPs adapt molten globule conformations (Dyson and Wright, 2005), hence they are not unstructured, but disordered. Also the other term, unfolded, is misleading, as it suggests that it is a protein not in its native state, or that there is a single folded state of that protein. This might be true for some IDPs, but as a general term it should be avoided, since not all IDPs fold-upon-binding (Fuxreiter, 2012). Finally, IDPs are clearly not misfolded proteins. Although they are more susceptible to proteolytic degradation, hence their half-life is shorter and are preferentially located is some cellular compartments, the ensemble of conformations which IDPs explore is their native state (Janin and Sternberg, 2013; Ward et al., 2004b).

1.1.3. *Abundance of IDPs*

Determining the prevalence of IDPs in different organisms and domains of life is a crucial piece of information. Because of the limited structural information and historic difficulties with characterizing IDPs experimentally, the estimate of abundance of IDPs is based on computational techniques. The methods that are used to predict disorder in proteins are described in detail in section 1.5. Here, let us concentrate on the conclusions from those computational studies on the abundance of intrinsically disordered proteins.

One of the first and still widely cited sources is based on a computational method (DISOPRED2) which was trained to yield low false positive rate (Ward et al., 2004b). It is a conservative approach, which is unlikely to overestimate the amount of disorder. Based on a DISOPRED2 survey of multiple eukaryotic and prokaryotic genomes, Ward et al. found that there is generally more disorder in eukaryotes, than in prokaryotes. There are around 30% of proteins with long disordered regions (more than 30 consecutive residues) in eukaryotes, and between 1% and 7% in prokaryotes. Some more recent studies confirm these findings and show that consistently IDPs in eukaryotes have more disorder content and are enriched in long disordered segments when compared to bacteria and archaea (Peng et al., 2014b).

Looking at the distribution of disordered proteins between different organisms, it is a general consensus that in higher organisms the abundance of IDPs (especially with long disordered regions) increases (Habchi et al., 2014; Pentony et al., 2010). For example, in humans disordered proteins with long disordered regions are estimated at 44% of the proteome (Oates et al., 2013). This observation is tightly linked to function and evolution (Dunker and Obradovic, 2001; Schlessinger et al., 2011).

An interesting addition to this, is the abundance of intrinsically disordered proteins encoded in viruses. In this case, a widespread abundance was observed, ranging from around 7% to more than 77% (Habchi et al., 2014; Xue et al., 2012b). In most cases, viral proteins are second to eukaryotic proteins in terms of their disorder content, or length of the disordered segments (Peng et al., 2014b).

1.2. IDPs functions and associations with diseases

1.2.1. *IDP functions*

Disordered proteins can be associated with the evolutionary functional achievements of eukaryotic cells. The functional hallmark of disorder is its ability to mediate specific interaction with multiple binding partners (Babu et al., 2012; Dyson and Wright, 2005). As a result, IDPs can perform molecular recognition associated with signalling and regulation, as well as binding (Babu et al., 2011; Cozzetto and Jones, 2013; Uversky and Dunker, 2013; Ward et al., 2004b).

Because of the dynamic nature of the disordered state, IDPs can provide a larger interaction surface than ordered proteins of similar size. IDPs are thus able to perform low affinity and high specificity binding (Dunker et al., 2002). The fact that in eukaryotes disorder is more prevalent may also be associated with unique eukaryotic cellular functions, such as organization and biogenesis of cytoskeleton, or functions associated with the development of the nucleus (e.g. disorder in histone proteins; Peng et al., 2012; Ward et al., 2004b). The lower content of disordered proteins in prokaryotes may be associated with a relatively short half-life of disordered proteins, which in prokaryotic organisms might have a much greater cost, and the lack of cellular compartments which would protect the disordered proteins from degradation (Ward et al., 2004b). Because of the higher susceptibility of IDPs to proteolysis, in eukaryotes most disordered proteins are located in cellular compartments, i.e. cellular cortex and the nucleus. This naturally associates disordered proteins with binding of DNA. Because of the dynamic features of disorder and its versatility, IDPs can facilitate transposition, transcription, packaging, replication and repair. Therefore, another feature of disorder from which the eukaryotes benefit is its contribution to cell differentiation.

As disorder is represented by many conformational states, IDPs can perform one-to-many binding (e.g. calmodulin and p53) (Oldfield et al., 2008; Romero et al., 1998; Wright and Dyson, 1999). p53 is known to interact with multiple partners, binding of which is performed by non-overlapping sets of amino acids that also do not share common secondary structure features in the bound state (Uversky and Dunker, 2013).

Although IDPs play many highly specific roles in organisms, their functions are even broader. Often they are utilized as flexible linkers and entropic chains (Uversky and Dunker, 2013; van der Lee et al., 2014). Flexible linkers contribute to overall protein plasticity allowing for interactions that would be otherwise impossible. A good example are zinc fingers, where flexible linkers enable the protein to wrap around its target DNA (Dyson and Wright, 2005).

From the protein interaction network perspective, IDPs are often found to be hubs of protein networks (Bellay et al., 2011; Cumberworth et al., 2013; Ward et al., 2004b). Their ability to perform one-to-many binding and adopt several folded states (as in the case of p53), as well as to form transient interactions makes them ideal for the task. It has been estimated that two thirds of all signalling proteins have long disordered regions (Iakoucheva et al., 2002; Latysheva et al., 2015).

This broad functional arsenal of IDPs and their efficacy requires that their expression and functional roles are tightly regulated. Disordered proteins are more prone to post-translational modifications (i.e. ubiquitination and phosphorylation), which enables cells to control the level of regulatory proteins (Edwards et al., 2009; Xue et al., 2012a). IDPs are also controlled on the transcript level by mRNA decay and tissue-specific alternative splicing (Babu et al., 2012; Buljan et al., 2012; Edwards et al., 2009). Fine-tuning of IDP availability usually keeps the IDP levels low and for short periods of time and imbalance is often a cause of IDP-related diseases (overviewed in section 1.2.2; Babu et al., 2011; Gsponer et al., 2008; Vavouri et al., 2009).

On the other hand, IDPs are rarely associated with catalytic functions. The requirement for a well-defined binding site precludes IDPs from performing this function (Babu et al., 2011). This does not include kinases, where disordered domains can be found. Kinases are involved in regulatory processes and need to bind multiple substrates (Ward et al., 2004b).

Another feature of IDPs, tightly associated with their function, is the presence of linear molecular recognition features (MoRFs) (Cozzetto and Jones, 2013; Cumberworth et al., 2013). These functional elements are also called preformed structural elements (PSEs), molecular recognition elements (MoREs), or pre-structured motifs (PreSMos) (van der Lee et al., 2014). MoRFs are short segments of protein disorder, usually no longer than 70 amino acids that bind to specific partners undergoing disorder-to-order transitions (Dunker et al., 2008; van der Lee et al., 2014). They have a lower mean net charge and higher hydrophobicity than other IDRs (Cozzetto and Jones, 2013; Vacic et al., 2007). Functionally, MoRFs are responsible for molecular recognition and binding. MoRFs themselves can be further divided into α-MoRFs, β-MoRFs and ι-MoRFs depending on their structure in their bound state (Vacic et al., 2007). Because of their functional relevance and distinct physicochemical characteristics, some predictors were developed to identify MoRFs. One of such predictors is MoRFpred (Disfani et al., 2012). It uses a SVM predictor based on a manually extracted set of data on MoRFs, from complexes enhanced with alignment information. The other, energy-based predictor of MoRFs, is ANCHOR (Dosztányi et al., 2009). It first computes local contacts in long disordered regions to ensure that the fragment is not prone to fold on its own. Then ANCHOR estimates per residue energy gain from interaction of disordered residues with a potential globular partner. In principle, ANCHOR is methodologically similar to disorder predictor IUpred (Dosztányi et al., 2005a). Later attempts to predict protein binding motifs found that predicting MoRFs is an extremely difficult task (Jones and Cozzetto, 2015). Although the computational methods are able to achieve high specificity, sensitivity and precision of current MoRF predictors are very low.

Features exhibited by MORFs can be more broadly classified as coupled folding and binding (Dyson and Wright, 2005; Schlessinger et al., 2011; van der Lee et al., 2014). An example of folding upon binding could be p21 and p27. These proteins regulate different cyclin-dependent kinases responsible for the control of cell-cycle progression in mammals (van der Lee et al., 2014).

An extreme example of folding and binding properties of IDPs is phase transitions observed in this class of proteins (Brangwynne et al., 2015; Latysheva et al., 2015; Toretsky and Wright, 2014). IDPs were shown to form “assemblages” thanks to self-association and low affinity binding. Those protein aggregates are in the form of hydrogels and, just like individual IDPs, are subject to tight regulation (Latysheva et al., 2015). IDP assemblages were shown to be associated with disease states and are now under active studies.

1.2.2. *IDPs and diseases*

It was found that the majority of protein disease-associated mutations are found in IDPs (Habchi et al., 2014; Uversky et al., 2008). IDPs are associated with many crucial cellular functions, especially in eukaryotes. Therefore, their dysfunction or inappropriate expression can result in pathological conditions (Babu et al., 2011).

Over-expressed flexible regions of IDPs are likely to cause molecular titration and bind uncontrollably to other molecular partners, or may produce fibrillar aggregates. (Babu et al., 2011; Diella et al., 2008). Under-expression on the other hand may perturb cell signalling, or regulation (Grimmler et al., 2007). It was also found that around 20% of disease-related mutations in IDPs cause local disorder-to-order transitions (Vacic et al., 2012).

The most widespread associations of IDPs are with cancer and neurodegenerative disorders (Uversky et al., 2008). Unsurprisingly, relations between IDPs and cancers are an intensive field of study.

Perhaps the most well-known cancer-related IDP is p53 – an apoptotic tumour suppressor. About 70% of its interactions are carried out by disordered regions (Oldfield et al., 2008). p53 is a large signalling hub and along with its close partner Hdm2/Mdm2 it is responsible for regulating expression of genes involved in the induction of apoptosis, DNA repair, response to stress and the progression of cell cycle (Anderson and Appella, 2003; Uversky et al., 2008; Vousden and Lu, 2002). Therefore, the loss of p53 function can easily lead to cancer. It has been observed for: lung, oesophagus, colon, breast, liver, hemopoietic and reticuloendothelial cells (Hollstein et al., 1991). Looking at the mutations at the protein domain level, p53 domain is the most prevalent mutant in breast and colon cancer (Nehrt et al., 2012).

Another widely studied disordered protein target is p27(Kip1), an inhibitor of cyclin-dependent kinases where under-expression is associated with various types of cancer (Grimmler et al., 2007).

Other prominent examples of associations between IDPs and cancer include: BRCA-1 – associated with breast cancer, which has 79% disordered residues (Mark et al., 2005); α-fetoprotein – marker of cancer and fetal abnormalities is an intrinsic molten globule; structured by its natural ligands (Abelev, 1971; Deutsch, 1991; Uversky et al., 2008).

A different group of diseases associated with malfunctioning of IDPs are neurodegenerative disorders, i.e. Parkinson's, Alzheimer's, dementia with Lewy bodies, or multiple system atrophy (Uversky et al., 2008). In particular, a fully disordered protein α-synuclein is associated with all of these diseases (also called synucleinopathies). Synuclein can form a variety of fibrillar aggregates in neurons; depending on the morphology they have different names (e.g. Lewy bodies, Lewy neuritis, glial cytoplasmic inclusions). Interestingly, α-synuclein is also a model for the development of many experimental techniques to study IDPs (Marsh et al., 2006; Rao et al., 2010; Tamiola and Mulder, 2012).

IDPs are also associated with prion diseases, cardiovascular disease, and type II diabetes. A comprehensive review of associations between intrinsic protein disorder and diseases is available (Uversky et al., 2008).

1.3. Experimental characterization of protein disorder

Theoretical considerations laid out in previous sections should be put into an experimental context. A wealth of experimental techniques enable identification and structural characterization of intrinsically disordered proteins. The techniques fall into 2 main groups:

(1) biophysical and biochemical,

(2) spectroscopic.

Those techniques usually provide a different level of detail about IDPs (Figure 2).

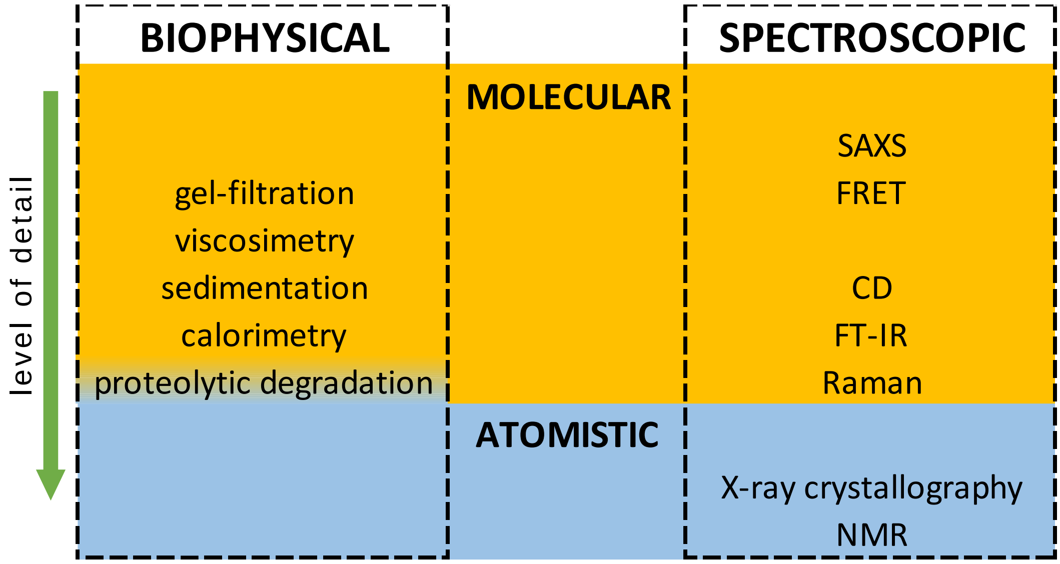


Figure 2. An overview of experimental techniques used to study intrinsically disordered proteins.

1.3.1. *Biophysical and biochemical methods*

Biophysical techniques allow the identification of disordered proteins and estimate the amount of intrinsic disorder within them (Figure 2). The most widespread methods include: gel-filtration, viscosimetry, sedimentation, calorimetry and proteolytic degradation (Eliezer, 2009; Habchi et al., 2014). They are all based on the fact that hydrodynamic volume, or the radius of gyration in the case of IDPs is increased in comparison to ordered, compact proteins of the same mass.

1.3.2. ***Spectroscopic techniques***

Comparable, yet higher impact information may be derived from small angle X-ray scattering (SAXS) and Foerster resonance energy transfer (FRET) experiments (Habchi et al., 2014; Mittag and Forman-Kay, 2007). These experiments can provide more detailed information about the shape of molecules; help to locate disordered regions within the molecule; or to gain long-range information on interacting sites. Often these results are used in combination with NMR experiments to derive ensembles of disordered proteins (Jensen et al., 2013; Krzeminski et al., 2013). Other spectroscopic techniques may be generally divided into 2 groups giving molecular or atomistic details (Figure 2). The first group includes circular dichroism (CD), infrared (FT-IR) or Raman optical activity (ROA) spectroscopies, with CD being the most widely used technique for this purpose. These methods enable the identification and quantification of intrinsic disorder in proteins, however they alone do not allow to locate or characterize the disorder in atomistic detail. Also, with CD spectroscopy it is difficult to distinguish disordered proteins from loopy proteins (low secondary structure content) with no repetitive secondary structure (Liu et al., 2002).

1.3.3. ***X-ray crystallography***

More detailed information come from X-ray crystallography, which is the most widespread technique for protein structure determination. It provides atomistic details averaged over time and space. Due to noncoherent X-ray scattering, disordered regions are not visible in the diffraction pattern (Vladimir N. Uversky, 2013). This way X-ray crystallography accounts for the indirect evidence of disorder. Such data has to be approached with caution, as intrinsic disorder is not the only cause of the lack of electron density.

There are some factors which may impact the disordered regions. Crystal packing can cause disorder-to-order transitions and result in under-determining disordered regions (Dunker et al., 2002). Another factor impacting the disordered state could be disordered binding segments crystallized along with the investigated protein. This concerns proteins and peptides, as well as small molecule ligands (Dunker et al., 2002).

On the other hand, there are factors which may cause ordered regions to appear disordered. This is mostly attributed to wobbly domains and crystal contacts which may account for the diffraction pattern to lack some reflections and therefore be interpreted as disordered (Dunker et al., 2001; Ward et al., 2004b).

1.3.4. *NMR characterization of protein structures and disorder*

1.3.4.1. What does an NMR spectrum tell us?

NMR, unlike X-ray crystallography, is capable of producing a set of output structures (an ensemble) giving insight into the dynamics of the protein (Jensen et al., 2013; Kosol et al., 2013; Lindorff-Larsen et al., 2005; Mittag and Forman-Kay, 2007). The ensembles are in fact alternative possible solutions of restraints obtained from the NMR experiment. In NMR, as in any other experimental technique, the structures are models fitting the experimental data. They are therefore constrained by the technique used and experimental conditions. This should be considered at all times.

In NMR, a pre-processed protein (deuterated solvent, host organism grown on special media to contain ^{15}N or more ^{13}C, etc.) is measured in solution (Roberts, 1993). All NMR parameters in some way reflect the molecular conformation of the studied system (Cavanagh et al., 2007). Most relevant parameters for the studies of proteins are gathered in Table 1 and summarized below.

From using Nuclear Overhauser Effect Spectroscopy (NOESY) a set of constraints on the structure can be derived (Mittag and Forman-Kay, 2007). They are sensitive to residue distances as their intensity is proportional to r^{-6}. Therefore, they provide some accurate constraints on the structure.

Chemical shifts (CS) give information on the local structural propensities and represent a population-weighted average over all interconverting conformers in an ensemble (Jensen et al., 2013). They are of particular use in analysing proteins undergoing transitions including changes in the amount of secondary structures (Tamiola and Mulder, 2012). Drawback of CS is the need of reference shifts. Different methods were developed for generating random-coil chemical shift libraries and for assessing experimental data using CS (Berjanskii and Wishart, 2006, 2007; Tamiola and Mulder, 2012).

Table 1. Summary of the most common NMR parameters.

parameter	information	IDP utility	requirements and difficulties
NOESY (Nuclear Overhauser Effect SpectroscopY)	Connectivity between residues	Conformation of ordered parts; transient interactions within disordered regions	High level of deuteration
CS (chemical shifts)	Local structural propensities (secondary structure)	Transient conformations in disordered regions	Reference needed: (1) back-calculations of CS from known structures; (2) random-coil models; (3) population-weighted averages of reference shifts from different conformations and a random coil shift (De Simone et al., 2009) CSs are temperature and pH sensitive
RDC (residual dipolar couplings)	Relative orientations of secondary structure elements; Long-range interactions;	Conformational sub-states in disordered ensemble; Fold in ordered parts	Aligning medium (could alter the ensemble (Dames et al., 2006)
PRE (paramagnetic relaxation enhancement)	Long-range interactions	Characterization of poorly populated states	Paramagnetic spin label in the sample (as above, could alter the ensemble (Mittag and Forman-Kay, 2007))
SC (scalar coupling)	Backbone dihedral angles	Characterization of topology	

based on: Jensen et al., 2013; Roberts, 1993; Rosato et al., 2013

Residual dipolar couplings (RDCs), similarly to CS, represent population-weighted average of the ensemble, but give information on the relative orientations of secondary structure elements and long-range interactions. RDCs require a reference frame, therefore an aligning medium is required (Jensen et al., 2013). This parameter can be back-calculated from structures (e.g. using Flexible-meccano (Kragelj et al., 2013; Ozenne et al., 2012)), hence is also useful for validation of derived ensembles.

Paramagnetic relaxation enhancement (PRE) provides information about long-range interactions. It requires a paramagnetic probe attached to an amino acid. PRE can give insight into sub-states probed by CS and RDCs because the line-broadening in PRE reflects the timescales responsible for the relaxation of interacting sites.

Scalar couplings (SC) give information on backbone dihedral angles and can help in identifying protein's topology and amino acid conformations.

Often, as an additional constraint in solving NMR ensembles, SAXS data is used to provide constraints on the shape of the protein determined by the radius of gyration (R_g) (Mittag and Forman-Kay, 2007; Sibille and Bernadó, 2012).

There are several avenues that can be followed in order to derive an ensemble of structures from this set of data (Cavanagh et al., 2007; Fisher et al., 2010; Jensen et al., 2013). They are discussed below in sections 1.3.4.2 and 1.3.4.3.

1.3.4.2. Generating models – restrained REMD

One way to obtain the structural ensemble of the studied protein is to use molecular dynamics (MD) restrained by the experimental observables. The MD simulation is usually performed as replica exchange molecular dynamics (REMD), meaning that multiple models run simultaneously and replicas are exchanged between simulations over time (Allison et al., 2009; Esteban-Martín et al., 2010; Wu et al., 2009). This approach requires the restraints to be weighted as pseudo-energy terms biasing the simulations. For robust results, it also requires a large number of distance restraints, the number of which is difficult to estimate in advance (Fisher et al., 2010).

1.3.4.3. Generating models – ensemble selection

An alternative approach is ensemble selection (Jensen et al., 2013). In this method, a library of conformations is pre-generated. Then, suitable conformations are selected on the basis of matching the experimental restraints. Initial conformations can come either from statistical coil models (Flexible-meccano (Ozenne et al., 2012), TraDES (Marsh and Forman-Kay, 2012), or others), fragment selection (Fisher et al., 2010), or various crude MD approaches (Lei and Duan, 2007). Selection is then made by minimizing the differences between the generated ensemble parameters and experimental values. Methods such as ENSEMBLE (Marsh and Forman-Kay, 2012), SAS (Chen et al., 2007), both utilising Monte Carlo for selection of sub-ensembles, or ASTEROIDS (Nodet et al., 2009) which utilizes an evolutionary algorithm, have been developed.

1.3.4.4. Assessment of NMR models

Derived NMR protein ensembles can then be assessed either by knowledge-based measures, or on the basis of model versus data comparison (Rosato et al., 2013).

Knowledge-based assessment relies on some prior knowledge of the protein conformations and protein biophysics. The knowledge-based information is usually derived from highly resolved X-ray structures. This causes some debate in the field, as the crystal environment is typically different to solutions used in NMR (Rosato et al., 2013). Knowledge-based methods assess either dihedral angle distribution (e.g. ProCheck (Laskowski et al., 1996)), atom packing (e.g. Molprobity (Chen et al., 2010)), or by energy refinement using energy potentials or structural fragments (van der Schot et al., 2013).

Model versus data assessment utilizes either all of the experimental data used to generate the model and checks how well it is fitted to the data, or preferably, is done by cross-validation on the data left out from the model generation. It can be done as a restraint analysis, or analysis of any of the NMR parameters (CS, PRE, RDC, or combinations thereof).

Some approaches were also developed to perform fully automated validation of models. But until now, blind experiments (e.g. CASD-NMR) proved that still there are many cases where manually validated and refined proteins have some important differences to automatically generated ones (Rosato et al., 2012).

1.3.4.5. NMR of IDPs

Considering the utility of NMR for investigations of intrinsically disordered proteins, it should be first noted that it gives a dynamic picture of the disordered ensemble. This wealth of dynamic information about the disordered state is not available from any other experimental technique. Secondly, NMR studies proved that, contrary to some opinions, disorder is not an artefact of sample preparation, but an actual *in vivo* phenomenon (Bertini et al., 2011; Ito et al., 2012; Uversky and Dunker, 2013). That is, *in vitro* observations do not represent an artificial state which would be otherwise impossible to maintain within the cell due to the crowding and the cellular mechanisms that degrade unfolded proteins (Janin and Sternberg, 2013),

Atomistic details of disordered proteins come either from X-ray crystallography or NMR. Therefore it is justified to seek comparisons between these two methods. Current capabilities for the size of proteins feasible for experiments to analyse are in favour of crystallography; where it is possible to solve even massive protein complexes (e.g. bacterial respiratory complex I (PDB id: 3RKO) having 6 unique chains and 2038 amino acids in total (Efremov and Sazanov, 2011)). NMR on the other hand is limited at present to chains of less than 300 amino acids (i.e. up to 50 kDa) and the structure set is dominated by chains of around 100 amino acids (Ota et al., 2013). Also, the cellular localization of the most abundant X-ray and NMR-solved proteins are different. X-ray structures are biased towards cytoplasmic proteins, whilst NMR solved structures are more populated with nuclear proteins (Ota et al., 2013). This bias is fortunate for NMR investigations of IDPs, as it is estimated that nuclear proteins contain many disordered regions (Ward et al., 2004b).

As mentioned before, crystallography provides only indirect evidence for protein disorder (i.e. regions of missing electron density – no structure), while NMR directly estimates the dynamic behaviour of disordered regions. Apart from this, NMR samples are measured in different environments, without the effects of crystal packing. Usual concentrations of protein used in NMR studies are in the millimolar range (0.1-3 mM[1]). For IDPs, the concentrations required are even lower (Dyson and Wright, 2004; Jensen et al., 2013). Because of this, as confirmed by in-cell experiments, the behaviour of IDPs observed by NMR *in vitro* likely corresponds to the actual functional states (Bertini et al., 2011; Ito et al., 2012). Nevertheless, the accurate determination of chain dynamics at atomistic level is demanding. Chemical shifts or residual dipolar couplings can unambiguously show which residues are disordered, but the determination of the actual ensembles is still a challenging task (Cino et al., 2012; Jensen et al., 2013; Tamiola et al., 2010).

One of the reasons for that is a relatively sparse set of constraints for the structure determination coming from NMR parameters. In this case, when the disordered region is under-determined many possible conformations that fulfil NMR constraints do not guarantee the validity of generated models (Fisher et al., 2010; Jensen et al., 2013). Even when the models agree with experimental data there is no guarantee that the generated ensemble is true. This problem can be bypassed to a certain extent using statistical methods (e.g. Bayesian Weighting). Bayesian Weighting can estimate the uncertainty of the weights assigned to each conformer in an ensemble. Relying on both experimental data and theoretical predictions, such methods can calculate probability densities over weights and estimate the uncertainties of each assignment (Fisher et al., 2010). Nevertheless, the identification of disordered regions is unambiguous and is an important take-home message for the following work.

An important drawback of the currently available NMR ensembles is the shortage of reference experimental data other than the models generated by the authors. Only recently did it become compulsory to deposit experimental constraints into BMRB

[1] http://www2.chemistry.msu.edu/facilities/nmr/900MHz/MCSB_NMR_sample.html

(Biological Magnetic Resonance Data Bank) alongside the deposition of the structures to the PDB. At present (September 2015), there are 219 entries (heteronuclear NOE values) in BMRB, compared to over 11,000 NMR-solved protein structures in the PDB (about 2% coverage).

1.3.4.6. Disorder classification from NMR structures

Because raw experimental NMR data are still scarce, disorder has to be analysed based on the PDB ensembles. To date, there have been several attempts to develop methods that would robustly identify disorder in NMR PDBs, as opposed to more commonly used crystallographic disorder (derived from missing densities from X-ray-solved structures).

One such method is MOBI (`http://protein.bio.unipd.it/mobi/`; (Martin et al., 2010)). MOBI classifies each residue as ordered or disordered based on a set of criteria derived from NMR PDB ensembles and heuristics. The criteria by which MOBI classifies residues as ordered or disordered are:

(1) Cα distance of superposed residues;

(2) DSSP annotation consistency;

(3) Standard deviations of internal coordinates (ϕ and ψ angles).

The method also uses heuristics in the form of regular expressions to make its annotations consistent (e.g. a single ordered residue flanked by disordered residues is re-annotated as disordered).

The original MOBI paper benchmarks the method against 19 NMR structures from the CASP8 disorder prediction category (manually annotated dataset; Noivirt-Brik et al., 2009). MOBI achieved the harmonic mean of precision and recall (F-value) of 0.939. This shows that automated MOBI method reliably reproduces manual annotations of NMR PDB ensembles.

The other approach aimed to determine the cut-off value for disorder-order classification using per residue deviations in IDPs (Ota et al., 2013). The authors were comparing X-ray and NMR PDB structures trying to determine the discriminating value maximizing the correlation (Matthews Correlation Coefficient; MCC) between X-ray and NMR sets of the same proteins. The consensus value of 3.2Å was found giving an MCC (Matthews Correlation Coefficient) of 0.63. Imperfect correlation between the two sets can be explained by a relatively small overlapping set of structures (55 proteins) and varying experimental conditions which have an impact on disorder (i.e. some regions become ordered in crowded, highly concentrated environments). The latter conclusion can be easily observed in disorder databases (e.g. DisProt or mobiDB; see section 1.4).

1.3.4.7. Future

Considering the future utility of NMR data there is definitely a need for a wider availability of raw experimental restraints. BMRB will undoubtedly grow in size and due to the progress of Structural Genomics centres the data is likely to accumulate at an even faster pace. Current protocols for solving NMR structures are far from perfect and as some experts point out, for IDPs there are many possible solutions to each set of experimental data. Therefore it is crucial that all of the future ensembles are not only stored as PDB structures, but also should have corresponding raw data freely and easily accessible.

There are also initiatives that concentrate on the use of NMR for IDP studies, such as EU FP7 initiative IDPbyNMR (`http://www.idpbynmr.eu`). This project was established to raise awareness of the importance of NMR in the field of IDPs, to train young scientist to gain expertise in this area and to network researchers across Europe interested in the IDP studies by organizing meetings and workshops. Initiatives like this and other possible future initiatives are likely to provide direction to the field, establish more robust NMR protocols to deal with IDPs and produce substantially more IDP ensembles.

1.4. Databases of structural information on intrinsic disorder

Given the abundance of intrinsically disordered proteins (section 1.1.3), their relevance (section 1.2) and structural features characterized by a variety of experimental techniques (section 1.3), several databases collecting the information on IDPs have been developed. They vary greatly in size and the extent of information they provide on intrinsic disorder (Table 2).

All of the databases summarized here concentrate on different aspects of the phenomenon of intrinsic protein disorder, sometimes also combining other databases within their own framework (e.g. mobiDB (section 1.4.2) and D^{2}P^2 (section 1.4.3) contain DisProt (section 1.4.1) annotations).

Table 2. Summary of IDP databases.

	DisProt	IDEAL	PED	mobiDB	D^{2}P^2
number of entries	694	582	26	80,370,243	10,429,761
last update	05/2013	06/2015	09/2015	09/2014	2012
manual annotations	+	+			
experimental data	+	+	+	+	
disorder predictions				+	+

1.4.1. ***DisProt***

DisProt was the first publicly available database of protein disorder (Sickmeier et al., 2007; `http://www.disprot.org/`). Its latest release (version 6.02 from 24 May 2013) contains 694 proteins with 1,539 disordered regions.

Disorder/order annotations in DisProt emerge from experimental data. However, unlike some other databases, or datasets used for disorder classification (e.g. missing X-ray density), DisProt does not only rely on 3D structural information. The database also combines an array of biophysical and spectroscopic methods, such as circular dichroism, fluorescence, hydrogen-deuterium exchange, sensitivity to proteolysis, SAXS and viscosimetry. A full list of experimental techniques used as disorder detection methods is available at the DisProt website: `http://www.disprot.org/view_detection.php`. All of the entries in DisProt are curated, the conclusion and data based on publications are author-approved.

DisProt also links all its entries to sequence databases (UniProt, UniGene, SwissPort, TrEMBL) and for each protein (or disordered region, if applicable) it includes a functional narrative section, sequence annotations (disordered regions, ordered regions, unknown regions) and links to relevant literature.

DisProt is considered the gold standard in disorder databases because of its manual curation and some other disorder databases contain DisProt entries (e.g. mobiDB and D^2P^2). Nevertheless, the database has some flaws.

DisProt entries are not updated when new experimental evidence becomes available. Once an entry is deposited and annotations are made, no further data appears to be added to it. DisProt is a manually curated database so amending data within the database, as well as creating new entries might be cumbersome. Nevertheless, comparing DisProt entries to other regularly updated databases, the discrepancies are apparent (for instance acyl carrier protein; DisProt record DP00416 and mobiDB entry P0A6A8).

Another issue with DisProt that is also apparent in other databases is the ambiguity with which ligands, cofactors or ions should be treated. A good example here is cytochrome c. The protein has a well-known structure, yet DisProt treats it as a fully disordered protein (DP00006; UniProt ID: P00004). This data is based on the observation that cytochrome c without the heme group becomes fully disordered (Stellwagen et al., 1972). In contrast, there are no disorder containing entries in mobiDB (maximum disorder content 5.71% from 2GIW NMR structure), which bases its annotations on solved 3D structures. Obviously, the biologically relevant structure of cytochrome c is heme-bound, hence all of the solved structures indicate the lack of disorder.

1.4.2. ***mobiDB***

mobiDB is a database of protein disorder and mobility annotations (Di Domenico et al., 2012; Potenza et al., 2015; `http://mobidb.bio.unipd.it/`). The database combines DisProt annotations, disorder/order classification derived from X-ray and NMR structures deposited in the PDB (annotations based on the MOBI method (section 1.3.4.6)) and the results of 10 disorder/order classification methods (VSL2B, GlobPlot, RONN, ESpritz-Xray, ESPritz-NMR, ESpritz-DisProt, DisEMBL-HL, DisEMBL-465, IUpred-short, IUpred-long; all methods discussed in section 1.5.1).

The latest version of mobiDB (version 2.2) was released in September 2014 and contains annotations for 80,370,243 entries from UniProt (Swiss-Prot: 546,000; TrEMBL: 79,824,243). UniProt also links directly to mobiDB in its cross-reference section of each UniProt entry.

Each entry in mobiDB includes all structural information on the given protein along with the results of disorder/order classification methods, DisProt annotations (where applicable) and the results from the STRING database of interactions (Szklarczyk et al., 2015), as well as known Pfam families (Finn et al., 2014). MobiDB also employs a naïve scheme providing users with a consensus disorder annotation of the protein. Regions where all experimental data agree (all X-ray and NMR structures, and DisProt annotations) the region is annotated as ordered or disordered, otherwise the regions is annotated as ambiguous.

MobiDB treats all entries at UniProt ID level. Mutants of a known UniProt entry are indexed under the same mobiDB entry. MobiDB therefore loses the ability to trace disorder-to-order or order-to-disorder transitions and, as mentioned above in DisProt (section 1.4.1), changes in the environment and presence/absence of ligands.

MobiDB is the most comprehensive database of information on intrinsic protein disorder to date. It combines multiple sources of structural information (no biophysical data, except from that inferred from DisProt), disorder predictors and domain information (Pfam).

1.4.3. ***D^{2}P^2***

D^{2}P^2 is the database of disordered protein predictions (Oates et al., 2013; `http://d2p2.pro/`). It contains the results of 9 disorder predictors (VL-XT, VSL2B, PrDOS, PV2, ESpritz (3 flavours), IUpred (2 flavours); the predictors are discussed in section 1.5.1). The database concentrates on providing disorder predictions for genomes on the transcript level. Its latest release contains 1,765 genomes with 10,429,761 sequences from 1,256 distinct species (no viral genomes). The sequence data is based on SUPERFAMILY 1.75 database (last update 8 November 2011; Gough et al., 2001). Unlike mobiDB, D^{2}P^2 concentrates not on UniProt-level data, but rather on SCOP superfamilies (predicted or annotated by SUPERFAMILY) and ENSEMBL genome data (Cunningham et al., 2014).

D^{2}P^2 also includes annotations coming from DisProt (as discussed previously), predictions of post-translational modification (from PhosphoSitePlus) and MoRFs (from ANCHOR (Dosztányi et al., 2009); discussed in 1.2). Some disorder predictors were not included in the database, because they were deemed too computationally expensive by the authors (e.g. DISOPRED2).

The 2 largest databases of protein disorder predictions – D^{2}P^2 and mobiDB – differ in data presentation style and concentrate on slightly different aspects of disorder. Notably, mobiDB attempts to annotate all possible proteins using data inferred from structural information (X-ray and NMR), while D^{2}P^2 uses only disorder predictors. D^{2}P^2 also tries to characterize individual genomes and enables users to define their own thresholds for consensus disorder annotations (25%, 50%, 75% or 100% agreement between different disorder predictors). Not only is the database constructed to be able to browse particular genomes, rather than UniProt IDs as in mobiDB. It also provides statistics on average disorder content within particular genomes and compares the predictions of disorder predictors used by the database. These genome and predictor-wide comparisons of data are provided as a possible source of future improvements to the disorder prediction methods.

Nevertheless, it is apparent that both databases share a large degree of redundancy and may benefit from joining forces, rather than competing with one another. See section 1.4.6 for summary and discussion of this issue.

1.4.4. *pE-DB (PED)*

Protein ensemble database (previously: pE-DB; now: PED; `http://pedb.vib.be/`) collects experimental data on IDPs (Varadi et al., 2015, 2014).

The current release of PED (version 2.0) contains 26 entries (as of 22 September 2015) with 70 ensembles. This includes protein fragments, e.g. heat shock protein beta-6 (HSPB6) residues 24-160, 40-160, 57-160 are listed as 3 entries.

PED relies on experimental data providing information on the ensembles of intrinsically disordered proteins from NMR (PREs, CSs, RDC, J-couplings, NOEs, paramagnetic shifts; for a discussion of how NMR is used to obtain the ensembles of IDPs see section 1.3.4) and SAXS. It also hosts some purely computational models coming from MD simulations (not necessarily NMR data-constrained simulations, as discussed in section 1.3.4.2).

Compared with other resources, PED is a small database (hundreds in DisProt and IDEAL, and millions in mobiDB and D^2P^2). It is however the only resource that concentrates on providing exhaustive experimental data on IDP ensembles. Provided that it is actively maintained, it may become a valuable resource which could enhance the understanding of intrinsically disordered proteins and the development of computational techniques to study them. At present, the computational methods are limited by the availability of data and usually resort to what is available in the PDB (either missing X-ray data or NMR PDB ensembles).

1.4.5. ***IDEAL***

IDEAL is the database of Intrinsically Disordered protein with Extensive Annotations and Literature (Fukuchi et al., 2014, 2012; `http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/`). It was first released in November 2011 and its latest update (version 12/Jun/2015) contains 582 entries.

The idea behind IDEAL is to provide the highest quality information on functional regions in IDPs (e.g. PTM sites, MoRFs undergoing coupled folding and binding, referred by the authors of the database as ProS – protean segments).

All of the entries in IDEAL have corresponding PDB structures. Disordered regions were identified based on several premises: missing X-ray density, regions interfering with protein crystallization, high RMSD regions in NMR PDB ensembles, and using other spectroscopic methods (i.e. CD) for disorder identification. The final approach most closely resembles the methodology used in DisProt.

ProS (regions undergoing disorder-to-order transitions upon binding) were categorized into 2 sub-groups: known ProS and circumstantial ProS. Known ProS were annotated if there is experimental evidence of a disordered state in unbound form and ordered state in bound form. In case of the lack of hard evidence (e.g. some data available only for protein’s homologue) the regions are annotated as circumstantial ProS.

IDEAL also provides SCOP and Pfam assignments and annotations of known binding sites and PTMs extracted from the literature and UniProt annotations. For proteins with known binding partners, IDEAL illustrates protein-protein interaction networks and shows the structures of protein complexes.

Again, as in the case of mobiDB and D^2P^2, IDEAL shows some significant similarities to DisProt. It is a manually curated database that relies on experimental data to annotate ordered and disordered regions. The novelty that IDEAL introduces are the ProS, regions undergoing disorder-to-order transitions upon interaction. The authors of the database claim that it is a broader concept than MoRFs and not limited to interactions with other proteins.

Overall, IDEAL provides a useful and reliable resource on IDPs, albeit rather small and biased (the first release of IDEAL consisted of 120 human nuclear proteins), due to the manual curation of the database.

1.4.6. ***Summary***

There are many resources that aggregate information on intrinsically disordered proteins. The three main groups of databases are:

(1) Manually curated databases: DisProt, IDEAL;

(2) Databases of disorder predictions and derived data: mobiDB, D^2P^2;

(3) Database of experimental data: PED.

These databases have distinct approaches and certainly add value to the understanding of intrinsic protein disorder. Researchers in the broad field of IDPs are interested in different aspects of disorder, and hence alternative sources of information aimed at different needs are available.

However, clearly there is some redundancy in what is available and some databases have become outdated (compare Table 2). With any database, the main issue is not to establish it, but to maintain and keep it up to date. The study of IDPs is still in the early days and it seems that it would be easier and more beneficial for the scientific community if more than a single group was involved in the development of any database that is supposed to last.

Extending the development of any given database to more than a single group would actually make it serve the community better. It could help to establish some standards and address the most important problems raised by the community, rather than perceived by any single research group.

At present, both mobiDB and D^2P^2 aggregate data from DisProt. Those databases provide less detailed information than DisProt itself, but it is a clear signal that further integrations are possible and desirable. In fact mobiDB and D^2P^2 share a great deal of redundancy (Table 3). There are some design and more substantial differences behind those resources, e.g. mobiDB is based on UniProt, while D^2P^2 uses SUPERFAMILY. Nevertheless, the majority of calculations required to produce the predictions for millions of proteins, as well as storage and infrastructure is common.

The same stands for other resources. IDEAL could be easily incorporated into mobiDB, just as DisProt was. PED also could also become a part of some larger database. It is more difficult to unambiguously interpret some of the experimental data deposited in PED, but it would be advantageous to see all of the experimental evidence on disorder for a given protein in one place.

Table 3. Comparison of mobiDB and D^2P^2 databases contents.

		mobiDB	**D^2P^2**
main unit		UniProt	ENSEMBL; SCOP
predictors	ESPritz (3 flavours)	+	+
	IUpred (2 flavours)	+	+
	PrDOS		+
	VSL2B	+	+
	VL-XT		+
	PV2		+
	DisEMBL	+	
	GlobPlot	+	
	RONN	+	
experimental data	X-ray	+	
	NMR	+	
annotations	Pfam	+	+
	STRING	+	
	SCOP		+
	DisProt	+	+
	IDEAL		+
	PhosphoSitePlus		+
	ANCHOR		+

Finally, today, mobiDB is leading in terms of its comprehensiveness and size, but it already lags behind the sequence databases. The latest UniProt release was in

September 2015, while the latest mobiDB release was in July 2014 (14 months behind UniProt). The amount of data available on intrinsically disordered proteins is already vast, however still far from complete, as these targets are difficult to characterize experimentally. Hopefully, in the future, some consortium that unifies the storage of information on IDPs will arise and the data will be kept up-to-date, well maintained and expertly curated.

1.5. Computational predictions of intrinsic disorder

In parallel, or even at times ahead of the development of experimental techniques there have been many attempts to study intrinsically disordered proteins computationally. The vast majority of these studies focused on the development of disorder/order classification methods. These methods take protein sequence as an input and provide a binary prediction of order or disorder on a per-residue level, often accompanied by various kinds of score or confidence values.

Disorder prediction methods provided several important conclusions. (1) They proved that disorder is not random, but has some important physical characteristics that make it possible to identify such regions. (2) Disorder is evolutionarily conserved. (3) Consequently, it was possible to estimate the amount of intrinsic disorder within various genomes (as described in 1.1.3). (4) That in turn enabled the conclusion that disorder is a development of evolution and is more prominent among higher organisms. (5) Some functional classes of proteins are enriched in IDPs (see above, section 1.2) (Kozlowski and Bujnicki, 2012; Ward et al., 2004b).

1.5.1. *Sequence-based disorder prediction methods*

1.5.1.1. Disorder as a classification problem

Sequence-based disorder prediction methods treat intrinsic protein disorder in a binary fashion. The aim of these methods is to classify each residue in the query sequence as ordered, or disordered. Treating disorder this way greatly simplifies the problem, as disorder can have many preferred conformations, functional roles and features (compare section 1.2; van der Lee et al., 2014). Still, this simplification proves stable and fruitful, as the performance and popularity of these methods show.

Since the prediction of disordered residues can be defined as a disorder/order (binary) classification problem, many algorithms developed in other areas of science are applicable in this case.

1.5.1.2. Approaches to disorder predictions

The first disorder prediction algorithms were developed by A. Keith Dunker's group between 1997 and 1999 (Li et al., 1999; Romero et al., 1998, 1997). Predictions were based on simple rules derived from the observations of IDPs known at the time and three neural network (NN) predictors which specifically identify disordered regions of different lengths (i.e. short, medium and long disorder).

As the interest in IDPs grew, a plethora of prediction methods arose. Disorder prediction has been one of the categories assessed during the biennial experiment CASP (Critical Assessment of protein Structure Prediction) between CASP5 in 2002 and CASP10 in 2012.

At present, sequence-based disorder prediction methods fall into one of the three categories:

(1) Physics-based methods,

(2) Machine learning-based methods,

(3) Meta methods.

Sub-sections below describe each of these approaches and give examples of some state-of-the-art methods. Some recent papers also review the disorder prediction methods (Ali et al., 2014; Dosztányi et al., 2010; Oates et al., 2013; Orosz and Ovadi, 2011).

1.5.1.3. Physics-based predictors

Intrinsically disordered regions in proteins have some characteristic features that make them distinct from ordered regions (Habchi et al., 2014; van der Lee et al., 2014). Physics-based disorder predictors take advantage of this fact and use one or more of the features of intrinsic disorder: low sequence complexity (Li et al., 1999); low hydrophobicity and high net charge (CH-plot) (Uversky et al., 2000); disorder promoting residues (such as proline, glycine, or charged amino acids) (Dunker et al., 2008); low probability to form energetically favourable contacts (Uversky et al., 2000).

Although physics-based methods are methodologically simpler than the machine learning approaches, they achieve excellent results that prove IDPs have clear physical characteristics. For comparison, refer to Mobility Continuous Assessment (MoCA; Walsh et al., 2015) and the results presented in Chapter 3. Some of the popular physics-based disorder predictors are discussed below and IUpred, the most popular and widely used physics-based method is described in more detail in sub-section 1.5.1.3.a.

iPDA (Su et al., 2007) combines DisPSSMP2 position-specific scoring matrices of amino acid physicochemical properties which are focused on the disorder propensities of residues, along with several other sequence predictors which account for sequence conservation, secondary structure, sequence complexity and the analysis of hydrophobic clusters.

GlobPlot (Linding, 2003) is based on a scale of residue propensities to form globular or non-globular states. It identifies disordered regions by calculating the propensity values over protein regions as differences between 'random coil' and 'secondary structure' propensities and applying a low-pass data filter.

FoldIdex (Prilusky et al., 2005) is based on hydrophobicity/net charge plots of sequence fragments. By using a sliding window approach it identifies putative intrinsically disordered regions within the sequences.

a. IUpred

IUpred (Dosztányi et al., 2005a, 2005b) attempts to predict regions that are unlikely to form stabilizing contacts and thereby form intrinsically disordered regions.

The method is based on a 20 by 20 table of interaction energies derived from a large non-redundant set of high resolution globular structures deposited in the PDB. By using a table of physicochemical properties and analysing the sequence environment of a given residue it assumes that the ordered regions make enough favourable contacts to maintain a stable 3D structure.

IUpred relies on a single sequence, rather than a MSA. The energy values are averaged over a window of 21 residues. The method uses two energy tables – one for short disorder and one for long (above 30 consecutive residues) disordered regions.

1.5.1.4. Machine learning-based predictors

Most of the current top performing disorder/order classification methods are based on supervised machine learning approaches; RONN (Yang et al., 2005), DisEMBL (Linding et al., 2003), VSL2 (Peng et al., 2006), ESpritz (Walsh et al., 2012), DISOPRED (Jones and Cozzetto, 2015; Jones and Ward, 2003; Ward et al., 2004b). This group of methods acknowledges the complexity of the phenomenon of intrinsic disordered (no single property can fully describe IDPs) and take advantage of the available experimental data to train the predictors.

The majority of machine learning-based methods use missing X-ray data for training (e.g. DISOPRED2, ESPritz-Xray, RONN, VSL2B, DisEMBL). Most of the methods use either neural networks or SVM approach.

Because X-ray data is biased towards short disordered regions (less than 30 residues), most machine learning-based methods achieve poorer performance on long disorder (Ali et al., 2014; Monastyrskyy et al., 2014). Some methods deal with long disordered regions by creating separate ‘flavours’ of predictors aimed at long disordered regions (e.g. ESpritz (Walsh et al., 2012) or CSpritz (Walsh et al., 2011)). Other methods try to combine short and long disorder predictions in one framework (e.g. VSL2 (Peng et al., 2006)). Finally, there are some methods aimed specifically at long disordered regions (e.g. SLIDER (Peng et al., 2014a)).

SLIDER is a method based on logistic regression (unique in disorder predictions), which uses the following features: amino acid composition, physicochemical properties of residues, sequence complexity and combinations of these features. Although the method was specifically designed to tackle the issue of long disordered regions, it achieves results comparable with a more universal method VSL2 (Peng et al., 2006).

Another interesting approach is s2D, which tries to combine secondary structure and disorder predictions within a single method (Sormanni et al., 2015). Secondary structure prediction is generally a solved problem now (Jones, 1999), but so far there were no methods that attempted to solve both of those issues simultaneously. The method utilizes a well-known (both in secondary structure and disorder predictions) neural network framework, but enhances it with the use of extreme learning machines (ELMs). ELMs speed-up the training process and allow for an evaluation of more models, as well as have a proven capability to perform universal approximations. The method was not validated or tested externally (e.g. in CASP experiment), but it is an interesting approach to the problem of disorder predictions.

DISOPRED is one of the most popular machine learning-based disorder predictors. It is described in more detail below, in sub-section 1.5.1.4.a.

a. DISOPRED

The DISOPRED method was initially developed in 2003 (Jones and Ward, 2003). It is based on neural networks and was trained on X-ray data.

DISOPRED2 became a successful disorder prediction method that offered low false positive rate and enabled an accurate estimation of the amount of disorder across different genomes and, in effect, between different domains of life (Ward et al., 2004a, 2004b). DISOPRED2 utilizes a combined approach taking advantage of SVM training and neural network classifier. At the time of its release it was a state-of-the-art method, but with time it became apparent that it under-predicts long disordered regions (CASP8 assessment; Noivirt-Brik et al., 2009).

DISOPRED3 was developed in 2014 and returned to the neural network framework to try and tackle the problems of previous DISOPRED releases in dealing with long disordered regions (Jones and Cozzetto, 2015). The method combines 3 approaches (neural network, SVM and nearest neighbour classifier) within an umbrella neural network.

DISOPRED3 outperforms DISOPRED2 according to all metrics and again it is a high specificity method (> 99% specificity, except at terminal protein regions).

1.5.1.5. Meta predictors

One of the most popular and effective ways to address the problem of disorder/order classification is through the use of meta methods. This class of methods addresses the problem by combining the prediction of different individual methods to produce a single consensus result. Meta methods usually perform very well and were highly ranked in recent CASP disorder evaluations (Monastyrskyy et al., 2014, 2011).

Some examples of disorder meta predictors include MetaDisorder (Kozlowski and Bujnicki, 2012), PONDR-FIT (Xue et al., 2010), PrDOS (Ishida and Kinoshita, 2007), MFDp (Mizianty et al., 2010), POODLE (Shimizu et al., 2007).

PrDOS uses conditional neural fields and metaPrDOS is a meta method using 5 other servers (`prdos.hgc.jp/cgi-bin/top.cgi`).

MFDp is a meta method using SVM fed with evolutionary profiles, secondary structure, solvent accessibility and dihedral angles (`http://biomine.ece.ualberta.ca/MFDp.html`).

POODLE is another SVM method combining 3 other SVMs: Poodle-S, Poodle-L and Poodle-W.

A different example of a meta server is MeDor (Lieutaud et al., 2008). It combines different predictors, such as VL3, VL3H (Radivojac et al., 2003), VSL2B, HCA (hydrophobic cluster analysis; Callebaut et al., 1997) but it does not provide a consensus prediction. Instead, it gives the user an opportunity to compare the predictions of those methods using a visual output.

1.5.2. *Computational simulations of intrinsic disorder*

Apart from one-dimensional sequence-based disorder predictors, some approaches to computationally model the dynamic nature of IDPs have also been attempted. All of the simulations are based either on the use of all-atom molecular dynamics (MD), some form of coarse-grained molecular dynamics, or on Metropolis Monte Carlo simulations in an implicit solvent. Some of the major computational simulations to date are gathered in Table 4.

1.5.2.1. MD-based disorder simulations

There are essentially no large scale studies of IDPs using molecular dynamics methods, as MD is still computationally expensive and limited to short proteins (Baker and Best, 2013). However, using MD simulations it is possible to gain an insight into intrinsic disorder that is not possible using sequence-based disorder predictors, such as information about the protein ensembles, protein dynamics or disorder-to-order transitions (Baker and Best, 2013; Bueren-Calabuig and Michel, 2015).

There was a single study thus far carried out by D.E. Shaw’s group to simulate using Anton – a purpose-build supercomputer for MD simulations – the dynamic behaviour of acyl-CoA-binding protein (ACBP), a 112 residue protein (Lindorff-Larsen et al., 2012). The computations showed acceptable agreement with NMR ensembles, however leaving the output structure more compact than its experimentally-solved counterpart. An important drawback of this approach is its exclusivity (there is only a handful of Anton machines with a very limited access) and the computation remains demanding, even though the MD timescales accessible via Anton are impressive (200 μs in the discussed case).

Generally, all-atom MD methods are thought to produce models that are too compact (Henriques et al., 2015). One of the proposed causes of this is an issue with protein-water interface parametrization (Henriques et al., 2015). Nevertheless, most of modern force fields are capable of reproducing the behaviour of IDPs reasonably well (Palazzesi et al., 2015).

Even given the issues above, folding-upon-binding (or disorder-to-order transition) simulations are possible using MD. They are usually limited to small binding motifs, as in the case of MDM2 (119 residues) and p53 TAD domain (12 residues) binding simulation (Bueren-Calabuig and Michel, 2015).

1.5.2.2. Coarse-grained disorder simulations

Coarse-grained models are also popular in the studies of IDPs (Baker and Best, 2013). They allow for more computationally tractable simulations of longer protein chains.

One of the most active groups in the field of coarse-grained simulations of IDP is Rohit V. Pappu's lab. They developed a continuum implicit solvation model called ABSINTH, tailored to the simulation of disordered proteins (Table 4) (Vitalis and Pappu, 2009b). In this model, the transfer of polypeptide from gaseous environment to solvent is the sum of direct mean field interactions. Then ABSINTH estimates the range of multi-body interactions by the size of per atom implicit solvation shell. It helps to select which interactions should be recalculated in each coming step in addition to the motions coming from the next step of simulation (Vitalis and Pappu, 2009a). Using this solvation model in combination with Metropolis Monte Carlo (MMC) simulations, Pappu's lab performed a series of experiments on peptides (20-49 residues) to investigate the behaviour of some natural and synthetic model IDPs (Das and Pappu, 2013; Das et al., 2012; Mao et al., 2010; Vitalis et al., 2008, 2007). The main purpose of these series of experiments was to discover some rules governing the behaviour of IDPs under physiological conditions. From these experiments it was found that:

(1) polyglutamine froms disordered, collapsed globules with a potential to form multimeric structures depending on the chain length and physicochemical environment (Vitalis et al., 2008);

(2) net charge per residue modulates the radius of gyration or compactness of disordered peptides (Mao et al., 2010);

(3) oppositely charged peptides depending on the intra-chain charge balance form either generic Flory random coils (mixed charges) or hairpin-like conformations (charges separated) (Das and Pappu, 2013).

A different study on a small, but more diverse system was carried out on basic regions of leucine zippers (bZIP-bRs) and revealed the relationships between the amino acid composition and the amount of helicity (Das et al., 2012). Clearly all of these studies provided some important insights into how small model IDP systems behave and to the driving forces in the behaviour of the disordered ensembles. Unfortunately, to date the group had not presented any studies on larger proteins (50 residues or more), nor on proteins containing disordered regions.

One of the other studies concerning disorder-to-order transition focused on a region of sortase by performing multiscale enhanced sampling (MSES) (Moritsugu et al., 2012). The method used is a modification of the classical MD approach that combines the all-atom approach with a MD force field and a coarse-grained system described by a simplified Hamiltonian that was arbitrarily assigned to the disordered region. The study revealed the flexibility and the transition of the investigated fragment correctly, but the need for an experimental starting structure and the somewhat arbitrary assignment of the disordered region remain as serious drawbacks of this approach.

Several other studies also used MD techniques to: describe a domain or fragment of a protein (Ganguly and Chen, 2009; Potoyan and Papoian, 2011); investigate the binding of a fragment of a disordered domain to a globular target (Higo et al., 2011; Staneva et al., 2012); determine the folding mechanism of a molten globule protein (Naganathan and Orozco, 2011). Refer to Table 4 for methodological details.

Table 4. Some major computational simulations of IDPs.

source	protein	PDB id (no. res.)	method	remarks
Lindorff-Larsen et al. 2012	acyl-CoA binding protein (ACBP)	1k19 (112)	MD CHARMM22* explicit TIP3P water	ANTON supercomputer 200 µs simulation
Vitalis et al., 2008	Polyglutamine peptides (Q_{5-45})	- (5-45)	Metropolis Monte Carlo (MMC) OPLS-AA/L ABSINTH solvation model	
Mao et al., 2010	21 positively charged protamines	- (24-49)	MMC OPLS-AA/L ABSINTH solvation model	
Das et al., 2012	15 bZIP-bRs	- (27-28)	MMC OPLS-AA ABSINTH solvation model	CAMPARI package TREx sampling
Das & Pappu, 2013	30 $(Glu\text{-}Lys)_{25}$ peptides	- (25)	MMC OPLS-AA ABSINTH solvation model	CAMPARI package TREx sampling
Moritsugu et al., 2012	Sortase	2kid (148; 90 used)	MD (MSES) AMBER ff03 + in-house V_{CG}	Arbitrary assignment of a different potential to disordered residues
Staneva et al., 2012	p53 & TRTK-12 (fragments)	1dt7 & 1mwn/1mq1 (15 & 12 used)	MD (GROMACS) + MC (PROFASI) OPLS-AA/L explicit SPC/E water	Binding of disordered peptides to a globular protein
Ganguly & Chen, 2009	CREB (KID & pKID)	1kdx (28)	REX-MD CHARMM22/CMAP GBSW implicit solvent	12 replicas (270-500 K) 200 ns simulations (160 ns KID folding)
Naganathan & Orozco, 2011	NCBD	2kkj & 1zoq & 1jjs (59 & 47 & 50)	Cα-Gō (GROMACS) & MD AMBER 99 SB* TIP3P water	MD only for 2kkj (residues 1-51) 100 ns simulations
Potoian & Papoian, 2011	Histone tails (H2A, H2B, H3, H4)	- (14-38)	REX-MD AMBER ff99SB TIP3P water	3 µs simulations
Higo et al., 2011	Neural restrictive silencer factor (NRSF; fragment)	- (15)	McMD AMBER parm94 & parm96 TIP3P water	Binding of a NRSF/REST fragment to Sin3 PRESTO package 10 ns runs (64 & 512 runs at 1,000 K)

Although the presented simulations make it possible to study IDP systems in detail they are currently limited to small proteins or peptides (in all but two cases smaller than 60 residues) and they often require a starting experimental structure as input. On the other hand, these methods provided some important insights into how IDPs behave in model systems and have revealed some physicochemical principles governing their behaviour. The presented MD simulations also proved that existing force fields are capable of capturing the disordered behaviour of proteins, although the output structures are often too compact. Another major difficulty in studying IDPs – sampling – is bypassed using more refined methods, such as replica exchange (REX), multiscale enhanced sampling (MSES) or multicanonical sampling (e.g. Higo et al., 2011). A methodologically different approach was mastered in Rohit V. Pappu's lab; it employs Metropolis Monte Carlo simulations using General Born approximation-based implicit solvent model (i.e. ABSINTH).

There is a recent review available which summarizes the recent advancements in simulating IDPs and relates the computational and experimental results to our understanding of the physical bases of intrinsic disorder and how they mediate protein function (Chen, 2012).

1.5.2.3. CABSflex

A different approach, designed specifically to simulate protein flexibility is CABSflex (Jamroz et al., 2014, 2013b), also available as a webserver (Jamroz et al., 2013a). CABSflex is a coarse-grained method based on the CABS model (Kolinski, 2004).

The CABS method (Carbon Alpha, Beta and Side-chain) is a coarse-grained model, which treats the polypeptide chain in a simplified manner (Kolinski, 2004). It uses Cα atoms and centres of mass of the peptide bond as a simplified backbone, and Cβs with a virtual side-chain pseudo-atom at the centre of mass to represent each side-chain. The space in CABS model is also simplified. The main chain moves along a cubic lattice, while the side-chains are placed off-lattice. The simulation additionally uses a knowledge-based force-field (Kolinski, 2004). After the simulation is finished, the chain is recalculated to an all-atom model. In CABSflex this rebuilding is performed using BBQ and ModRefiner methods (Gront et al., 2007; Xu and Zhang, 2011).

CABSflex was recently made available as a webserver for simulations of protein chain flexibility (Jamroz et al., 2013a). Starting from a submitted structure it attempts to sample available neighbouring conformations by a series of random Monte Carlo conformational transitions modulated by the force-field. In a series of comparisons this method produced results in good agreement with MD simulations (Jamroz et al., 2013b), accurately reproducing the behaviour observed in NMR ensembles (Jamroz et al., 2014). Their latter paper shows that starting from a rigid structure CABSflex is even more accurate in predicting the protein fluctuations than MD simulations which are far more computationally expensive (Jamroz et al., 2014). A very good agreement with NMR ensembles was achieved for both ordered but flexible proteins and disordered proteins with a mean R_S (Spearman's rank correlation) of 0.72 (± 0.15) on a dataset of 140 non-redundant proteins (0.64 ± 0.23 for MD).

Results from the CABSflex papers show that stochastic simulations based on knowledge-based potentials are capable of reproducing what is observed by NMR experiments even for disordered proteins. The papers also show that MD simulations have no advantage over coarse-grained methods and CABSflex was able to produce more accurate results than MD in most cases. Although the method proved to be quick and reliable, it also poses a serious drawback – it requires a high-quality starting structure as an input.

1.5.2.4. DynaMine

DynaMine is a machine learning linear regression model that predicts NMR order parameters (S^2) from protein sequence (Cilia et al., 2013). The NMR order parameter is an experimental NMR parameter which represents how restricted is the movement of an atomic bond vector with respect to the reference frame. In case of DynaMine, the method predicts backbone (N-H bond) order parameters (Cilia et al., 2013).

There have been previous attempts to derive order parameters from structures using either derived relationships between the distances within experimental structures (Zhang and Brüschweiler, 2002), or machine learning (neural network) methods (Trott et al., 2008). DynaMine, however, is the first approach to try and obtain order parameter values from sequence alone.

The order parameter determines the level of constraints in an atomic bond, a value of 1 means that the movement is completely restricted (rigid) and 0 means there are no constraints on the movement of the bond (highly disordered) (Berjanskii and Wishart, 2008). Experimentally, order parameters can represent movements at different time scales – from femtosecond to low millisecond. In DynaMine, because reference ("experimental") order parameter values are derived from chemical shifts, the predicted order parameter values represent a mix of different timescales (Cilia et al., 2013).

In DynaMine, no training data comes directly from experimentally derived S^2 values. Instead, the data is calculated from reported experimental chemical shifts (which are easily obtained and commonly deposited in BMRB) using the RCI method and then re-scaled to match S^2 values best (Berjanskii and Wishart, 2005). This re-calculation may contribute to a bias in DynaMine, because, as the authors determined, comparing S^2 values re-calculated using RCI with experimentally determined S^2 values on a set of 16 proteins, Pearson's correlation equals 0.685 between the two sets.

DynaMine is intended to serve as a disorder predictor (when S^2 values are binned into disordered/flexible, context-dependent and ordered/rigid regions), helping to distinguish regions of different structural organization (e.g. folded domains, disordered linkers, molten globules, pre-structured binding motifs) and ultimately to re-evaluate residue propensities previously derived from X-ray and spectroscopic data (Uversky and Dunker, 2010).

1.6. Aims and outline of thesis

The focus of this PhD is to further bridge the gap between the experimental data on IDPs and computational methods. Most of current computational techniques rely on sequence data alone to provide a binary disorder/order classification and do not yield any higher-dimensional information. Few attempts have been made to computationally simulate the behaviour of disordered regions *in silico.* To date, only MD or Monte Carlo approaches have been used. These methods require a starting structure they may attempt to simulate and significant computational power and time, while being applicable only to small or medium-sized proteins. This significantly impairs any advances for large-scale computations or the routine use of these techniques.

In this work, I address the issue of utility of protein structure prediction techniques to the prediction of intrinsically disordered protein ensembles and protein backbone dynamics.

Chapter 2 describes FRAGFOLD-IDP approach, which takes advantage of FRAGFOLD fragment assembly method to predict the ensembles of intrinsically disordered proteins from sequence. The ensembles are then clustered and analysed in terms of their per-residue fluctuations. FRAGFOLD-IDP is assessed on a set of 200 NMR PDB ensembles. The method is compared to a naïve method and analysed on a basis of protein CATH class, disorder content and the correlation between the quality of structure and backbone dynamics predictions. Finally, FRAGFOLD-IDP is evaluated against other methods that provide similar information, or data that can be related to protein backbone dynamics – crystallographic B-factor predictions, NMR order parameter predictions and disorder classification methods. I show that FRAGFOLD-IDP achieves superior performance than any of those methods and only FRAGFOLD-IDP and DynaMine, an NMR order parameter predictor, achieve results that are significantly better than the naïve method.

Chapter 3 explores FRAGFOLD-IDP performance on the task of disorder/order classification. It is a well-established problem in bioinformatics and it gives a broader spectrum for comparisons of the method. The assessment includes a wide array of top performing disorder classification methods which utilize different approaches and were trained on different data. I show that although FRAGFOLD-IDP is a protein backbone dynamics prediction method it performs well on the task and is on par with the top performing binary disorder/order classification methods.

Chapter 4 describes a consensus protein backbone dynamics predictor which integrates the predictions of FRAGFOLD-IDP and DynaMine. Using a neural network approach, I combine the input methods to produce the predictions which are significantly better than any of the input methods. The consensus predictor established a new state-of-the-art in intrinsically disordered protein backbone dynamics predictions.

Chapter 5 summarises the major contributions of this work in a biological context, possible applications and future directions where the use of FRAGFOLD-IDP and the consensus predictor could help to address some of the outstanding problems in the fields of intrinsically disordered proteins and structural bioinformatics. I also provide a general discussion of the limitations in studying intrinsically disordered proteins using computational techniques.

The results of this work are likely to be of great interest to the whole IDP research community, both experimentalists and theoreticians alike. This work should help to obtain dynamic information for proteins of unknown structure. It could also help to gain a better understanding of the physical bases of intrinsic disorder in proteins. This would not only enable us to predict disorder using a novel approach, but also provide a better insight into how intrinsically disordered regions behave in different classes of proteins, or how to quantify disorder in a different manner.

Finally, I expect these findings to facilitate the design of disorder-to-order transitions, what in turn could give a better understanding of the phenomenon and help to study some of the IDP-related diseases.

Chapter 2. FRAGFOLD-IDP *DE NOVO* PREDICTIONS OF INTRINSICALLY DISORDERED PROTEIN ENSEMBLES

2.1. Background

2.1.1. *Purpose and challenges*

Chapter 1 laid out the current state of knowledge about intrinsically disordered proteins and the techniques used to study them. At the core of the computational techniques that are currently used to study protein disorder (section 1.5) are disorder predictors and simulation techniques. The predictors help to quantify the amount of disorder within genomes and led to the realization that intrinsic protein disorder is a prevalent phenomenon which has its characteristics (see section 1.5.1). Simulation techniques, on the other hand, provide dynamic information on protein disorder. Those methods showed that modern force fields are accurate enough to reproduce the behaviour of disordered ensembles observed from NMR experiments (see section 1.5.2). Nevertheless, these 2 approaches have some serious limitations. Disorder predictors limit our knowledge to the classification of residues as ordered, or disordered – they provide no structural or dynamic information, treating disorder as a binary property. Simulation techniques, either Molecular Dynamics (MD)-based, or Metropolis Monte Carlo (MMC)-based, provide a greater level of insight into intrinsic disorder than disorder predictors, but at the same time they require starting structures to simulate. This itself poses a serious limitation and leaves those simulation techniques at a proof-of-concept stage, since they cannot be used to study the behaviour of proteins of unknown structures. The other limitation that simulation techniques suffer from is their high computational cost. As discussed in section 1.5.2, most current simulation approaches are limited to small proteins, usually shorter than 100 residues. The simulations of even those small targets are computationally expensive and non-trivial, requiring days of simulations on modern computer clusters.

Disorder is not a binary property and there are different degrees of protein disorder, preferred conformations, or intra-chain interactions (Dyson and Wright, 2005; van der Lee et al., 2014). In this work I address the problems of current computational approaches to enable a greater insight into the dynamics of intrinsically disordered proteins of unknown structure. To achieve this, FRAGFOLD, a *de novo* fragment-based protein folding approach, is used (see section 2.1.3). By using a *de novo* approach, the study is not limited to proteins of known structure, but for the purpose of benchmarking NMR results from the PDB database are utilized. The heterogeneity of the intrinsically disordered protein ensembles are measured using per-residue RMSD. It quantifies the behaviour of disordered ensembles and provides more information than disorder classification.

2.1.2. *Other approaches used to predict the properties of IDP ensembles*

Although the approach presented in this work is innovative in terms of the methodologies used, it is not the first attempt to address the issue of predicting backbone protein dynamics in intrinsically disordered proteins, or the properties of intrinsic protein disorder from sequence.

To the best of the author’s knowledge there are two other approaches that address this problem – the DynaMine method (Cilia et al., 2014, 2013) and the attempts made by David Baker’s group using Rosetta (Wang et al., 2011). Hence, before moving on to describing the details of the method developed in this work, let us discuss the basics and results of those approaches.

2.1.2.1. DynaMine – summary of the results

The method is described in detail in section 1.5.2.4. Briefly, DynaMine uses a linear regression model to predict NMR order parameter (S^2) for each residue in the query sequence. Unlike most modern disorder predictors, DynaMine relies on a single sequence, instead of a sequence profile (compare section 1.5.1) and uses a 51 residue sliding window (a shorter window for smaller targets) as an input to the model.

The method was assessed in several ways: its ability to reproduce known experimental order parameter values, as a disorder/order classifier and in a more exploratory context predicting the properties of targets not having structural information available, but having known domain annotations. The last aspect was studied to elucidate the biological relevance of the predictions and to show the potential of the method in recognizing the regions of different structural organizations, such as disordered linkers, molten globules, or pre-structured binding motifs (Cilia et al., 2013).

DynaMine training relies on order parameter values calculated from chemical shift data, rather than directly from experimental order parameters. The authors evaluated the calculations of order parameters using the RCI method (Berjanskii and Wishart, 2008) and achieved Pearson's correlation of 0.68 on a set of 16 proteins (1,581 residues) with known chemical shift and order parameter data.

Disorder/order classification was evaluated on two DisProt-derived datasets (DisProt is described in section 1.4.1). The first DisProt dataset overlapped with available chemical shift data to verify the utility of calculated order parameters in reproducing DisProt annotations (< 90% sequence identity with DynaMine training data). The second DisProt dataset consisted of other entries not having related NMR data.

The first assessment showed that calculated order parameters reproduce DisProt binary classification remarkably well (AUC = 0.92) and that optimal order/disorder threshold values for both calculated order parameters and DynaMine predictions are similar (between 0.77 and 0.80 S^2 values). The second evaluation compared the results of 6 disorder predictors (IUpred (Dosztányi et al., 2005b); RONN (Yang et al., 2005); PrDOS2 (Ishida and Kinoshita, 2007); VSL2 (Obradovic et al., 2005); FoldIndex (Prilusky et al., 2005) and ESpritz (Walsh et al., 2012)) with DynaMine. In this evaluation DynaMine performed very well, on par with the top disorder predictors

(PrDOS2, ESpritz). The drawback of this analysis is that DynaMine was trained on a closely related set of proteins (< 90% sequence identity) in one approach and in the other the sequence similarity between the training and test sets is unknown. This suggests a possibility that the model was overtrained on the data and the evaluation gave the advantage to the previous methods that were trained on DisProt data.

The DynaMine case studies include 8 proteins: human p53, human CREB-binding protein, human cyclin-dependent kinase inhibitor p27, E1A protein from human adenovirus 5, human calpastatin, HIV Nef protein, PaaA2 human antitoxin and Phd human antitoxin. For those targets, it is possible to estimate the boundaries of folded domains basing on DynaMine predictions. The authors also assessed their predictions looking at the distributions of predicted S^2 values in known domains and disordered regions. The distributions of folded and disordered domains are partially overlapping, but in most cases have distinct maxima.

Overall, DynaMine predictions provide a good qualitative tool for the analysis of sequences. The predictor does not have a bias towards proteins of known structures and can also be used as a disorder/order classification tool (Cilia et al., 2013). Because the underlying data (order parameters calculated from chemical shifts) is a source of initial bias, DynaMine predictions should rather be interpreted in terms of relative, not absolute values. Indeed, in most analysed cases, known secondary structure elements correspond to peaks in predicted S^2 values, so do some known interaction motifs. The authors also suggest that DynaMine predicted S^2 values might have a meaning in the sense of domain stability.

2.1.2.2. Study of IDPs using Rosetta

The study by Wang et al. attempts to use Rosetta to model disordered regions in proteins (Wang et al., 2011). The authors observe that protein structure prediction methods attempt to find the lowest potential energy conformations, whereas the native state of the protein corresponds to the free energy minimum. In the case of ordered proteins this discrepancy does not play a major role, since the entropies of ordered (folded) proteins are similar. Whereas for disordered proteins, the entropic effects should play a more important role. To address this, the authors arrive at two possible solutions. The first solution, starts from predicted low energy models and attempts to enhance the models with disorder identification by optimising a free energy function. The free energy function takes different forms in the case of internal (mid-sequence) disorder and protein termini. Minimization of the free energy function is achieved by enumerating possible disorder/order assignments along the sequence. The second approach relies on a pre-computed disorder/order classification. Residues classified as ordered follow normal Rosetta potentials (Leaver-Fay et al., 2011; Rohl et al., 2004; Simons et al., 1999, 1997). Disordered residues are treated as interacting via repulsive interactions only – van der Waals interactions when the simulations is carried out using only centroids and Lennard-Jones potential at all-atom level.

The results for the first approach (calculating the free energy from low-energy Rosetta models) are assessed for the termini and internal loops separately on a set of 38 proteins in total. In the assessment of 8 targets with disordered termini the methodology showed some minor improvements in discriminating native structures considering the free energy rather than the potential energy. For the internal disorder set (30 proteins), the authors enumerated all possible stretches of disorder using the free energy calculations, starting from pre-calculated low energy structures. Comparing the results to a null model (assuming all residues are ordered), improvements were made in 16 out of 30 cases. The shapes of potential energy and free energy landscapes are similar. Overall, this approach requires a significant computational cost. It first generates an ensemble of structures and then the ensemble is recalculated to obtain the free energy values. The latter calculations enumerate all possible stretches of disorder (within given constraints).

The results of the second approach (using repulsive-only interactions on pre-annotated disordered regions) were illustrated on 4 cases (1 overlapping with the previous 38 case test set) using 3 methodologies: Rosetta *ab initio* with DISOPRED2 predictions, CS-Rosetta (Shen et al., 2009) with chemical shift data and a comparative modelling approach. All those approaches yielded better results in terms of overall structure prediction quality over the standard Rosetta approach. Nevertheless, given the small size of the test set it is difficult to draw general conclusion as to the utility of this approach.

Overall, the method is aimed at improving the predictions of ordered proteins, accounting for disordered regions in a different fashion. The authors achieved this goal, at the same time gaining some insights into the behaviour of the disordered regions themselves. However, the work does not go beyond the binary order/disorder scheme and shows modest improvements over a typical approach to the structure prediction problem, regardless of the solution used.

2.1.3. ***FRAGFOLD folding engine***

The method developed in this chapter relies on FRAGFOLD calculations, hence the method is introduced here. FRAGFOLD is a state-of-the-art *de novo* fragment-assembly method for protein structure prediction (Jones and McGuffin, 2003; Jones, 2001, 1997; Jones et al., 2005). It bases on similar principles as Rosetta (described above) and was shown to be effective in *de novo* structure predictions of globular proteins (Jones and McGuffin, 2003; Jones, 2001).

Because of the complexity of the protein folding problem, most protein structure prediction approaches utilize some form of coarse graining to limit the conformational search. In FRAGFOLD, the coarse graining is achieved by using structural fragments (peptides), instead of individual residues. This allows for both spatial and structural coarse-graining. FRAGFOLD fragment library consists of three types of fragments:

(1) Supersecondary fragments;

(2) 9-residue (fixed-length) fragments;

(3) Dipeptide and tripeptide fragments.

The fragments are pre-computed and assembled into FRAGFOLD fragment library (Figure 3 greyed area). They come from a set of highly resolved protein structures. The selection of large fragments (supersecondary and 9-residue) is determined by a threading score based on the multiple sequence alignment (MSA) and secondary structure predictions provided as input for FRAGFOLD (Figure 3 top). Small dipeptide and tripeptide fragments are universal and do not depend on the input protein sequence. At each position in the target sequence, a shortlist of large fragments that both agree with the prediction of secondary structure and which have a favourable threading energy are produced. The lists of large and small fragments are sampled randomly during the folding run to generate each new conformation. 50% of fragments are taken from the pre-calculated large fragment list and the other half is taken from the set of small fragments.

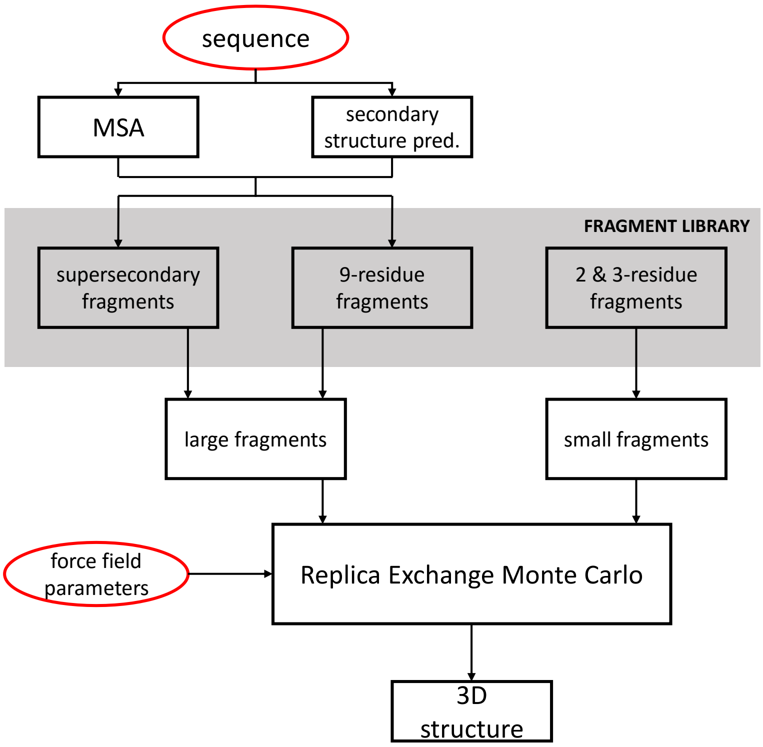


Figure 3. FRAGFOLD flow diagram. Elements defined by the user are shown as red ellipses.

The sampling is performed using Replica Exchange Monte Carlo (REMC) with Simulated Annealing (SA) approach (Figure 3 bottom). The simulation starts from a random conformation. All fragment insertions are made at random positions within the sequence and are accepted with a probability $e^{-\Delta E/kT}$. The temperature of the simulation decreases as the simulation progresses. The starting temperature (T_0) is determined by taking 10 times ΔE of the largest energy difference between a pool of random conformations. The temperature is decreased by $0.05T_0$ after each 10% of the maximum number of moves. After each fragment is inserted, side-chain conformations are also generated from a library of rotamers.

The FRAGFOLD objective energy function force field embodies pair-wise potentials of mean force determined by inverse Boltzmann equation – short range (6 residues or less apart) and long range (7 or more residues separation) potentials; solvation potential, hydrogen bonding, structure compactness and steric terms (Jones and McGuffin, 2003; Jones, 2001). The final energy function is in the form:

$$\begin{aligned}E = {} & W_1 \cdot SR \cdot E_{short-range} + W_2 \cdot LR \cdot E_{long-range} + W_3 \cdot SOLV \cdot E_{solvation} + W_4 \cdot STERIC \cdot E_{steric} \\ & + W_5 \cdot HB \cdot E_{H-bond} + W_6 \cdot COMPACT \cdot E_{compactness}\end{aligned}$$

Weighting components of the potential function (W_1 to W_6) are determined by comparing the standard deviations of each term, across an ensemble of random conformations, to that of the short range (W_1) term. Random conformations are generated by threading fragments randomly from N- to C-terminus and performing a steric clash check. If steric clashes are observed the conformation is discarded and a new one is generated in its place. Further weighting (SR, LR, SOLV, STERIC, HB, COMPACT) can be user-defined, and by default the STERIC terms are weighted by an additional factor of 3.0, while all other terms (SR, LR, SOLV, HB, COMPACT) are left unweighted.

The simulations run up to a pre-defined number of maximum annealing steps or until simulation time runs out. The number of steps is typically defined depending on the sequence length. For proteins shorter than 120 residues 5,000,000 steps are made and for longer proteins 10,000,000 steps.

Each FRAGFOLD run generates a single three-dimensional protein model. The whole FRAGFOLD workflow is summarized in Figure 3.

2.2. Dataset

2.2.1. *Construction*

Because this work moves away from the binary disorder/order classification, NMR PDB ensembles are used, instead of relying on the classical DisProt dataset (Sickmeier et al., 2007), or missing electron densities from X-ray data (e.g. as in DISOPRED2 (Ward et al., 2004b)).

As described in the Introduction (section 1.3.4), NMR ensembles contain dynamic information about the disordered state. It is also straightforward to extract per-residue backbone dynamics data from the ensembles deposited in the PDB (described further in section 2.3.4).

The dataset of NMR-resolved disordered proteins was constructed based on mobiDB (Di Domenico et al., 2012) – a comprehensive database of experimentally solved disordered proteins and sequence-based disorder predictions (described in section 1.4.2). The database assesses the disorder content in proteins on the basis of the MOBI method (described in section 1.3.4.6; Martin et al., 2010). It also contains DisProt annotations (Sickmeier et al., 2007), where applicable.

The dataset was extracted from mobiDB version 1.2 (accessible via: `http://mobidb.bio.unipd.it/`). The database was queried to extract only the proteins:

(1) solved by NMR;

(2) that have at least 95% coverage of PDB sequence with UniProt;

(3) between 50 and 150 amino acids long;

(4) that have no other molecules in the PDB file, as indicated by `COMPND` PDB keyword;

(5) that have at least 5 consecutive disordered residues, as indicated by MOBI;

The criterion for using only NMR-derived structures was chosen to include the dynamic information that X-ray PDB structures lack. The coverage criterion was set to exclude possible protein constructs, or single domains of multi-domain proteins, where the native state might have different disorder characteristics. Limits on size were set, so that the simulations that are carried out on those sequences would be computationally tractable. PDB files with more than 1 molecule were excluded, to remove potential binding-induced changes in the behaviour of the disordered ensembles. Finally, the criterion of disorder content was set, so that proteins without a significant disordered region are not considered.

Despite the criteria above, some other proteins that were previously studied using different computational techniques (Table 5) were added to the set for comparison.

2.2.2. *Characterization*

Applying the criteria from 2.2.1 resulted in a dataset of 200 proteins in the dataset. The average length of the protein in the dataset is 105 residues (Figure 4) and the average disorder content is 33.7% (Figure 5). There are 28 proteins (14%) with at least 50% of disorder content and 3 fully disordered proteins. The disorder distribution is close to what is predicted for the human proteome (Pentony et al., 2010).

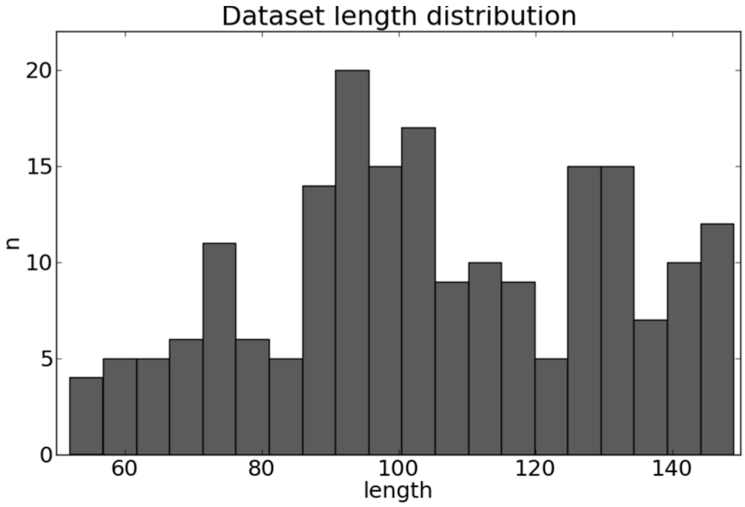


Figure 4. Length distribution of proteins in the dataset.

Considering only long disorder regions (30 amino acids or more) the dataset contains 29 such proteins.

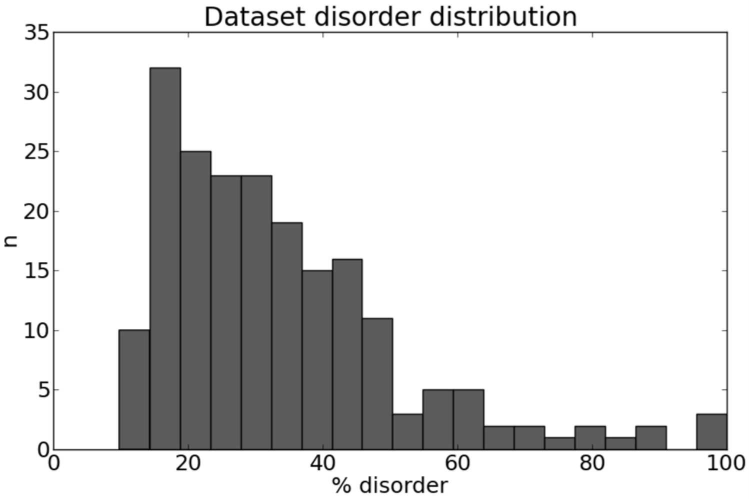


Figure 5. Disorder content distribution in the dataset.

2.2.3. ***Benchmark set***

Based on the initial dataset constructed on the basis of mobiDB, a separate subset was extracted from it. The subset is designed to serve as a benchmark of the method, so that an evaluation of the parameters can be performed (sections 2.3 and 2.4). It contains 28 proteins of varied sizes and includes all localizations of disorder (N-terminal, C-terminal, mid-sequence and full disorder). The average length of the proteins in the benchmark set is 110 residues and average disorder content is 32.2%. It also contains some of the previously computationally studied targets that do not meet the criteria set on the full dataset (described in 2.2.1). For a full list of benchmark set proteins refer to Table 5.

Table 5. Summary of benchmark set proteins.

PDB id	UniProt id	name	length	no. disordered residues	disorder location	remarks
1b75	P68919	ribosomal protein L25	94	16	m	
1dmo	P62155	calmodulin	148	58	N+m+C	
1dvd	P01040	stefin A	98	11	N+m	
1fkr	P62942	fk506 and rapamycin-binding protein	107	12	M	
1jw3	O27635	conserved hypothetical protein MTH1598	140	34	M	
1k19	Q9NG96	chemosensory protein CSP2	112	38	m+C	
1ni7	P0AGF2	hypothetical protein ygdK	149	15	N+m	
1nin	P0C178	plastocyanin	105	37	m	molten globule
1nnv	P44199	hypothetical protein HI1450	107	18	N+m	
1nti	P07107	acyl-CoA-binding protein	86	13	m	from Lindorff-Larsen et al., 2012
1ovq	P0A8I1	hypothetical protein yqgF	138	41	m	
1soy	P27838	CyaY protein	106	10	m	
1tac	P12506	HIV-1 transactivator (TAT) protein	86	67	N+m+C	fully disordered; molten globule
1tiv	P12506	HIV-1 transactivator (TAT) protein	86	86	N+m+C	fully disordered
1xhs	P0AE48	hypothetical UPF0131 protein ytfP	113	29	m+C	
1xpw	Q9Y547	LOC51668 protein	143	17	m+C	
1y6d	P0C5S4	phosphorelay protein luxU	114	11	m	
1zza	O75324	stannin	90	44	N+m+C	
2aqa	Q6Q547	H/ACA ribonucleoprotein complex subunit 3	57	57	N+m+C	fully disordered
2fki	P0AF50	protein yjbR	118	27	m	
2hgk	Q46919	hypothetical protein yqcC	117	27	m	
2jo6	P0A9I8	nitrite reductase [NAD(P)H] small subunit	110	11	m	
2ju4	P04972	retinal rod rhodopsin-sensitive cGMP 3',5'-cyclic phosphodiesterase subunit gamma	87	86	N+m+C	fully disordered
2k36	P68466	protein K7	149	25	N+m	
2k5t	P37613	uncharacterized protein yhhK	128	13	m	
2kkj	P45481	CREB-binding protein	59	29	N+m+C	non-bound part of CBP; from Naganathan & Orozco, 2011
2kx4	P03714	tail attachment protein	117	41	N+m+C	
2py1	P0A552	deoxyuridine 5'-triphosphate nucleotidohydrolase	129	10	N+m	molten globule

annotations of disordered residues are taken from mobiDB
disorder location: N - N-terminal disorder; m - mid-sequence; C - C-terminal disorder

2.3. Methods

The general concept behind this computational method, which will be referred to as FRAGFOLD-IDP, is to generate an ensemble of structures using FRAGFOLD and compare the backbone dynamics emerging from this ensemble of models to that of the NMR experimental ensembles. In this section, the details of the methodology are described.

2.3.1. *FRAGFOLD-IDP workflow*

The models are generated using FRAGFOLD (described in 2.1.3). For each sequence in the dataset (section 2.2) FRAGFOLD generates a desired number of structures (section 2.3.2). These structures correspond to the energetic minima found by FRAGFOLD during the simulation. Because of the complexity of the protein energy landscape determined by the available conformational space, some models may not correspond to any of the structures found in the experimental NMR ensemble, i.e. false predictions. Therefore, all of FRAGFOLD-generated structures for a given sequence constitute its raw ensemble, which should contain the desired experimental ensemble (Figure 6). This raw ensemble needs to be processed to extract what would be a final ensemble and the result of the method. The process of obtaining the final ensemble from the initial set of FRAGFOLD-generated structures is called ensemble extraction (section 2.3.3). The final step of the methodology is to compare the final FRAGFOLD ensemble to its experimental counterpart – ensemble comparison (section 2.3.4) – consisting of superposition of individual structures within an ensemble (section 2.3.4.1) and scoring the agreement between FRAGFOLD and NMR ensembles (sections 2.3.4.2 and 2.3.4.3). The whole process is outlined in the diagram below (Figure 6).

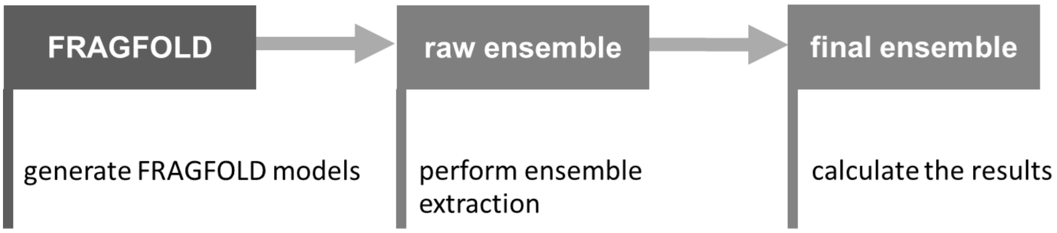


Figure 6. Schematic workflow used in FRAGFOLD-IDP method.

2.3.2. *Ensemble generation*

All FRAGFOLD simulations were ran using the all-atom representation, Replica Exchange Monte Carlo (REMC) to search for low energy conformations and relative weighting of the energy function terms determined by considering the standard deviations of each term across an ensemble random chain conformations for the target, as described by Jones *et al.* (Jones et al., 2005). The total number of annealing steps was set to 10,000,000 steps per simulation. An ensemble of 200 models per protein was generated to ensure reasonable sampling of conformational space.

For each sequence in the dataset, secondary structure predictions were generated using PSIPRED (Jones, 1999) and HHblits was used to generate the input multiple sequence alignments (Remmert et al., 2011). Standard HHblits parameters were used – 3 iterations, E-value threshold of 10^{-3}, minimum sequence coverage of 50% and minimum sequence identity to query of 30%.

FRAGFOLD takes advantage of a set of parameters (as described in 2.1.3). During method’s optimisation alternative sets of potentials were tested:

(1) All potentials (as in the case of globular proteins; referred to as ALL);

(2) Removing STERIC potential (STERIC);

(3) Removing COMPACTNESS potential (COMPACT);

(4) Using all potentials, but excluding secondary structure predictions from the input (noSS).

2.3.3. ***Ensemble extraction***

Each FRAGFOLD run generates a requested number of models per target sequence. The models correspond to minimized structures according to FRAGFOLD potential terms, and represent various minima across the free energy landscape. Therefore, to obtain a representative and (presumably) correct ensemble corresponding to the native disordered ensemble, it is necessary to extract a set of related structures. There are several issues that need to be dealt with in order to achieve this – choice and optimisation of a clustering method (structural clustering or other approach to clustering), determination of the optimal features to be clustered and the size of the ensemble. The methods to tackle each of these issues are described below.

2.3.3.1. Structural clustering

Structural clustering is an essential technique in structural bioinformatics. The name structural is used because some form of structural similarity between models is used as a metric in the feature space. Structural clustering is typically used as a final step in many modern template-free protein modelling pipelines (Jones and McGuffin, 2003; Shortle et al., 1998). Structure prediction algorithms (e.g. FRAGFOLD) first generate a large number of possible models (decoys) that correspond to minimized structures. Then, to account for entropic contributions a structural clustering method is employed to group related structures and find patterns in structural preferences. In an ideal case, the largest cluster would correspond to the correct structure – most of the structures are minimized around the global energy minimum (Shortle et al., 1998). Here, based on a similar predicate – the correct disordered ensemble should lay in a broad global minimum of the free energy – several structural clustering methods were evaluated.

2.3.3.2. Hierarchical clustering

In the most popular clustering algorithm – *K*-means clustering – the number of clusters to which the algorithm converges needs to be pre-defined. Therefore, this approach is not suitable for problems where the structure of the data is unknown in advance. More suitable for this task is hierarchical clustering, where a criterion on the relationships between and/or within the clusters can be imposed. In hierarchical clustering, the initial set of clusters is composed of pairs of nearest neighbours according to the defined distance metric. Then, iteratively, the clusters are merged and populated with more distant neighbours. This process takes place until some cut-off value for the distance between elements in the cluster is reached, or until all clustered elements form a single cluster and a hierarchical tree is formed.

Two in-house clustering methods were used, basing on either RMSD or TM-score (described in section 2.3.4.4) as a distance metric – RMSDclust and TMclust (Jones and McGuffin, 2003).

2.3.3.3. SPICKER

SPICKER is a clustering method tailored to protein structure prediction problem of selecting near-native folds from a library of decoys (Zhang and Skolnick, 2004a). The method heuristically accounts for the energy landscape and combines clustering with model evaluation to iteratively search for an optimal pair-wise RMSD values in order to output the most representative structures from the library of decoys. SPICKER automatically outputs a single set of best decoys using a combination of pair-wise self-adjusting RMSD assessment, neighbour clustering and filtering.

SPICKER is used in some of the current top-performing protein modelling pipelines, such as I-TASSER (Roy et al., 2010), QUARK (Xu and Zhang, 2012), or in EvoDeisgn *de novo* protein design (Mitra et al., 2013).

2.3.3.4. MaxCluster

MaxCluster is a versatile clustering program that enhances the repertoire of more traditional clustering methods than SPICKER. It contains a variety of hierarchical and nearest-neighbour clustering algorithms (`http://www.sbg.bio.ic.ac.uk/~maxcluster`). A total of 15 variations of parameters were tested, modifying clustering method (hierarchical clustering: single linkage, average linkage, maximum linkage; and two neighbour pairs algorithms) and RMSD of the initial clustering thresholds (4, 5, 8 Å).

2.3.3.5. PFClust

The final approach tested was parameter-free clustering (Mavridis et al., 2013; Musayeva et al., 2014). PFClust is a partitional algorithm that aims to tackle the problem of the lack of unique definition of optimal number of clusters. Instead, based on the similarity matrix provided, it attempts to automatically determine an optimal number of clusters the data should be split into.

The method is based on the idea that each cluster can be represented as a non-predetermined distribution of the intra-cluster similarities of its members (Mavridis et al., 2013). The clustering criterion PFClust uses is the expectation value of the similarity distribution of its cluster members. Because the distribution is determined by all possible clusterings, which in turn are determined by the dataset size, PFClust uses random sampling to determine initial parameters. From this, the method performs threshold selection and proper clustering. Because the process is stochastic, determined by the initial randomization, both randomization and calculation steps are repeated until convergence.

Thresholds are selected from the most significant values of the distribution of expectation values calculated from the distribution of intra-cluster similarities. Proper clustering is then performed based on those values by an agglomerative algorithm and validated based on the Silhouette width of obtained clusterings. Finally, runs starting from different randomizations are compared and if they converge according to Rand Index criterion the clustering is finished (Rand, 1971). Otherwise the lowest scoring (Silhouette width) clustering is discarded and the process is repeated.

Although PFClust is heuristic in the sense that it does not optimise any simple metric, the method is robust. It is well-suited to the problem studied here, where it is difficult to pre-define *a priori* clustering parameters, such as the number of desired clusters, number of cluster members or distance metric cut-off for intra- or inter-cluster similarity.

In this work, RMSD was used as the distance metric for PFClust.

2.3.4. *Ensemble comparison*

2.3.4.1. Structural superposition

Clustering methods rely on some global similarity or distance metrics to generate the final ensemble. Having obtained the final ensemble it is necessary to superpose all structures within each of the ensembles – the extracted final FRAGFOLD-IDP ensemble and experimental (NMR) ensemble – in the same way, as a necessary step to calculate per-residue backbone dynamics within each of the analysed ensembles.

A set of possible superposition algorithms was evaluated. Two global superposition methods: a least-squares method – ProFit (`http://www.bioinf.org.uk/software/profit/`) and a maximum-likelihood method – Theseus (Theobald and Steindel, 2012). Each of these methods uses a different approach to structural alignment, hence estimating different variations of residues along the sequence. Additionally, a sliding window superposition method, as an alternative to published and freely available methods, was implemented and tested.

a. Least squares fitting – ProFit

ProFit performs least squares superposition using the McLachlan algorithm (McLachlan, 1982). Optimal superposition of two proteins is determined by minimizing the sum of the squared distances between all of the pre-defined atoms (Cα trace). The McLachlan algorithm provides an efficient mean to do so, exploiting the conjugate gradients method. Therefore, ProFit tries to minimize all inter-residue distances at once not accounting for any protein-like features. In the case of heterogeneous structures (having high RMSD), results generated by ProFit tend to have variations uniformly distributed along the superposed structures (compare Figure 7). However, when considering multiple structures to be aligned, ProFit resorts to a simplification that increases the algorithm's efficiency. In each step of multiple structural superposition, ProFit averages previously superposed structures and fits the considered structure to the average. This implies that the order of the superposition may impact the end result.

Considering overall RMSD across the whole structure, this has a small effect, but considering per-residue results, the results can vary substantially over complete enumeration of all possible structures, especially in the case of highly variable regions (see sub-section 2.3.4.1.c below).

b. Maximum likelihood fitting – Theseus

Theseus performs maximum likelihood superposition which weighs more variable regions in structures differently to the more conserved ones (Theobald and Steindel, 2012; Theobald and Wuttke, 2008). This results in more tightly aligned rigid regions and more variable mobile and terminal regions (what shows agreement with principal component analysis; Figure 7). The method also enables the superposition of multiple structures onto each other simultaneously, unlike the sequential approach in ProFit.

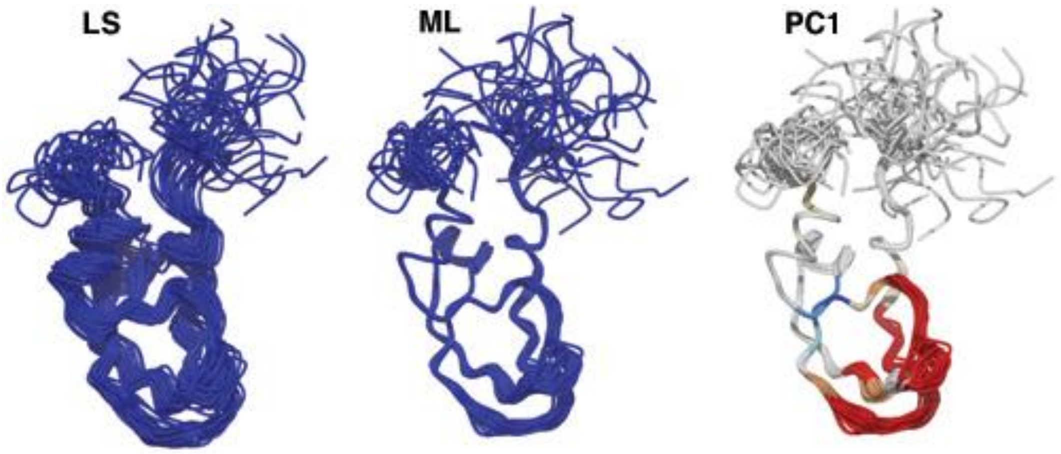


Figure 7. Comparison of least-squares (LS), maximum-likelihood (ML) superposition and principal component analysis (PC1). Figure reproduced from `http://www.theseus3d.org/`. Superposed protein is Kunitz domain from PDB 1ADZ.

c. Generalized least squares calculations

Instead of the approach to least squares fitting presented in ProFit, an analytical version of the fitting using a generalized form of RMSD could be considered. It can be expressed as:

$$RMSD = \sqrt{\frac{2}{l \cdot n \cdot (n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \sum_{k=1}^{l} \left\| p_{ik} - p_{jk} \right\|^2}$$

where *l* is the length of the studied structure and *n* is the number of superposed structures. Let us define δ as

$$\delta_{i,ab} = \sqrt{\left(a_{ix} - b_{ix}\right)^2 + \left(a_{iy} - b_{iy}\right)^2 + \left(a_{iz} - b_{iz}\right)^2} \equiv RMSD(a_i, b_i)$$

then:

$$RMSD = \sqrt{\frac{2}{l \cdot n \cdot (n-1)} \sum_{i \neq j}^{n} \sum_{k=1}^{l} \delta_{k,ij}^2}$$

δ in these equations represents per-residue RMSD. This approach is more computationally expensive, as it requires calculation RMSD over all pairs of structures (or residues) in an ensemble of structures, but it ensures the results are robust regardless of the order of superposition, or the flexibility of considered structures.

d. Sliding-window superposition

As an alternative to the global superposition approaches discussed in points (a) and (b), sliding-window superposition was used to complement them. In this approach, only a selected subset of residues (within the window) is superposed at a time. The window progresses along the sequence, residue by residue, until it reaches the end of the sequence. The superpositions were performed using the generalized least squares fitting described above. This approach enables to minimize the impact of the quality of structure prediction, separating predictions of per-residue fluctuations (disorder) and predictions of the overall structure (which can be assessed using TM-score; section 2.3.4.4). Alternatively, sliding window superposition can be interpreted as a method to remove the effects of rigid body motions from the comparisons.

If a residue, or a fragment of the structure, is highly flexible (disordered) it will not superpose well to its neighbours. In contrast, when superposing the entire structure at a time using a global superposition method, misplaced fragments (poorly predicted regions) can make the RMSD profile gain a background signal propagating along the sequence (example Figure 16 residues 1-40). In other words, thanks to applying the sliding window superposition it is possible to reduce the rate of residues erroneously assigned as disordered.

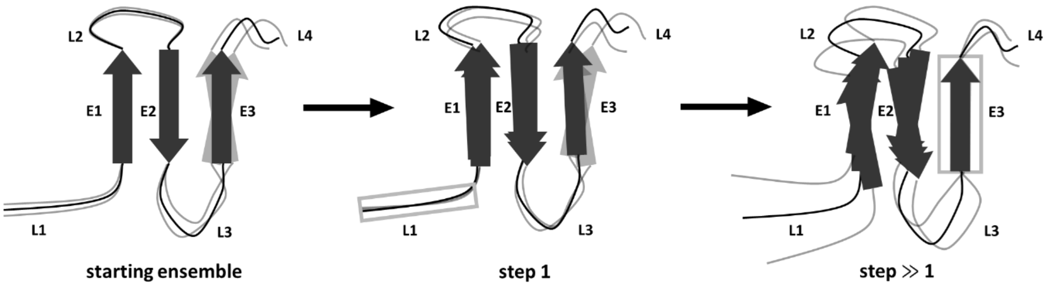


Figure 8. A toy example of sliding window superposition procedure. Loops in the protein structure (L1 to L4) are shown as black and grey lines; anti-parallel *β*-sheets (E1 to E3) are shown as red arrows. The sliding window is depicted as a green rectangle in middle and right panels. Global superposition (starting) ensemble is shown on the left. The first step of the sliding window superposition procedure (middle panel) superposes the L1 loop removing the flexibility visible in the starting ensemble. Similarly, as the window proceeds further down the sequence (right panel), E3 flexible sheet becomes well-aligned.

To make this description more tangible, let us consider a schematic example presented in Figure 8. The starting ensemble (left panel) shows a possible outcome of a global superposition method (least squares or maximum likelihood superposition). Generating a per-residue RMSD profile basing on that superposition, one would assume that L1 is moderately disordered; E1, L2 and E2 are ordered, while L3, E3 and L4 are disordered. However, by applying the sliding window superposition it becomes clear that L1 in fact oscillates around an equilibrium conformation and therefore is not disordered (middle panel). Similarly, E3 is not disordered (right panel), but the flexibility observed in the starting ensemble (left) might have been caused by: an adjacent disordered region L3 due to rigid body motions, the lack of stabilizing contacts on the exterior of the sheet, or if it was a model, an error in structure prediction. Doing a systematic sliding window analysis of this example, one would see that only L3 and L4 are disordered, i.e. these regions would not align across the ensemble, as there is no equilibrium conformation.

A window size needs to be determined in the sliding window methodology. Let us consider boundary conditions first. If a window size is equal to the length of the sequence, then the alignment becomes effectively a least squares fit of the whole structure (because a generalized least squares superposition is used to calculate per-residue fluctuations). Whereas, if the window size is 1, then each residue is perfectly superposed across the ensemble (considering only the main-chain) and there are no fluctuations along the whole chain in each of the ensembles. As may be expected, the relation of the window size to the disorder match metric (described in section 2.3.4.3) is largely monotonic (Figure 9). Therefore, simply maximizing the fit of the window size basing on comparisons between predicted and experimental ensembles is pointless. As it turns out, it is possible to find an optimal sliding window size based on the sequence and structural information content (Šikić et al., 2009). Šikić et al. concentrate their efforts on finding a sequence-based method for the prediction of protein binding sites. Nevertheless, they perform a systematic survey of a sliding window approach and find that window sizes between 7 and 10 residues possess the highest amount of structural information content per sequence length. Also, there is some insight from older surveys concentrating on the preferred lengths of secondary structure elements in proteins (Penel et al., 2003, 1999). Penel et al. show that a typical length of a β-strand is 6 residues, whereas a typical α-helix spans 12 residues. Over 90% of strands and 50% of helices analysed by Penel et al. fall into a 10-residue threshold. Basing on these studies, it was found it to be a reasonable compromise, which would not diminish the sensitivity of the sliding window superposition, to select a window size of 10 residues.

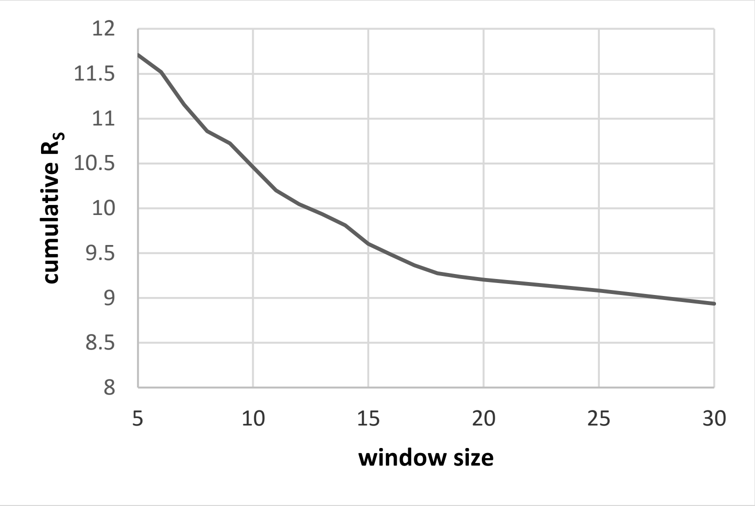


Figure 9. Comparison of sliding window size versus the fit of predicted ensembles to NMR ensembles. For the purpose of the comparison a benchmark set was used (section 2.2.3). “cumulative” refers to the cumulative value of the predictions made on the benchmark set.

2.3.4.2. Per-residue RMSD profile

Starting from any of the structural superposition methods described in section 2.3.4.1, it is feasible to obtain a per-residue RMSD profile. All three analysed methods make it possible to obtain RMSD values for individual residues along the sequence, regardless if the methods perform global (ProFit, Theseus) or local superposition (sliding window). The profiles give information about the degree of fluctuations of individual residues along the protein chain. It is a direct measure of protein backbone flexibility and can be interpreted as the degree of intrinsic disorder within protein regions. The relationship between per-residue RMSD and disorder/order classification is studied in Chapter 3.

Because the per-residue RMSD profile measures the degree of protein disorder, it will be also referred to as disorder profile throughout this work.

2.3.4.3. Disorder match metric – Spearman’s rank correlation

To compare how well the generated ensemble matches its experimental counterpart, a goodness of fit metric is necessary. In this research, disorder profiles are considered, therefore it reasonable to compare them rather than the structures explicitly. Also, the exact flexibility values along the chain are not as important as their relative values within an ensemble. Looking at the relative intensities can account for insufficient sampling of ensembles on FRAGFOLD side, or on the experimental side for inaccuracies emerging from under-determined NMR ensembles (refer to section 1.3.4.5). For these reasons Spearman's rank correlation coefficient (R_S) was selected as the most appropriate metric for the problem. The correlation determines if the order of corresponding data points is consistent between the sets. The metric was already used in similar cases comparing a coarse-grained method (CABSflex) with MD and NMR results (section 1.5.2.3; Jamroz et al., 2014, 2013a, 2013b).

2.3.4.4. Structural comparison – TM-score

Spearman's rank correlation gives the information about the agreement of the disorder profiles between FRAGFOLD-IDP and NMR ensembles. It does not account for the fold or tertiary structure of the protein. To make structural comparisons, i.e. determine whether the folds of proteins in FRGFOLD-IDP and NMR ensembles are similar and to what degree, a different measure is required. For this purpose, TM-score was used (Zhang and Skolnick, 2004b).

TM-score is a robust, protein length-independent similarity metric that is routinely used in many structure prediction problems (e.g. Kosciolek and Jones, 2014). TM-score is defined as:

$$TM-score=\max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

where L_N is the length of the native structure, L_T is the length of the aligned residues from the template (model) structure, d_i is the Cα-Cα distance of the *i*-th pair of aligned residues and d_0 normalizes the match difference.

The higher the TM-score is (on the scale of 0 to 1), the closer is the predicted structure to its experimentally-solved equivalent. Generally, it is assumed that structures with TM-score 0.5 or higher have a correct fold and can be considered successful predictions (Xu and Zhang, 2010).

For the current problem, because ensemble versus ensemble comparisons are performed, TM-scores need to be calculated differently and accumulated accordingly. For each structure in the predicted ensemble, TM-score was calculated against each of the NMR models included in the PDB file. For each structure in the final FRAGFOLD-IDP ensemble, the highest TM-score was selected. As all of the structures in the predicted ensemble had their TM-scores calculated, the mean TM-score was then computed. This averaging procedure was performed to account for the fact that FRAGFOLD ensembles not necessarily include all of the conformational states included in the NMR ensemble. Also, the NMR ensemble may not include all of the naturally occurring conformations, as it itself rather represents one of the sets of conformations that fit the experimental data. This problem was discussed in the Introduction, section 1.3.4.5.

2.3.5. *Borderline results extraction*

Borderline results are useful to assess the effectiveness of the developed method. In case of FRAGFOLD-IDP there are several borderline cases that could be considered. A baseline method – needed to assess whether and to what extent the predictions made using FRAGFOLD-IDP improve over a simplistic approach that would not require any simulations. An upper-bound method – to assess how far from an optimal solution within a given framework the predictions are.

One baseline method was constructed (described in 2.3.5.1) and one approach to extract top-scoring ensembles – based on extraction of the top result from a set of randomly generated ensembles (section 2.3.5.2).

2.3.5.1. Baseline method – naïve disorder assignment

A naïve assumption made for the problem of predicting protein backbone dynamics is that all loops are disordered, while all helices and sheets are ordered. In real proteins, the helices are most rigid because of local hydrogen-bonding patterns (i+4 → i in α-helix), whereas sheets allow for some degree of flexibility due to possible bending and twisting motions between individual strands. Secondary structure predictions were used for this purpose, because of the assumption that only the protein sequence is known. The secondary structure predictions were carried out using PSIPRED (Jones, 1999).

Since Spearman's rank correlation (R_S) is used as the disorder match metric (section 2.3.4.3), the order of fluctuations along the backbone is important, but not their absolute values. Therefore, following the rationale presented in the previous paragraph, an arbitrary set of values was assigned to each predicted secondary structure element along the protein chain: 2 for loops (C), 1 for sheets (E) and 0 for helices (H). To verify if the assumptions were correct other variants were also tested, e.g. where sheets and helices were assumed to have equal flexibility, or where helices were more flexible than loops (Table 6). The initial assumptions proved to be correct and robust (i.e. produced the highest cumulative and mean R_S values on the benchmark set compared to experimental ensembles) regardless of the assigned flexibility value (Table 6).

Table 6. Naïve predictions optimisation.

secondary structure type	**per-residue RMSD values**								
H	2	0	1	1	**0**	1	0	0	1
E	1	2	1	2	**1**	0	1	0	0
C	0	1	0	0	**2**	2	1	1	1
mean R_S	**-0.35**	**0.06**	**-0.34**	**-0.31**	**0.35**	**0.31**	**0.24**	**0.34**	**0.18**
cumulative R_S	**-9.67**	**1.79**	**-9.62**	**-8.67**	**9.67**	**8.67**	**6.34**	**9.62**	**4.03**

The table compares alternative helix (H), sheet (E) and coil (C) per-residue RMSD values to find an optimal set of parameters for the naïve approach.

2.3.5.2. Random ensemble generation

To assess the ensembles from the complete dataset, 1,000 random ensembles of 10 structures for each protein in the dataset were generated. It is a reasonable compromise between effective sampling of possible ensembles and computational efficiency.

Since the random ensembles can serve as a proxy to the top scoring ensemble in terms of the possible per protein performance, top R_S and median R_S for each target was calculated to be used as an indicator of FRAGFOLD-IDP performance (discussed in section 2.5.4).

2.3.6. *Relationship between NMR order parameter (S^2) and per-residue RMSD*

Order parameter S^2 is an NMR experimental value, therefore it is necessary to understand its relation to per-residue RMSD in order to make direct comparisons to NMR ensembles deposited in the PDB, or generated by the method developed in this work. S^2 is only sensitive to angular motions and not translations, because it is calculated from the second-order spherical harmonics (Brueschweiler and Wright, 1994). Formally, order parameter can be defined as:

$$S^2 = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \left\langle Y_{2,m}(\Omega) \right\rangle \right|^2$$

where $Y_{2,m}(\Omega)$ are the second order spherical harmonics, and Ω describes the orientation of the N-H bond vector (Lipari and Szabo, 1982). Hence, high values of the order parameter correspond to rigid structures, while low S^2 values represent flexible regions. In borderline cases, when $S^2 = 0$ the orientation of the bond vector is completely isotropic, i.e. the residue is completely disordered and in case when $S^2 = 1$ the orientation of the bond vector is fixed, i.e. the residue is completely rigid.

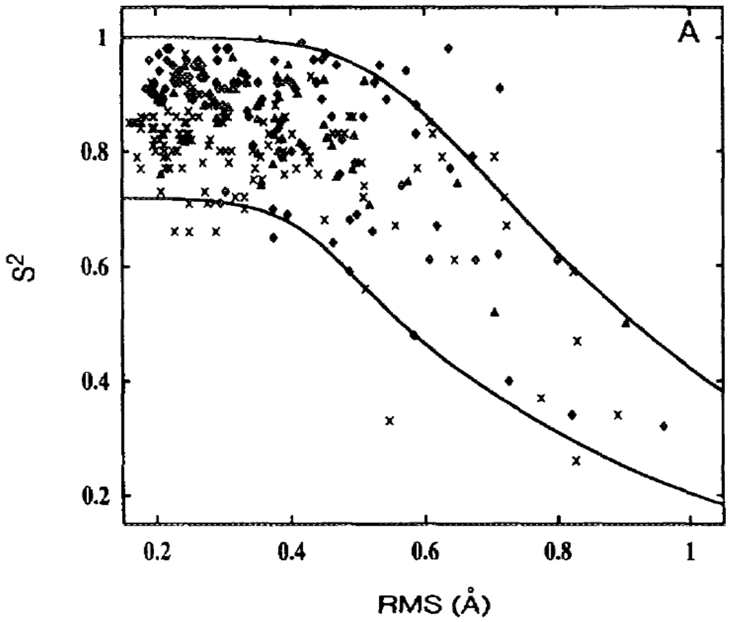


Figure 10. Relation between order parameter S^2 and backbone RMSD. Figure reproduced from Powers et al., 1993.

Because S^2 is an experimental parameter, it has no analytical correspondence with per-residue RMSD (Figure 10; Powers et al., 1993). This somewhat hinders direct comparison between other methods (e.g. the method developed in this work) and methods producing order parameter values as an output (e.g. DynaMine described in 1.5.2.4). But since the main assessment metric in this work is Spearman's rank correlation, then as long as the relationship between RMSD and S^2 is monotonic the conclusions would not be affected. Simple inversion the of the S^2 function (i.e. $1\text{-}S^2$, the higher the value the more disordered the residue) enables a robust comparison with per-residue RMSD values in terms of R_S.

Another possibility is to consider the methods that try to estimate the order parameter values from other quantities. An example can be Random Coil Index method (Berjanskii and Wishart, 2006). RCI uses chemical shift data to calculate either S^2 values, or RMSF (root mean square fluctuations). The authors of the method give analytical expressions how to achieve both of these values from RCI and do a comparison showing a correlation between $1\text{-}S^2$ and RMSF calculations. In RCI method RMSF are proportional to RCI values, whereas order parameter is given by:

$$S^2 = 1 - 0.5\ln(1 + RCI * 10)$$

The relationship is therefore again monotonic and given by a negative logarithm.

In conclusion, since $1\text{-}S^2$ is monotonically related to RMSD, these quantities can be directly compared using R_S as a metric.

2.4. FRAGFOLD-IDP optimisation

The FRAGFOLD-IDP workflow involves 3 steps: generating raw ensemble, final ensemble extraction and assessment of the results (Figure 6). All those steps need to be optimised in order to maximize the performance of the method. Although the workflow is sequential, each of the steps of the method is inter-dependent on each other. For example, the features of the initial raw ensemble (determined by FRAGFOLD parameters) may impact the conclusions about which comparison method is optimal. For this reason, the optimisation of FRAGFOLD-IDP was performed iteratively, until adequate approaches and parameters for each of the steps were found. In this section however, the process is described in a sequential manner in the same order as FRAGFOLD-IDP proceeds.

To assess the performance of FRAGFOLD-IDP, three aspects need to be optimised:

(1) FRAGFOLD parameters;

(2) ensemble extraction method;

(3) structural superposition method.

As mentioned previously, in section 2.3.4.3, FRAGFOLD-IDP results are assessed on the basis of Spearman's rank correlation (R_S) values between the per-residue RMSD profiles predicted by FRAGFOLD-IDP and corresponding NMR ensemble.

All calculations in this section were performed on the benchmark dataset of 28 proteins extracted from the full 200 protein NMR PDB dataset (section 2.2.3), unless noted otherwise.

2.4.1. *FRAGFOLD folding parameters*

FRAGFOLD was designed to effectively predict protein structures of globular proteins (Jones and McGuffin, 2003; Jones et al., 2005; Kosciolek and Jones, 2014). The parameters embodied in FRAGFOLD were developed and tested on sets of highly resolved globular proteins using an inverse Boltzmann approach. Therefore, it is possible that FRAGFOLD is biased towards more globular and compact structures and the way it was initially designed might not be optimal for generating the ensembles of intrinsically disordered proteins.

To verify this hypothesis and find an optimal set of FRAGFOLD parameters for generating ensembles of intrinsically disordered proteins, a set of alternatives was tested. An initial set of calculations on the benchmark set was ran with a set of parameters typical for globular protein targets (section 2.1.3) – ALL parameters. Further runs involved different combinations of parameters: without the compactness potential, but with all other potentials – COMPACT; without the steric potential – STERIC. FRAGFOLD also utilizes secondary structure predictions that aid the selection of fragments during the simulation. The runs without secondary structure predictions included in the input were also carried out – noSS.

The choice to explore FRAGFOLD simulations without steric or compactness terms was dictated by the fact that intrinsically disordered proteins exhibit larger than expected radii of gyration given their length. The compactness term in FRAGFOLD was explicitly developed to ensure the globular fold of the simulated models (Jones, 2001). A similar case concerns the steric terms. Although they ensure that no conformations involving steric clashes are permitted, the way that FRAGFOLD generates new conformations, permitting sterically unfavourable conformations, could allow for less conservative sampling. Finally, eliminating secondary structure predictions (noSS) was dictated by a possible bias from predicted secondary structure elements overestimating the amount of order within predicted structures.

Results of the calculations are gathered below (Table 7). 'Good' predictions ($R_S \geq 0.5$) are shown in bold. Outstanding results ($R_S > 0.7$) are additionally highlighted in green. The column best result shows which approaches (ALL, COMPACT, STERIC, noSS) produced top scoring results in terms of R_S. All approaches that achieved results within 10% of the highest score are included. If more than one approach is listed in best result column, the R_S value (best) is presented for the highest scoring approach (listed first).

Table 7. Results from the initial studies using FRAGFOLD-IDP on the benchmark set.

protein	length	no. disordered residues	best result top10%	TM-score	R_S value (best)
1b75	94	16	A + C	0.71	0.53
1dmo	148	58	A + S	0.45	0.62
1dvd	98	11	S + N	0.51	0.67
1fkr	107	12	A + C	0.85	0.19
1jw3	140	34	S + C	0.37	0.41
1k19	112	38	A	0.36	0.71
1ni7	149	15	A + S	0.36	0.41
1nin	105	37	A	0.41	0.51
1nnv	107	18	S + A + N	0.31	0.40
1nti	86	13	S	0.39	0.46
1ovq	138	41	A	0.32	0.53
1soy	106	10	N + C	0.63	0.61
1tac	86	67	N	0.18	0.23
1tiv	86	86	S + N	0.19	0.44
1xhs	113	29	N	0.25	0.53
1xpw	143	17	A	0.33	0.32
1y6d	114	11	A	0.43	0.27
1zza	90	44	A	0.21	0.26
2aqa	57	57	A + N + C	0.23	0.51
2fki	118	27	S	0.42	0.47
2hgk	117	27	A	0.45	0.75
2jo6	110	11	C + S + A	0.36	0.30
2ju4	87	86	A	0.22	0.37
2k36	149	25	S	0.25	0.60
2k5t	128	13	N + A	0.39	0.46
2kkj	59	29	A + N	0.36	0.69
2kx4	117	41	N	0.29	0.28
2py1	129	10	S	0.63	0.15

best result column: A (ALL potentials); C (no COMPACTNESS potential); S (no STERIC potential); N (no secondary structure predictions). The order in the best result column corresponds to the order of R_S values. Clustering was performed using RMSDclust.

Overall, 12 out of 28 predictions should be considered as ‘good’. All of the tested approaches contribute to this set (ALL, COMPACT, STERIC and noSS). Interestingly, one of the fully disordered proteins (2AQA) is among the well-predicted targets. Since all FRAGFOLD parameters were optimised for globular proteins it was expected, that highly disordered proteins should be difficult targets for the method.

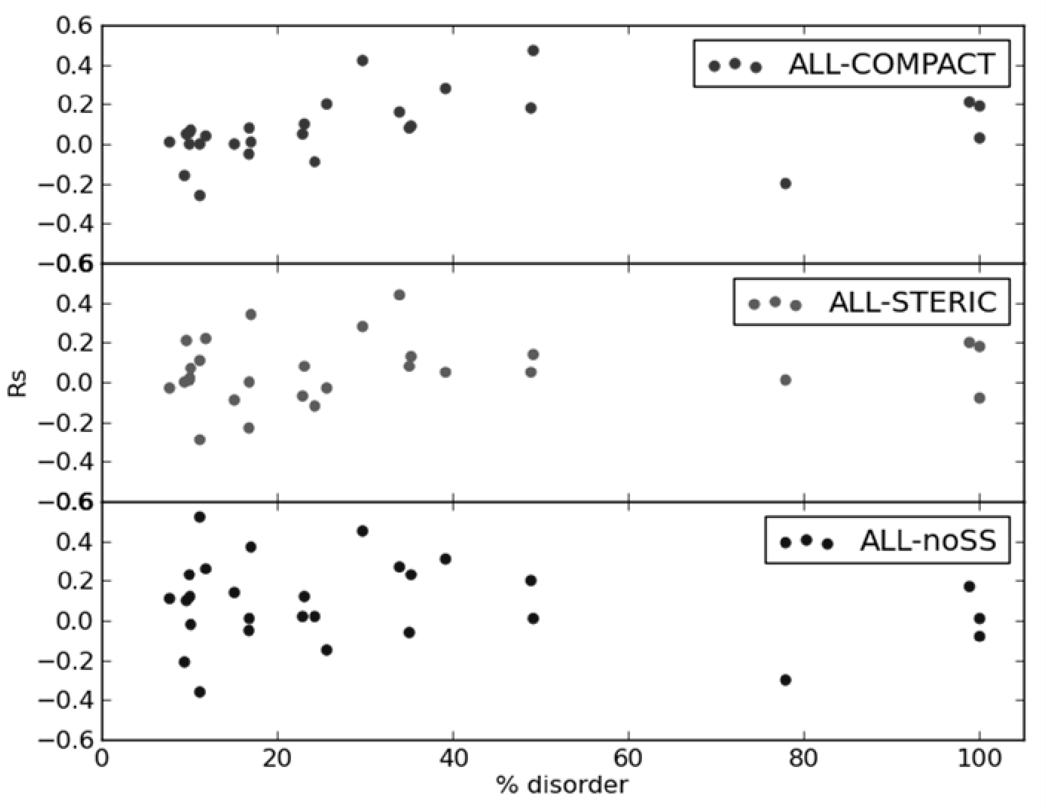


Figure 11. Differences in FRAGFOLD performance using different sets of potentials.

Overall, the benchmark set results do not indicate that any other combination of parameters other than using ALL parameters consistently improves the predictions. From Table 7 17 out of 44 predictions in best results come from ALL potentials (6 from COMPACT, 11 from STERIC, 10 from noSS). There are no other indications that would point to cases where using other combinations of potentials would yield better results (e.g. disorder content; Figure 11). Also, considering the distribution of R_S values, ALL parameters perform best on average, although not significantly (Figure 12).

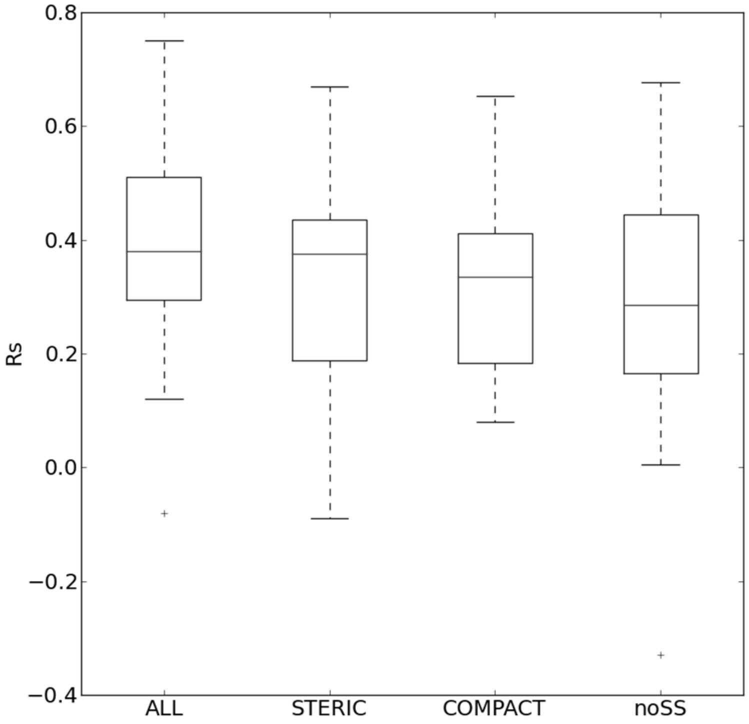


Figure 12. Box-and-whiskers plot comparing the distribution of R_S values between different sets of FRAGFOLD parameters. The analysis was performed on the full benchmark set.

In conclusion, using ALL FRAGFOLD parameters is an optimal method for generating the ensembles of intrinsically disordered proteins. There are no clear indicators of the situations where using some other sets of parameters (COMPACT, STERIC, noSS) would be beneficial over ALL parameters.

2.4.2. *Ensemble extraction method*

In order to obtain the final FRAGFOLD-IDP ensemble, it is necessary to perform ensemble extraction which takes the raw ensemble of models generated by FRAGFOLD and limits it to the best set of models, which should correspond to the ensemble of structures inferred by NMR (Figure 6). To achieve this, an optimal clustering algorithm needs to be found and appropriate cluster selection criteria.

The results described below are based on the ensemble generation method described in 2.4.1, i.e. all FRAGFOLD parameters are used and the raw ensemble is composed of 200 models per protein.

2.4.2.1. Clustering algorithm

A natural choice was to evaluate some standard structural clustering algorithms that are commonly used to in protein structure prediction applications – RMSDclust and TMclust, SPICKER and MaxCluster. The algorithms were described in section 2.3.3.

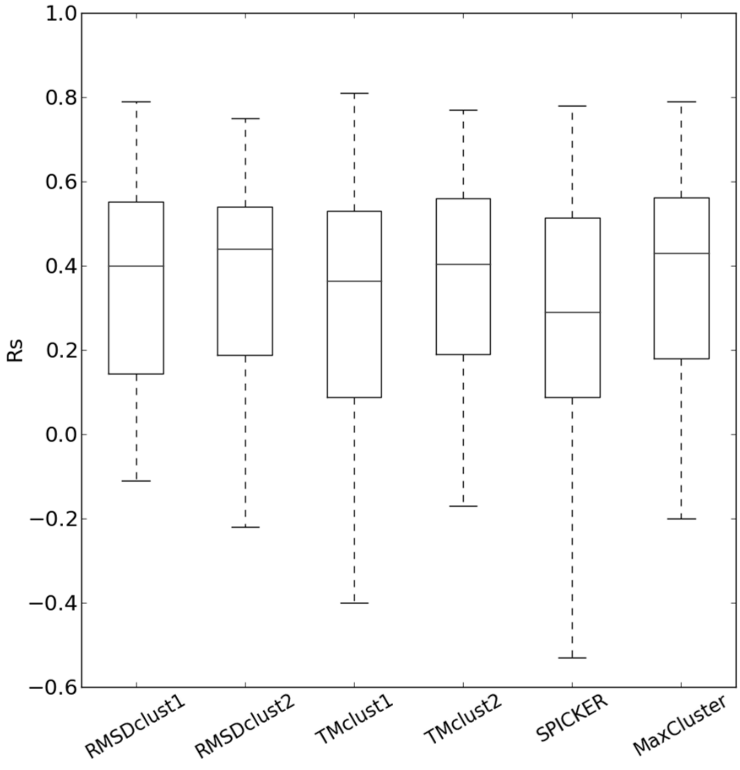


Figure 13. Comparison of standard clustering algorithms.

Following a systematic survey, an evaluation was made to determine how the selection of a clustering algorithm impacts the quality of predictions in terms of R_S. It was found that overall the algorithms perform similarly (Figure 13). SPICKER generally produces inferior results, whereas RMSDclust, TMclust and MaxCluster perform equally well. For clarity, only the best MaxCluster set of parameters is shown here. As an alternative to the distribution of R_S values in Figure 13, a comparison of top results for each target was made. In this approach, the number of methods giving results in the top10% for each target is compared (Table 8). This approach enables a more balanced comparison – if the top R_S value is high (e.g. 0.80), a relatively permissive threshold is applied (results between 0.72 and 0.80) and all high quality predictions are counted. Whereas, if the results are relatively poor (e.g. 0.20), usually only the result of a single clustering approach is counted (results between 0.18 and 0.20). As a result of this evaluation, it was found that RMSDclust produces consistently the best results (Table 8), and was selected as the best classical structural clustering approach. However, it should be noted that none of the approaches, except from clearly under-performing SPICKER, is outstanding and consistently outperforming other methods.

Table 8. Top clustering method performance on the benchmark set.

	RMSDclust1	RMSDclust2	TMclust1	TMclust2	SPICKER	MaxCluster
count top10%	18	15	12	12	7	15

As an alternative to the classical structural clustering methods, PFClust was tested (described in 2.3.3.5). It is an attractive method, because it does not rely on any external parameters except for the distance metric provided (RMSD). Unlike hierarchical clustering approaches, which are based on a distance threshold parameter to separate the clusters, PFClust should be able to adapt to ensembles of different disorder content and effectively extract them without altering any parameters. This argument is especially important, since it is impossible to *a priori* estimate the quality of predictions, or (in real life cases) disorder content and impose parameters on the clustering.

Comparing the top PFClust to RMSDclust results (Figure 14), again it is clear that the methods achieve comparable results, but PFClust shows superior performance comparing both mean (0.42 PFClust and 0.40 RMSDclust) and median (0.48 PFClust and 0.44 RMSDclust) R_S values. Interestingly, differences in the results (PFClust improvement over RMSDclust) are not related to the disorder content of the analysed proteins.

Although RMSDclust and PFClust perform similarly, the latter method does not require any external parameters and is an attractive approach to clustering. Hence, it was selected as the optimal clustering method to extract the final FRAGFOLD-IDP ensembles from raw ensembles of FRAGFOLD models.

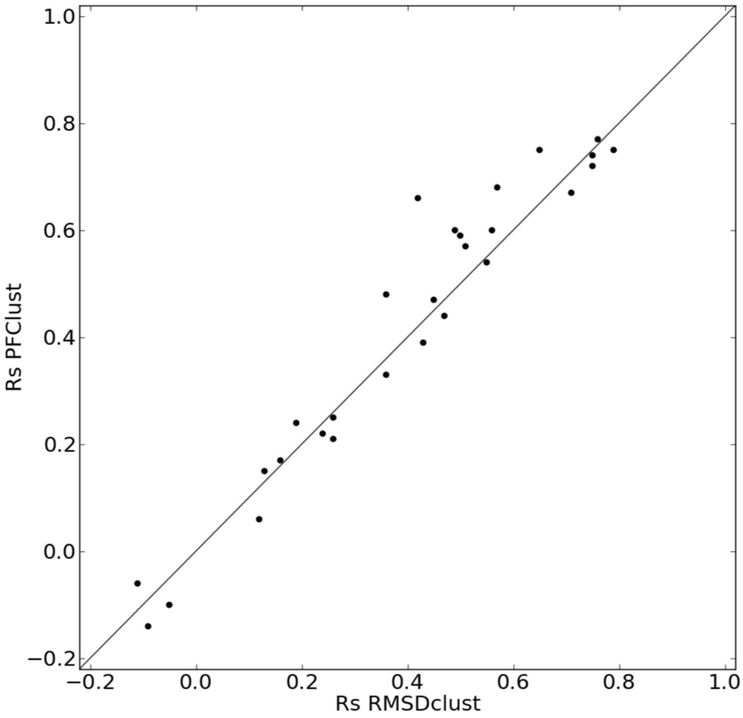


Figure 14. Comparison of RMSDclust and PFClust performance on the benchmark set.

2.4.2.2. Cluster selection criteria

The default and most typical cluster selection criterion is cluster size. The evaluation of the classical structural clustering methods was carried out using cluster size criteria, so was the PFClust evaluation in the previous section.

However, considering the theoretical background behind IDPs (section 1.1), intrinsically disordered protein ensembles should consist of structurally heterogeneous conformations (depending on the amount of disorder) having similar energy (broad energetic minimum of alternative conformational states).

Therefore, to account for both structural heterogeneity and assumed energetic cohesion, several cluster selection criteria were tested:

(1) Cluster size ($|clust|$);

(2) Mean energy ($\langle E \rangle$);

(3) Median energy (median E);

(4) Energy difference (ΔE);

(5) Ratio of mean energy to energy difference (${\langle E \rangle}/{\Delta E}$);

(6) Product of cluster size and mean energy ($|clust| \cdot \langle E \rangle$).

As energy, the final total FRAGFOLD energy of each model was considered (compare section 2.1.3).

For this comparison, instead of the benchmark set, the results were compared on the complete dataset (section 2.2.1). The most computationally expensive step of FRAGFOLD-IDP is generating raw ensembles and since FRAGFOLD folding parameters were established previously (section 2.4.1), it was feasible to perform a large scale comparison of cluster selection criteria.

A comparison of the cluster selection criteria shows that their performance is similar (Figure 15). Overall, the selection criteria including cluster size perform best. Both of those methods achieve highest mean R_S values (Table 9). Overall, 2 (out of 200) results between cluster size alone and cluster size and energy are different – one selection is better using the cluster size and one if both cluster size and energy are considered.

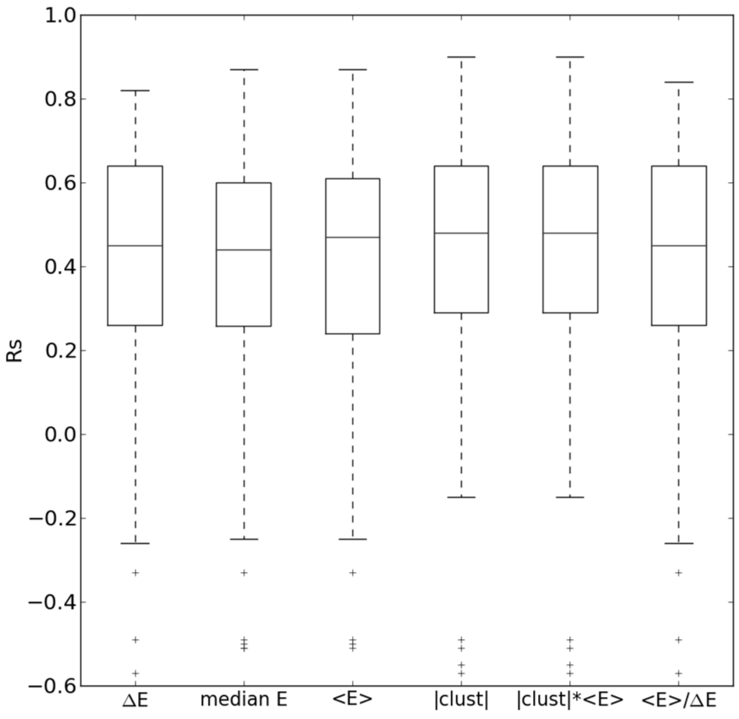


Figure 15. Comparison of cluster selection criteria using PFClust on the whole dataset.

The distributions of the results using all approaches are similar, but there are some notable exceptions. Interestingly, 3 out of 4 outliers (values -0.3 and below; 1G6M, 1K0T, 1XN7) are shared between all cluster selection criteria. One target (1TCP) is significantly better predicted thanks to the use of cluster size criterion (R_S = 0.51 using cluster size; R_S = -0.50 using mean energy) and one target (2K02) is significantly worse (R_S = -0.51 using cluster size; R_S = 0.67 using mean energy). Also, there are 22 cases where using any of the cluster selection criteria yields identical results.

Table 9. Comparison of mean R_S values for different cluster selection criteria.

	ΔE	median E	<E>	|cluster|	|clust|*<E>	<E>/ΔE
mean R_S	0.42	0.40	0.41	0.44	0.44	0.42

The two top methods for cluster selection are cluster size and combining cluster size with mean cluster energy. Since cluster size alone is the most popular cluster selection criterion and simpler of the two (adding mean energy information does not add any significant value), it was selected as the final cluster selection method.

2.4.3. ***Structural superposition method***

Considering structural superposition methods, it was found that Theseus generally produces results consistent with previous disorder annotations (from mobiDB), while ProFit generates more diverse results, that not in all cases stay consistent with Theseus or mobiDB. Nevertheless, both of these methods produce a signal coming from labile, but not disordered regions (e.g. Figure 16). These motions can be either a result of poor FRAGFOLD predictions, or artefacts of rigid body motions that leave a background signal propagating along the sequence (example Figure 16 residues 1-40 for Theseus and 1-50 for ProFit). The third superposition method that was evaluated – using a sliding window – does not suffer from these effects.

Sliding window superposition makes it possible to minimize the effects of rigid body motions and decouple disorder signal from them. In other words, thanks to applying sliding window superposition, it is possible to reduce the rate of residues erroneously assigned as disordered.

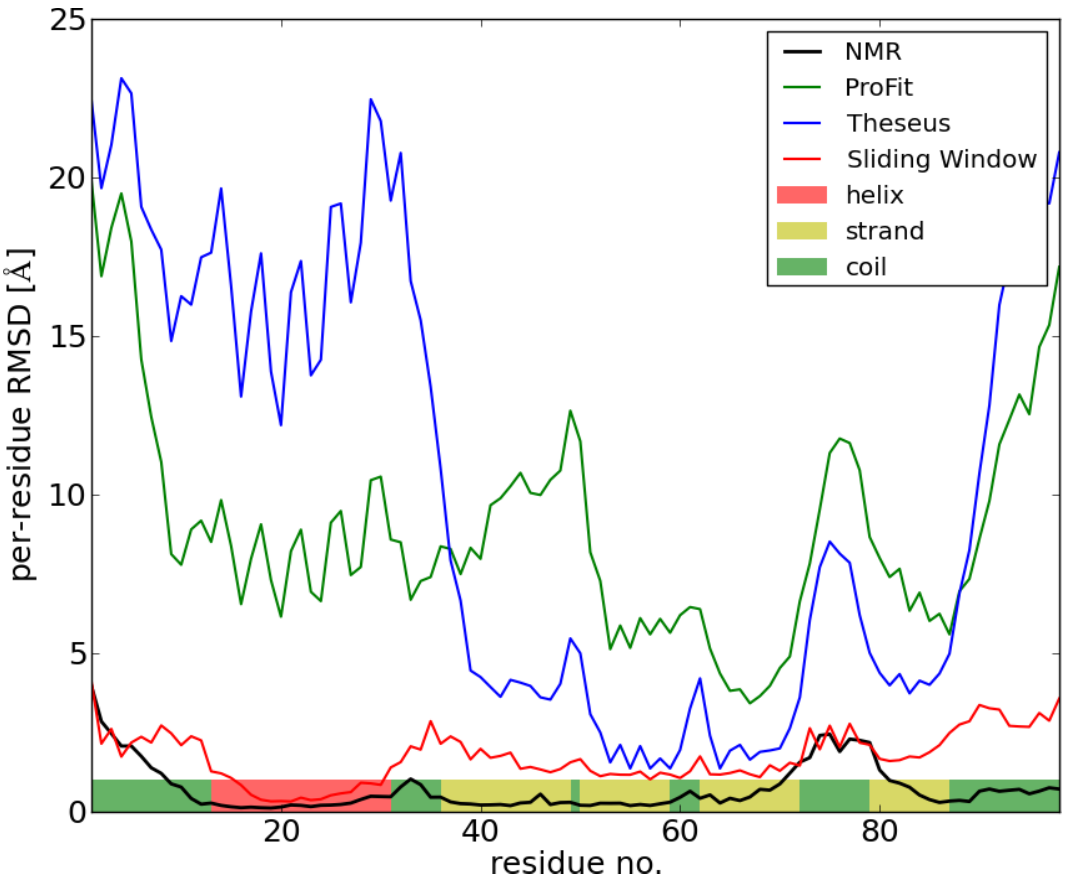


Figure 16. Disorder profile of stefin A (1DVD). NMR ensemble is compared with the same ensemble using alternative structural superposition methods – global superposition (ProFit and Theseus) and sliding window superposition. Predicted secondary structure elements are also highlighted.

Sliding window superposition produces more noisy results in some cases (the results show more local maxima; e.g. residues 70-80 in Figure 16), but because this approach performs multiple local structural superpositions it separates disordered regions from poorly predicted, but not disordered ones.

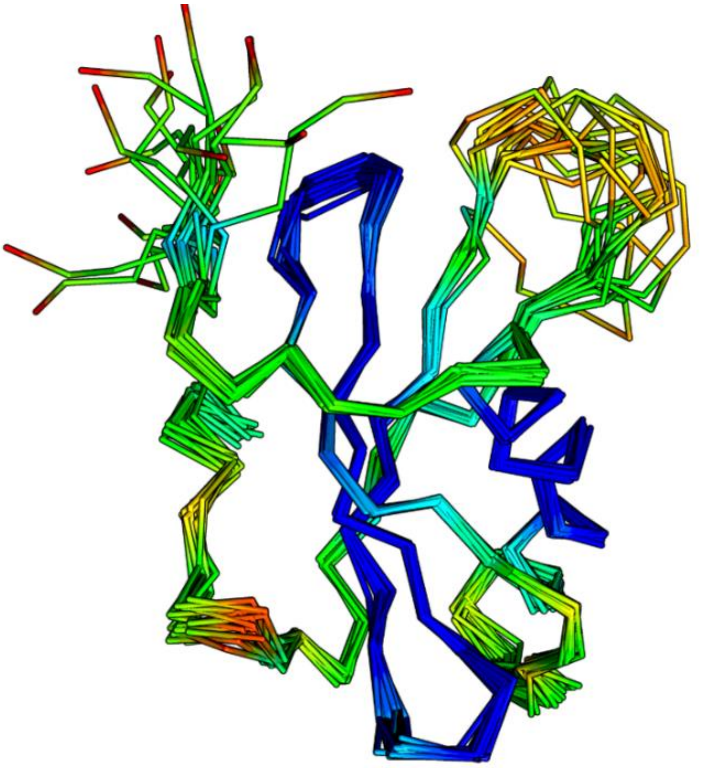


Figure 17. Visualisation of an NMR ensemble and FRAGFOLD predictions of 1DVD (stefin A). The original NMR ensemble deposited in the PDB was split, so that all of the conformations are overlaid onto each other. FRAGFOLD predictions (top-scoring cluster produced by RMSDclust and superposed using sliding window approach) are represented here by a spectrum of colours, where blue represents the most rigid predicted residues, followed by green and yellow. Orange and red represents the most disordered regions according to FRAGFOLD predictions.

In a PDB NMR ensemble of 1DVD (Figure 16) residues 1-10 are disordered, but in the FRAGFOLD-IDP predictions, because of a poorly predicted region, the initial superposition methods (ProFit, Theseus) identify residues 1-40 as disordered. When the sliding window superposition is applied, the method separates disorder from a poor prediction (see Figure 16 and visualisation in Figure 17). This proves that the sliding window disorder profiles are robust and effectively decouple backbone dynamics from overall structure predictions. Because of that, sliding window superposition was selected as an optimal structural superposition method for future analyses.

2.4.4. ***Disorder match metric***

A crucial parameter of the method is the disorder match metric. It should robustly assess alternative ensembles of structures and tell which selection or which set of FRAGFOLD parameters produces qualitatively (comparing relative residue fluctuations along the protein chain) best matching results with an NMR ensemble. For the purpose of my current research, Spearman's rank correlation coefficient (R_S) was found to be the most reliable and works well in terms of distinguishing correct and incorrect predictions. It was also used in other studies concerning the assessment of protein dynamics (Jamroz et al., 2014, 2013b). The metric is described in section 2.3.4.3.

2.4.5. ***Optimal ensemble size***

Given the criteria for an ensemble, it is an arbitrary choice of how large the ensemble should be. The most natural choice is selecting 10 or 20 conformers as these are typical sizes for experimental NMR ensembles. However, to determine how the quality of predictions depends on the size of an extracted ensemble, ensembles of different sizes generated by random ensemble generation were compared on the benchmark set (described in section 2.3.5.2).

Ensembles of 2 to 6 structures are sub-optimal and show inferior R_S values than larger ensembles (Figure 18). 7 to 10 structures in an ensemble show identical quality having the highest mean R_S value. However, since 10 conformers are more typical to experimental NMR ensembles 10 models were selected as a representative size for analyses and comparisons using random generation of ensembles (section 2.3.5.2).

The optimal ensemble size concerns only the ensembles constructed using the random generation (section 2.3.5.2). Whenever clustering results are used (FRAGFOLD-IDP method), all of the structures belonging to the selected cluster are considered as the final ensemble.

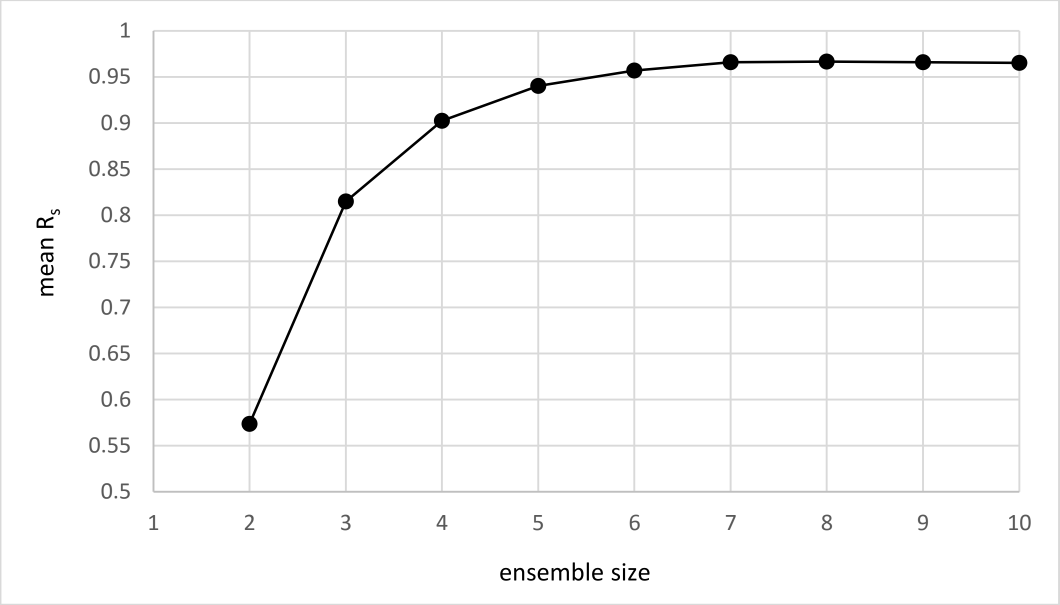
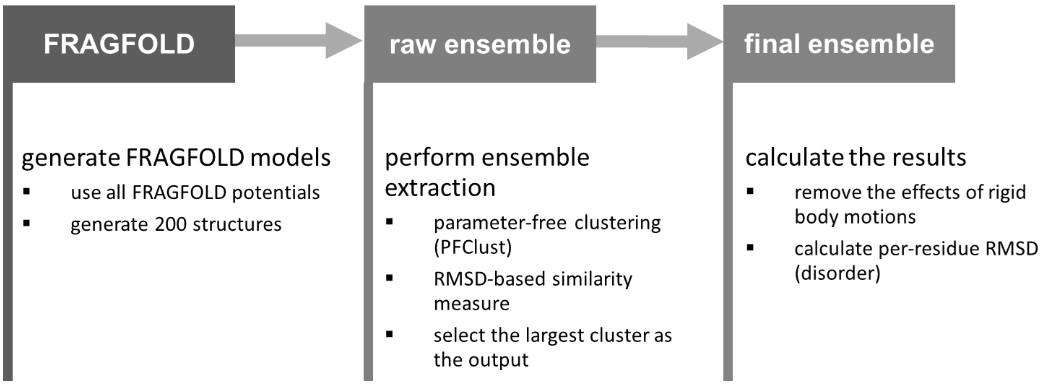


Figure 18. Optimal ensemble size determination. Values presented are calculated on the basis of the benchmark set with mean R_S values calculated using the top-scoring ensembles.

2.4.6. *Final FRAGFOLD-IDP workflow*

All FRAGFOLD-IDP parameters were optimised and a robust workflow was established. It was built upon the initial framework (Figure 6) beginning with FRAGFOLD simulations generating a raw ensemble of models, which are subsequently extracted to form the final ensemble which can be compared to NMR PDB results (Figure 19).



FRAGFOLD

generate FRAGFOLD models
- use all FRAGFOLD potentials
- generate 200 structures

raw ensemble

perform ensemble extraction
- parameter-free clustering (PFClust)
- RMSD-based similarity measure
- select the largest cluster as the output

final ensemble

calculate the results
- remove the effects of rigid body motions
- calculate per-residue RMSD (disorder)

Figure 19. Final FRAGFOLD-IDP workflow.

The final FRAGFOLD-IDP workflow proceeds as follows. For each protein sequence, a multiple sequence alignment and a secondary structure prediction are generated (section 2.4.1). MSA and secondary structure predictions serve as an input to FRAGFOLD, which generates 200 models per input sequence, using all potentials. This set constitutes the raw ensemble (section 2.4.1). The ensemble is then extracted using PFClust with RMSD as the distance metric. The largest cluster is selected as the output (section 2.4.2). The final ensemble is compared to its experimental counterpart by generating a per-residue RMSD profile using a sliding window superposition (section 2.4.3). The disorder profiles are then assessed on the basis of Spearman's rank correlation coefficient (section 2.4.4).

2.5. Full dataset results

In the previous section a set of optimal parameters for FRAGFOLD-IDP and the entire workflow was established. This section describes the results obtained using the optimised FRAGFOLD-IDP on the full dataset (described in section 2.2).

First, an interpretation of R_S values is described (section 2.5.1). R_S is used to quantify the quality of FRAGFOLD-IDP predictions and it is crucial to understand the limitations of this metric and draw the boundaries between high and low quality predictions.

Equipped with an intuition of how to interpret R_S values, I describe the overall performance of FRAGFOLD-IDP (section 2.5.2). Some examples of poor, moderate and outstanding predictions are also discussed in this section to provide an outlook of the results.

All predictions need some baseline method to which they can be compared to. Section 2.5.3 describes the comparison of FRAGFOLD-IDP predictions to a naïve method which bases on the secondary structure predictions and eliminates the need to carry out any FRAGFOLD simulations.

In section 2.5.4, further context is put into the predictions. FRAGFOLD-IDP is compared to a set of randomly generated ensembles. Median and top results from these ensembles are compared to FRAGFOLD-IDP predictions to assess the entire method and to gain insight into the effectiveness of ensemble extraction methodology.

Next, the predictions are put into a structural context and I discuss the impact of protein fold (section 2.5.5) and disorder content (section 2.5.6) on the quality of FRAGFOLD-IDP predictions.

Because of the way FRAGFOLD-IDP is designed, it decouples protein backbone dynamics predictions (by using a sliding window approach) from the predictions of protein structures. Therefore, in section 2.5.7 the relationship between the quality of protein backbone dynamics predictions and the quality of structure predictions is discussed.

Finally, in section 2.5.8 FRAGFOLD-IDP is compared to other methods that either explicitly attempt to predict the protein backbone dynamics from sequence (i.e. DynaMine; described in 1.5.2.4 and 2.1.2.1), or predict qualities that were shown to be related to protein backbone dynamics.

2.5.1. *Interpretation of R_S values*

FRAGFOLD-IDP uses Spearman's rank correlation values (R_S) as a method to score the predictions. It is a reliable comparative metric which works well in a qualitative setting. For example, it is suitable while attempting to find the best solution from a set of alternatives, as it was the case in optimising FRAGFOLD-IDP methodology (section 2.4). However, considering the predictions of protein backbone dynamics R_S values themselves are difficult to interpret, i.e. does $R_S = 0.5$ represent a good prediction? This difference is apparent comparing the interpretation of R_S values to TM-score, which has clear statistical and structural interpretation (compare section 2.3.4.4). TM-score of 0.5 and above is typically interpreted as a good prediction and the two compared proteins share the same fold (Xu and Zhang, 2010).

It is difficult to state the boundary between 'good' and 'bad' predictions for the problem of protein backbone dynamics predictions. In structural classification, there are terms such as class, fold or topology (and respective databases, e.g. CATH (Orengo et al., 1997; Sillitoe et al., 2015) and SCOP (Andreeva et al., 2007; Murzin et al., 1995)). In protein backbone dynamics there is still no such classification, except for descriptive identification of disordered states, such as molten globule, entropic chain, etc. (van der Lee et al., 2014).

Nevertheless, from the visual analysis of the results and from previous studies attempting to predict $1\text{-}S^2$ (NMR order parameter) from NMR ensembles, some intuition can be derived. Zhang & Bruschweiler derived an analytical expression to calculate the order parameter from NMR and X-ray structures (Zhang and

Brüschweiler, 2002).Their method achieves a mean R_S value of 0.61 comparing S^2 values calculated from X-ray structures against experimental S^2 values for 5 proteins and a mean of 0.67 for comparisons with NMR structures on the same set. However the test set is small, the results suggest that R_S values of above 0.6 indicate very good predictions.

Another way to estimate R_S values typical of 'good' predictions is to compare them to some other prediction methods. One of such methods is CABSflex (described in section 1.5.2.3; Jamroz et al., 2014, 2013a, 2013b). It is a coarse-grained method that attempts to predict protein backbone dynamics from a single structure. It was shown to perform very well in comparison with both NMR and MD results (Jamroz et al., 2013b). Unlike FRAGFOLD-IDP, CABSflex is given a significant advantage of starting from a known structure. Therefore, it can be assumed that CABSflex predictions should constitute what can be assumed excellent FRAGFOLD-IDP predictions. Since the entire dataset used in this study has corresponding experimental structures, for each case in the benchmark set a single structure (`MODEL 1`) was extracted from the PDB file and submitted to the CABSflex server (`http://biocomp.chem.uw.edu.pl/CABSflex/`; Jamroz et al., 2013a). After obtaining CABSflex per-residue RMSD predictions, the results were again evaluated using R_S. Since CABSflex starts from a known structure, there is no need to remove the rigid body motions, as in the case of FRAGFOLD-IDP. Therefore, global superposition profiles were compared, instead of the sliding window superposition results (compare section 2.4.3). Mean R_S achieved this way on the benchmark set is 0.66 (median 0.70). The results are close to the ones reported in the CABSflex paper, comparing CABSflex simulations to NMR per-residue fluctuations using RMSF, instead of RMSD (Jamroz et al., 2014). The paper reports R_S values = 0.72 (±0.15). Again, this confirms that R_S values of around 0.6 could be considered typical of very good predictions and 0.7 and above, excellent predictions.

To put this discussion into the context of FRAGFOLD-IDP results, let us consider an example from the benchmark set. An example of an excellent prediction would be 2KKJ, the nuclear coactivator binding domain of CBP (Figure 20). The prediction achieves R_S = 0.76. FRAGFOLD-IDP correctly identifies the highly disordered termini and a disordered region between residues 15 and 25. It only fails to correctly predict a short region of disorder between residues 35 and 40. Although, per-residue RMSD values between the NMR ensemble and FRAGFOLD-IDP predictions do not match exactly, the trends can be clearly seen. The fluctuations at the N-terminus (residues 1-12) follow the same descending trend. So do the predictions of the C-terminal region (residues 44-59), including 2 troughs at residues 50 and 55. The large mid-sequence disordered region (residues 15-25) is also well reproduced in FRAGFOLD-IDP predictions – per-reside RMSD values are smaller than those in the termini, but 2 local maxima and the breadth of the region are well predicted.

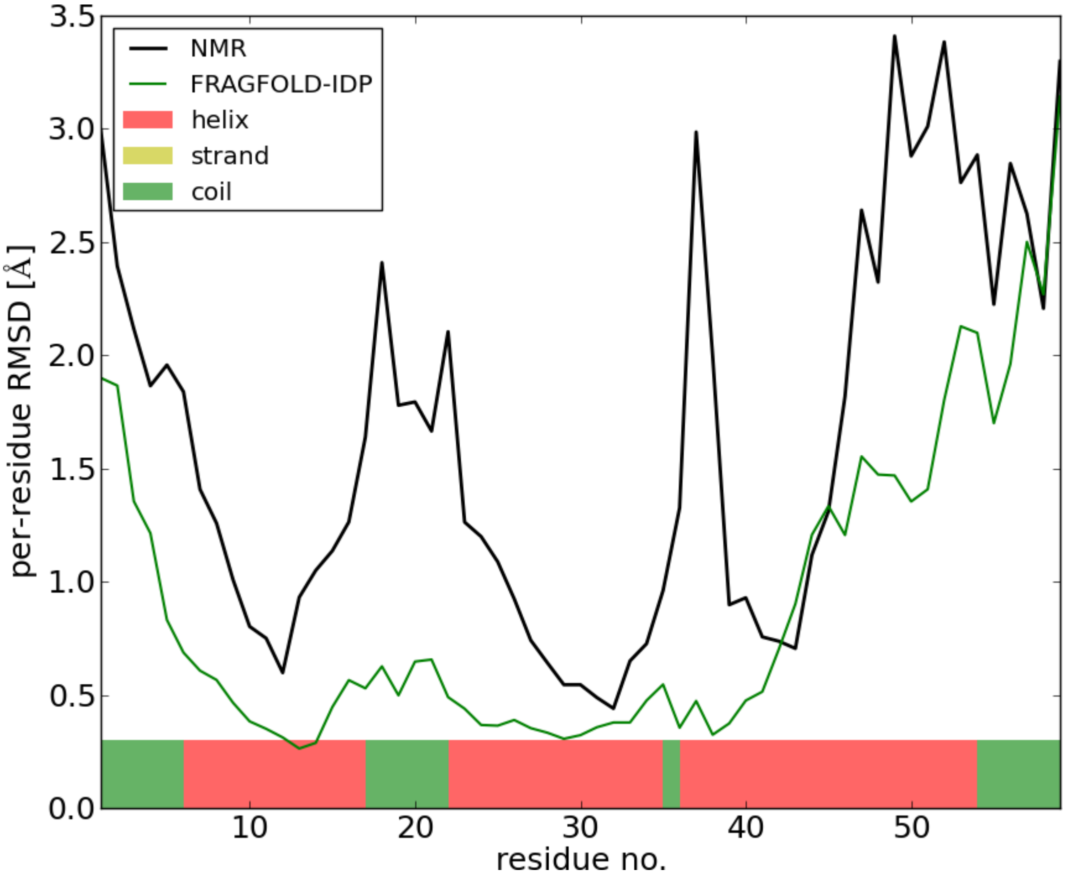


Figure 20. Disorder profile of 2KKJ (nuclear coactivator binding domain of CBP). PSIPRED secondary structure predictions are represented as a colour bar at the bottom of the plot.

2.5.2. *Overall performance*

Equipped with intuition as to how to interpret the R_S values, it is possible to discuss the overall performance of FRAGFOLD-IDP on the entire dataset (Figure 21). The mean R_S value is 0.44 and median is 0.48.

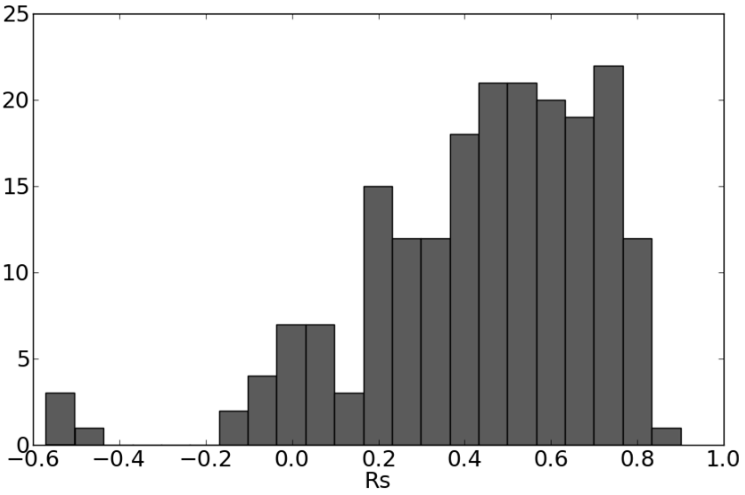


Figure 21. Distribution of FRAGFOLD-IDP results on the 200 protein dataset.

Out of 200 proteins in the dataset, 187 predictions have $R_S > 0$ (93.5%). There are 4 clear outliers, having predictions with $R_S < -0.4$ (discussed in section 2.5.2.1). On the other hand, there are 67 very good predictions with $R_S \geq 0.6$ (33.5%). They include 35 excellent predictions with $R_S \geq 0.7$ (17.5%).

An example of a poor prediction is 1SIY – lipid transfer protein 1 (Figure 22). The prediction achieves an R_S value of 0.21. Indeed, the disorder profile is not informative. Although the disordered region between residues 50 and 62 is correctly identified, the noise coming from false positives makes it lost in 4 other highly disordered regions predicted by FRAGFOLD-IDP. Also, the short disordered region around residue 20 is completely missed in the prediction.

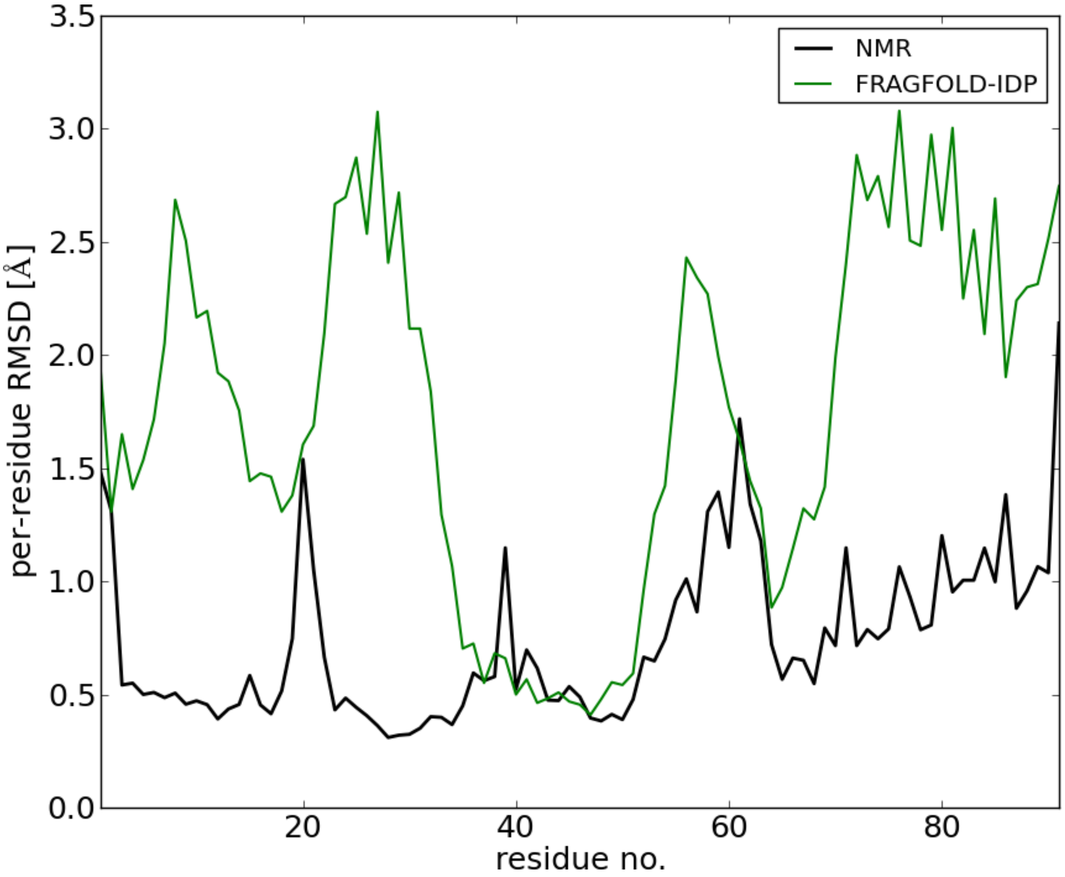


Figure 22. Disorder profile of a poor FRAGFOLD-IDP prediction (1SIY; R_S = 0.21).

An example of a medium quality prediction is 1P94 – ParG protein (Figure 23). The prediction achieved R_S = 0.54, which is close to the median value of the predictions on the entire dataset. Here, FRAGFOLD-IDP correctly identifies first 15 residues as highly disordered, but underestimates the breadth of this region, which spans 35 residues. Finally, the predictions from around residue 48 to 76 are correctly identified as ordered and the disorder profile shows low per-residue RMSD values.

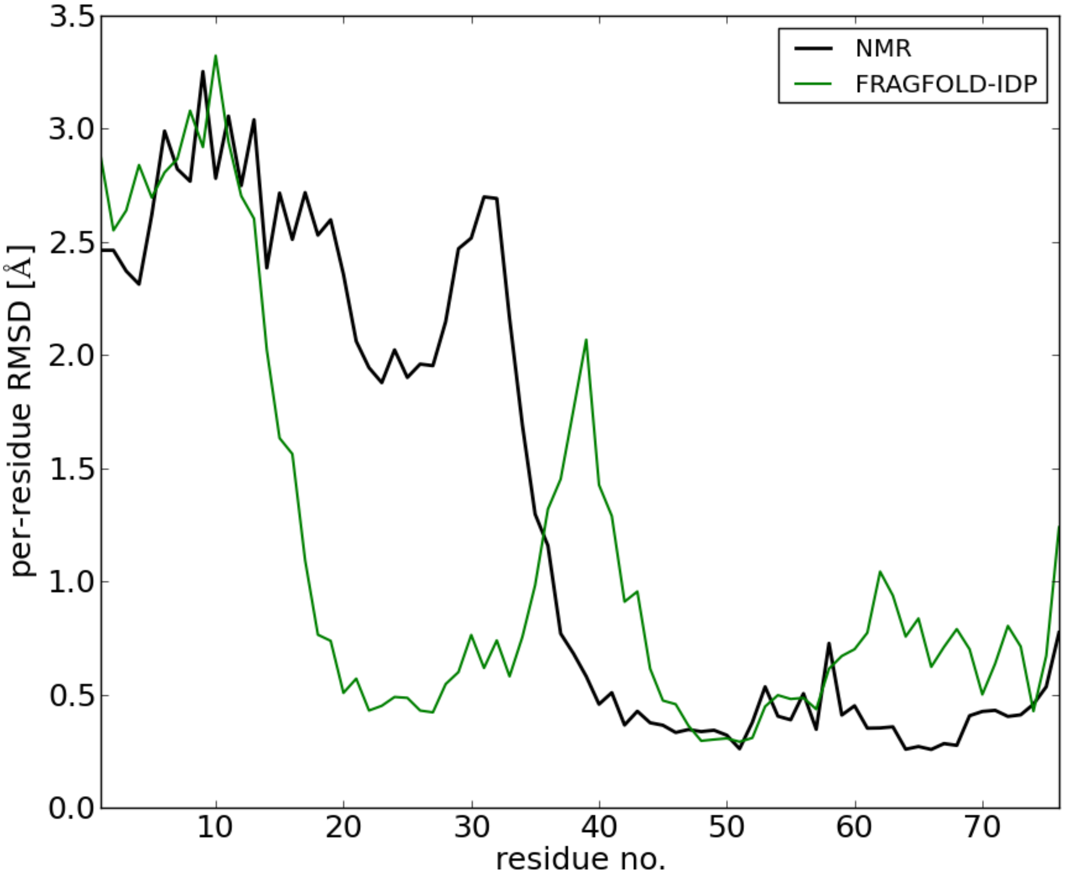


Figure 23. Disorder profile of a medium quality FRAGFOLD-IDP prediction (1P94; R_S = 0.54).

An example of an excellent prediction is 2KJV – ribosomal protein S6 (Figure 24). It achieves an R_S value of 0.82. FRAGFOLD-IDP captures all of the features of the NMR disorder profile remarkably well. The large disordered region between residues 40 and 60 is well reproduced, although FRAGFOLD-IDP slightly overestimates it, extending the region to around residue 35. The C-terminal region (residues 82-101) is also slightly overestimated and in FRAGFOLD-IDP it starts around residue 79. Finally, a small medium disorder region around residue 10 is captured by FRAGFOLD-IDP, but it spans from residue 1 to 15, instead of residue 7 to 12. The increase in per-residue RMSD signal could be partially attributed to the way sliding window (window size = 10) superposition works, i.e. from residues 1 to 9 there are less averaging steps, because of the sliding window size – residue 1 is superposed only once, residue 2 twice, etc.

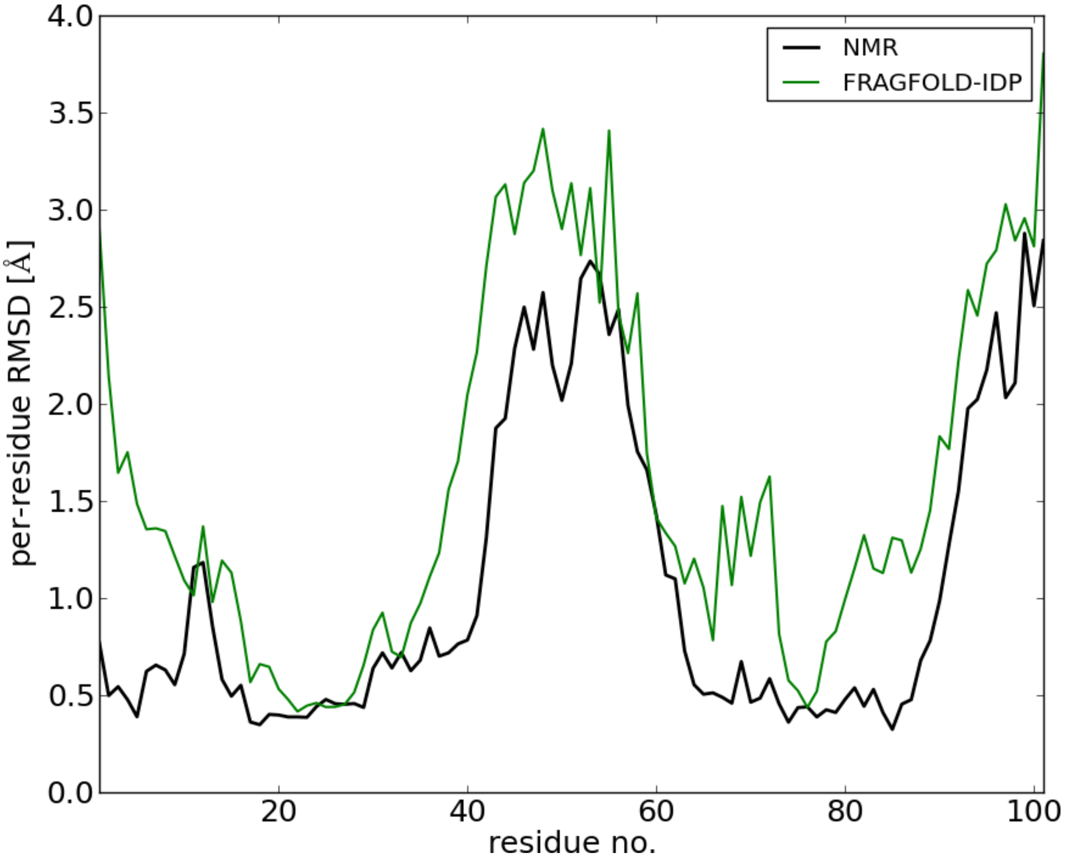


Figure 24. Disorder profile of an excellent FRAGFOLD-IDP prediction (2KJV; R_S = 0.82).

2.5.2.1. Outliers

An interesting aspect of the initial results are also the outliers in the distribution (Figure 21). There are 4 proteins that can be identified as such ($R_S < -0.4$). The results of the outliers are gathered in Table 10. The set contains proteins shorter than an average in the dataset (75 residues in outliers and 105 residues in the dataset), but have a typical disorder content (29% in outliers, 33% in the dataset). FRAGFOLD-IDP R_S is the output of the FRAGFOLD-IDP method, best cluster R_S represents the highest R_S result generated on the same set of models as FRAGFOLD-IDP R_S, but selecting the highest R_S among the clusters generated by PFClust. Top and median R_S values come from 1,000 random ensembles (as described in section 2.3.5.2 and discussed in section 2.5.4) generated from the same raw ensemble, as previously. Naïve R_S are the results of the naïve approach that uses only secondary structure prediction, but does not require any simulations (described in section 2.3.5.1 and discussed later, in section 2.5.3).

Table 10. Outliers in FRAGFOLD-IDP predictions.

protein	length	% disorder	FRAGFOLD-IDP R_S	best cluster R_S	top R_S	median R_S	naïve R_S
1G6M	62	33.87	-0.55	-0.16	-0.23	-0.50	0.26
1K0T	80	36.25	-0.57	-0.33	0.19	-0.48	-0.02
1XN7	78	20.51	-0.49	-0.49	0.69	-0.54	0.66
2K02	79	24.05	-0.51	0.67	0.87	0.71	0.72

Two of the cases among the outliers are clearly related to the ensemble extraction method – 1XN7 and 2K02 (Table 10). The final FRAGFOLD-IDP ensemble results are low, but among FRAGFOLD-generated models (raw ensembles) there are some with excellent R_S values (top R_S). In case of 2K02, the poor result can be attributed to cluster selection criteria, as among the PFClust-generated clusters there is one which achieves a very good results (best cluster R_S = 0.67). 1XN7 is a more general ensemble extraction problem, as the clustering algorithms do not extract a high quality cluster at all – both FRAGFOLD-IDP R_S and best cluster R_S are -0.49. However, FRAGFOLD is able to generate better ensembles for this target, with top R_S reaching a very good R_S value = 0.69. Also, the naïve approach deals well with this target (naïve R_S = 0.66).

The remaining cases – 1G6M and 1K0T – are more challenging (Table 10). Although all results – best cluster, top cluster and median results are better than the selected cluster, the R_S values are still very low (the highest R_S for 1G6M = -0.16 and for 1K0T R_S = 0.19). Comparison with the naïve approach hints that some FRAGFOLD problems are likely, as for 1G6M the naïve approach generated the best result of all of the attempts (naïve R_S = 0.26), and for 1K0T only the top R_S is higher than the naïve result (top random cluster R_S = 0.19). Still, even the naïve calculations produce results far lower than for the 2 cases discussed previously (1XN7 and 2K02).

1G6M is a snake Cobrotoxin II from *Naja kaouthia* (Monocled cobra). It is a mostly beta sheet protein. From the NMR PDB ensemble of 1G6M it can be inferred there are 4 disulphide bridges that constrain the structure making it more ordered (Figure 25). The bridges are evenly spaced (bridge 1: residues 3 & 24; bridge 2: 17 & 41; bridge 3: 43 & 54; bridge 4: 55 & 60) within the protein structure and constrain the loop regions. There are no disulphide bridges in the beta-hairpin region (residues 25-

40). Those bridges are the likely cause of poor predictions of the backbone dynamics achieved by FRAGFOLD-IDP.

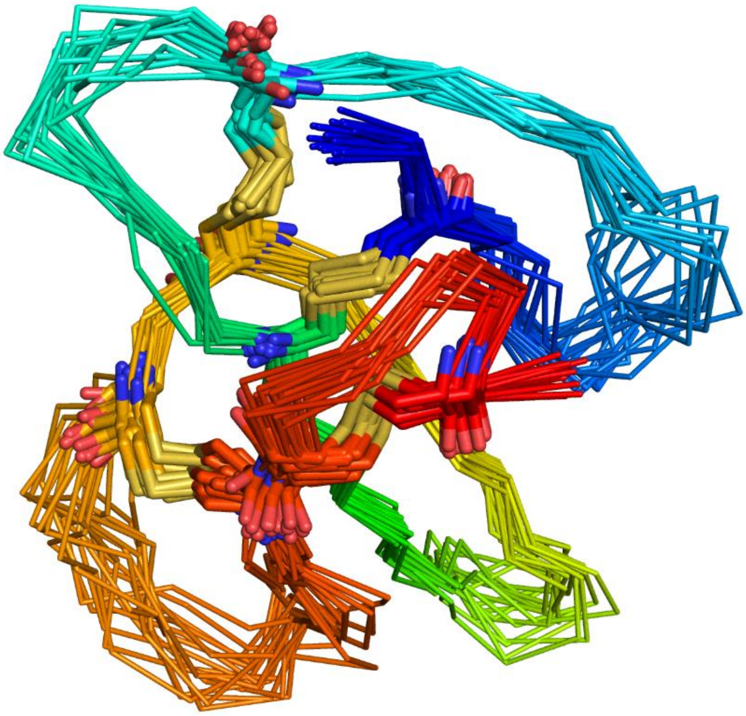


Figure 25. NMR PDB ensemble of 1G6M (Cobrotoxin II). Disulphide bridges are represented as yellow sticks in the ensemble. The rest of the structure is shown in ribbon representation and each conformation in the PDB ensemble is shown as a separate structure.

1K0T is photosystem I subunit PsaC from *Synechococcus sp*. Although the protein passed all of the dataset criteria (section 2.2.1), it has two inorganic clusters (Fe_4S_4) covalently bound to the protein (Figure 26). Such modification is likely to alter backbone dynamics of the protein. It can also be confirmed by the fact that other backbone dynamics predictors evaluated fail to significantly improve (e.g. DynaMine R_S = 0.23) over FRAGFOLD-IDP predictions (comparison with other methods is presented later, in section 2.5.8).

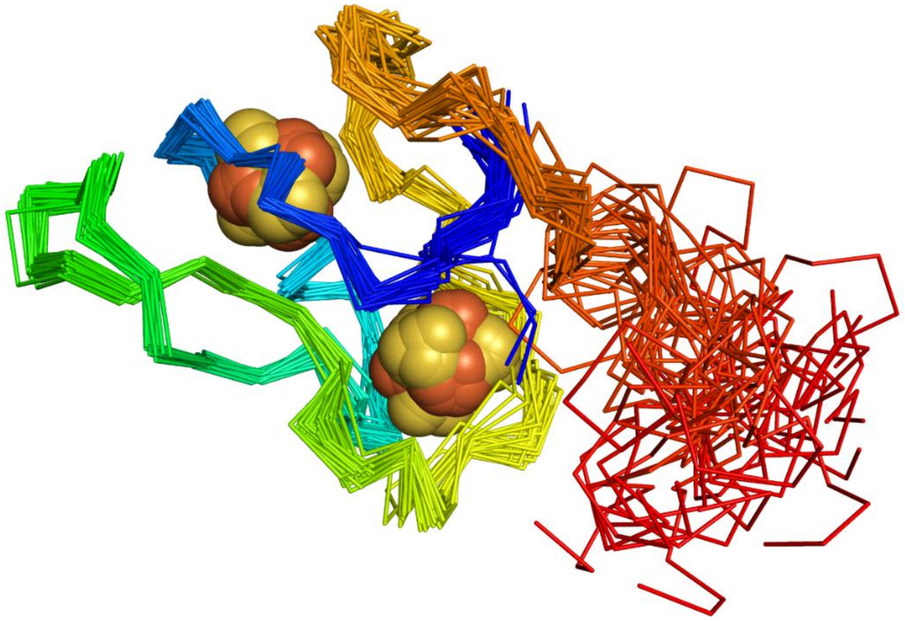


Figure 26. NMR ensemble of 1K0T (photosystem I Subunit PsaC). The ensemble is rainbow-coloured (N-terminus – blue, C-terminus – red). Inorganic (Fe_4S_4) clusters are represented as yellow and orange spheres.

2.5.3. ***FRAGFOLD-IDP and naïve predictions***

For every computational method, it is crucial to estimate how well it performs in relation to a baseline method, i.e. a naïve method. The approach developed for FRAGFOLD-IDP (described in section 2.3.5.1) assumes that all predicted loops are disordered, all beta-sheets allow for some flexibility, whilst all helices are rigid. The method does not require any simulations to be carried out, but it requires a sequence profile to be computed in order to perform secondary structure predictions.

A method that would not require any computations could make assumptions that all residues at the protein termini are disordered, while all mid-sequence residues are ordered. An example of such approach is the naïve method used for disorder predictions assessment in CASP10 (Monastyrskyy et al., 2014). The authors assumed that the first nine and final four residues are disordered and that all remaining residues are ordered.

Therefore, the method used in this work is more sophisticated than what was used in CASP10 disorder assessment. For each protein sequence, no arbitrary choice is made, but instead the simplified disorder profile is dictated by the predicted secondary structure of the protein. This is justified by the problem at hand. In CASP disorder predictions, the goal is to make disorder/order classification, whilst with FRAGFOLD-IDP the goal is to predict protein backbone dynamics.

The naïve method was optimised to maximize its performance on the benchmark set (section 2.3.5.1). Overall, the naïve method achieves a mean R_S = 0.37 (median R_S = 0.38) on the complete dataset, while FRAGFOLD-IDP achieves a mean R_S = 0.44 (median R_S = 0.48). FRAGFOLD-IDP results are significantly better than the naïve method (p-value = 0.004).

Although, FRAGFOLD-IDP predictions are better than the naïve method, 76 out of 200 predictions on the dataset are better or equal to FRAGFOLD-IDP predictions in terms of R_S using the naïve method. Considering only very good predictions ($R_S \geq 0.6$), the naïve approach is better than FRAGFOLD-IDP in 16 cases. The comparison of R_S values between FRAGFOLD-IDP and the naïve approach is summarized in Figure 27.

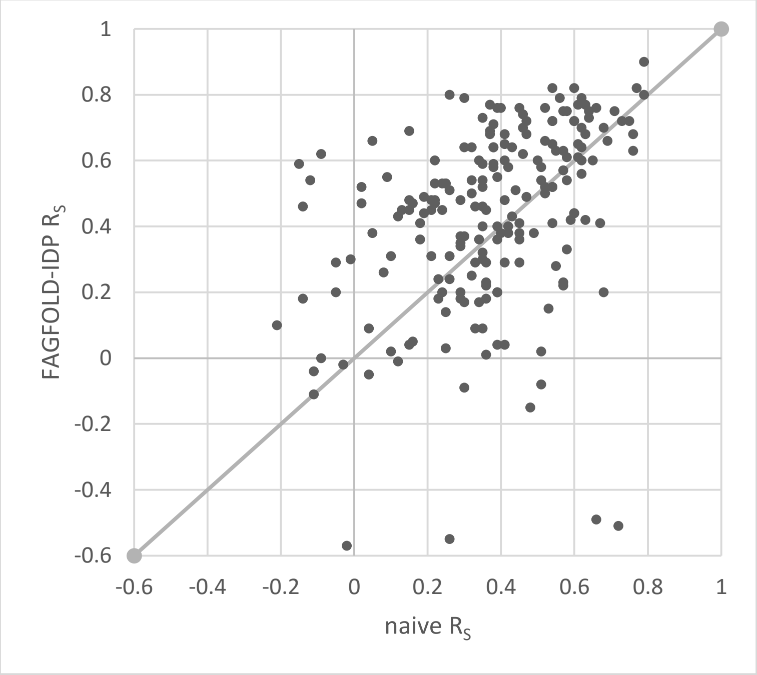


Figure 27. Comparison of R_S values between a naive approach and FRAGFOLD-IDP.

From the results mentioned above, it is clear that FRAGFOLD-IDP goes beyond predicting only the flexibility of loops (what the naïve method assumes) and FRAGFOLD simulations add value to the predictions.

An example of such added value could be the previously discussed nuclear coactivator binding domain of CBP (Figure 20). Using FRAGFOLD-IDP, the protein backbone dynamics are predicted with R_S of 0.76, while the naïve prediction achieves R_S of 0.46. The two terminal regions in 2KKJ both partially overlap with predicted helices. FRAGFOLD-IDP accurately reproduces those regions at both termini. At the N-terminus, part of the predicted helix (residues 5-10) is included in the disordered region that spans residues 1 to 10. At the C-terminus the long disordered region begins 5 residues into the predicted helix, spans the rest of it and includes the C-terminal predicted loop. Also, the mid-sequence low disorder region between residues 15 and 25 includes parts of helices flanking the coil region which is the centre of this disordered region.

Most of the disordered regions are within predicted loops, hence the performance of the naïve approach is relatively good. An example of a case where the naïve method performs remarkably well is 1NSH (naïve R_S = 0.69; FRAGFOLD-IDP R_S = 0.66; Figure 28). In this protein, secondary structure predictions precisely reflect protein backbone dynamics and there is no space for FRAGFOLD-IDP to improve over those predictions. Regardless, the method is able to reproduce the backbone dynamics well. Nevertheless, the previous example (2KKJ) exemplifies that FRAGFOLD-IDP as able to go beyond the predictions of loop flexibility and adds important information to the predictions of protein backbone dynamics.

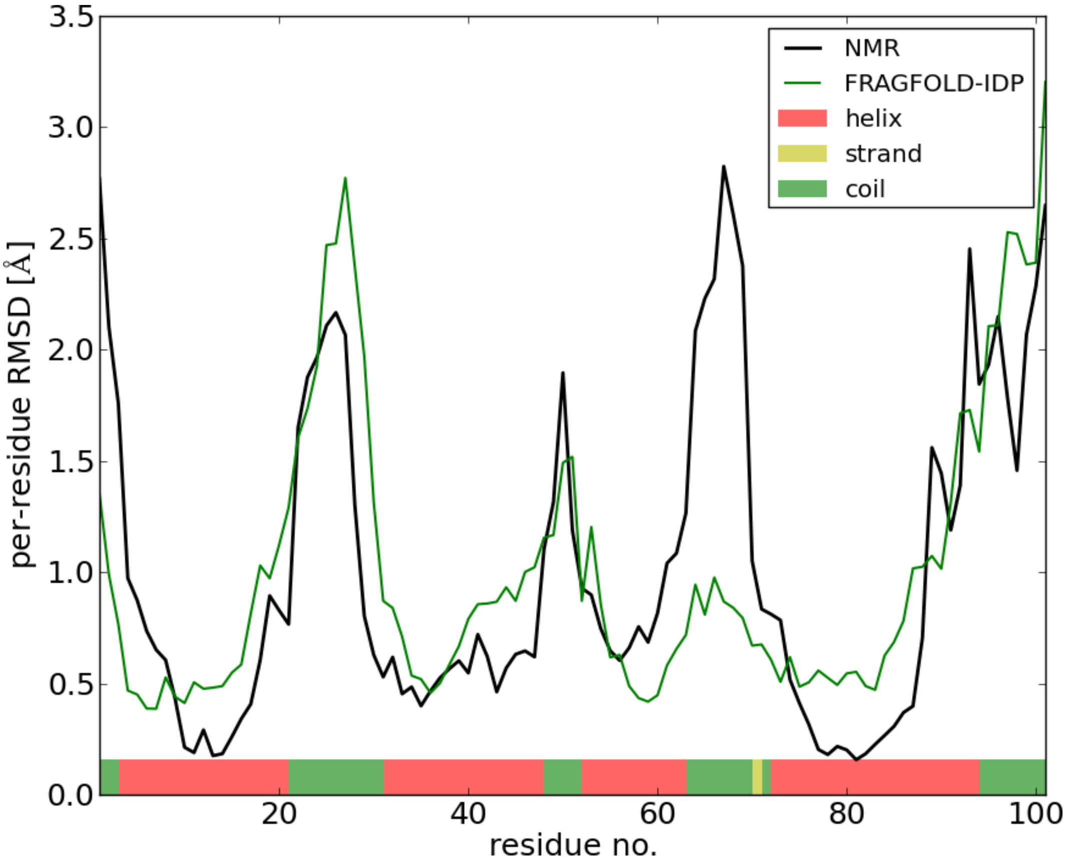


Figure 28. Disorder profile of 1NSH (Rabbit protein S100-A11).

2.5.4. *Comparison of final FRAGFOLD-IDP ensembles with random and best clusters*

Using FRAGFOLD-IDP, a single output ensemble is generated for each protein sequence (section 2.4.6). However, to analyse the results in a broader context it is useful to look also at alternative ensembles that could be generated from FRAGFOLD raw ensembles. It can help to better assess the performance of the elements of FRAGFOLD-IDP workflow and to determine which parts of the method could be improved in future (e.g. ensemble generation, or ensemble extraction).

To perform this comparison, three sets of alternative ensembles were extracted:

(1) best FRAGFOLD-IDP cluster (highest R_S cluster from the set of clusters generated by PFClust from the raw ensemble);

(2) median R_S cluster from the set of 1,000 random ensembles (described in section 2.3.5.2);

(3) top R_S cluster from the 1,000 random ensembles set.

Each cluster in the 1,000 random ensembles consists of 10 structures per protein. The optimal ensemble size was discussed in section 2.4.5.

The results from 1,000 random ensembles serve as a baseline for the comparison of the ensemble extraction methodology. Figure 29 compares the performance of FRAGFOLD-IDP against median random cluster.

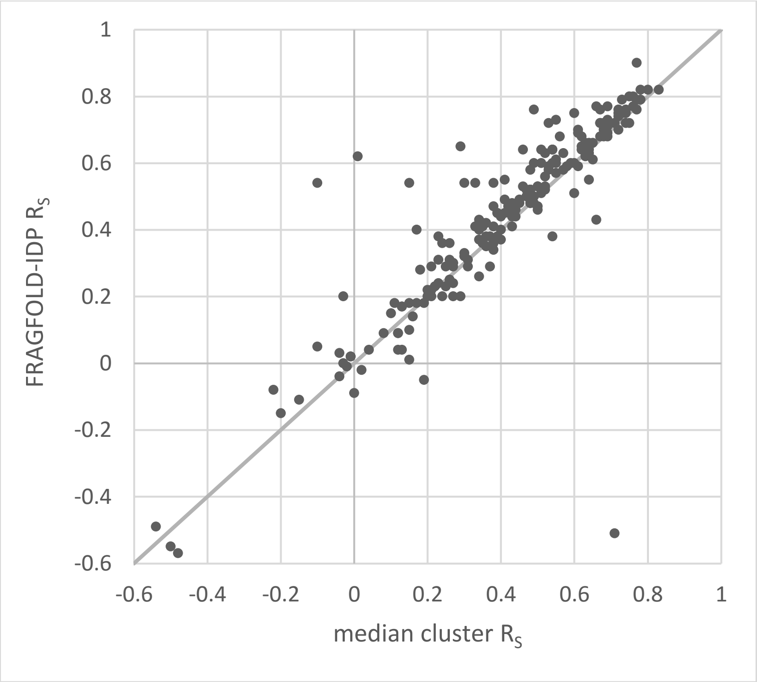


Figure 29. Comparison of R_S from the final cluster extracted using PFClust with median R_S extracted from 1,000 randomly generated ensembles.

Overall, the current clustering methodology performs slightly better than median cluster selection. The mean R_S value for median cluster is 0.40 and median R_S is 0.43 (mean FRAGFOLD-IDP R_S = 0.44; median R_S = 0.48). For 139 out of 200 proteins in the dataset, higher R_S is achieved with FRAGFOLD-IDP than with median random cluster. This shows that the cluster selection methodology employed in FRAGFOLD-IDP works well and improves the predictions. The clear outliers in favour of median cluster are the same as indicated previously in section 2.5.2.1. There are also some outliers in favour of FRAGFOLD-IDP, e.g. 1NY4, 30S ribosomal protein S28e from *Pyrococcus horikoshii* (median cluster R_S = 0.01; FRAGFOLD-IDP R_S = 0.62; naïve R_S = -0.09; Figure 30). It is a small, 71 residue mostly beta protein. In this case, FRAGFOLD-IDP is able to precisely differentiate disordered loops from ordered ones, e.g. loop at residues 34-39 and 46-52. The median cluster achieves similar results to the naïve method.

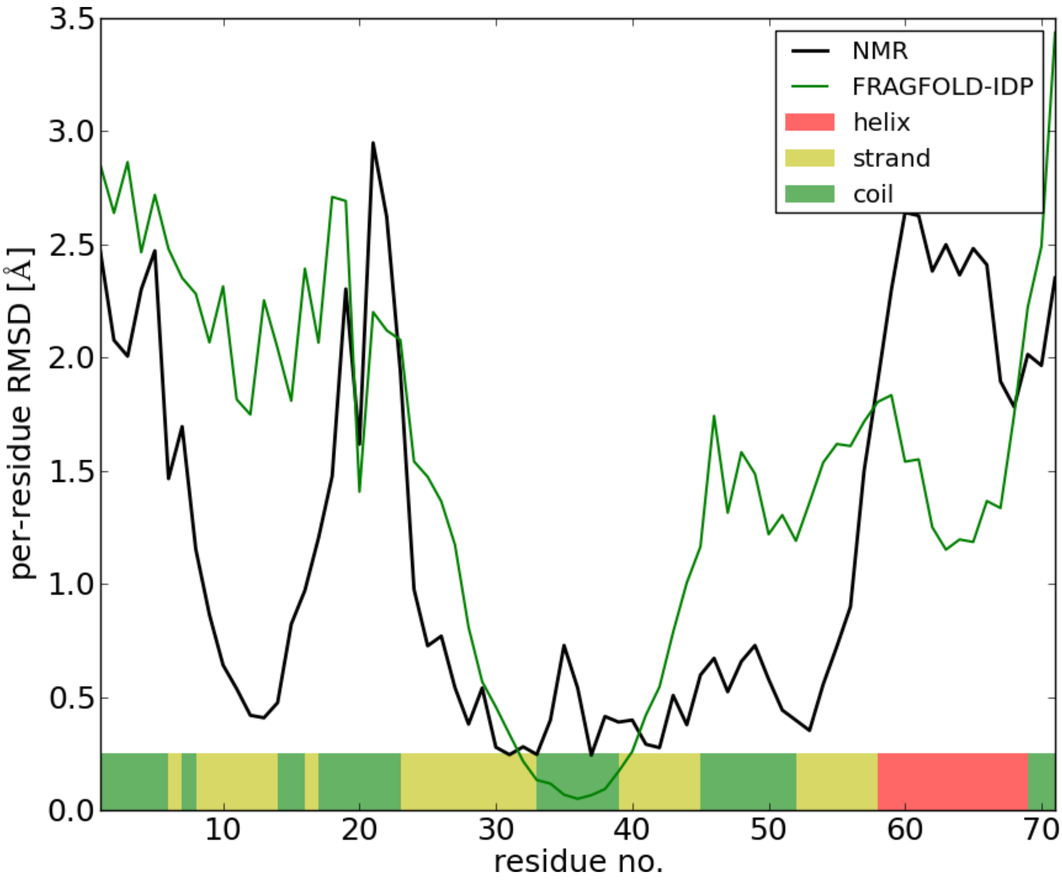


Figure 30. Disorder profile of 1NY4 (30S ribosomal protein S28e from Pyrococcus horikoshii).

Another aspect is to compare how FRAGFOLD-IDP ensembles compare to the best FRAGFOLD-IDP cluster ensembles (Figure 31). In 42 cases the FRAGFOLD-IDP result is the best cluster and in 102 cases FRAGFOLD-IDP result is within 0.05 R_S of the best cluster. This confirms that the cluster selection criterion works reasonably well. The outlier in Figure 31 is the case which was already discussed in section 2.5.2.1 (PDB id: 2K02).

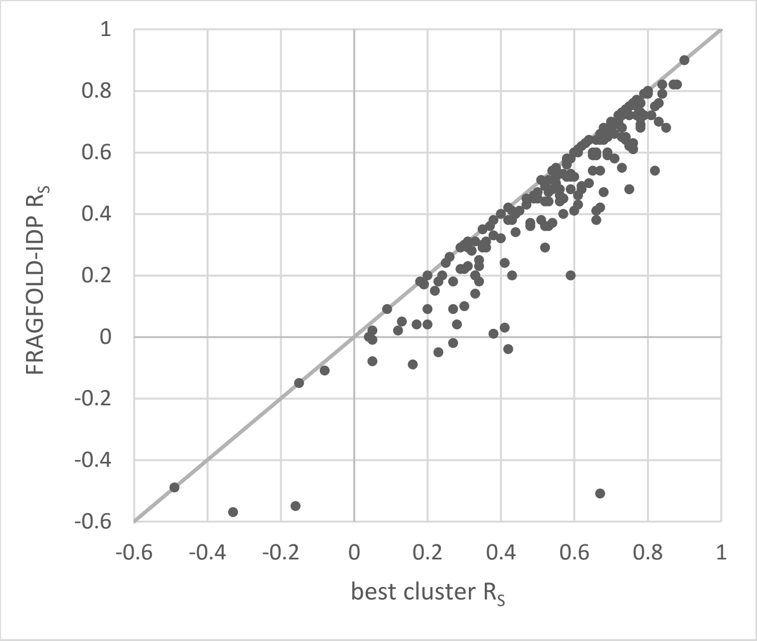


Figure 31. Comparison of the best cluster from PFClust in terms of R_S with FRAGFOLD-IDP.

The final comparison is between the best FRAGFOLD-IDP clusters and the top random clusters (Figure 32). The best cluster achieved by PFClust is close to the top cluster extracted from 1,000 random ensembles. One notable outlier is 1XN7 (ferrous iron transport protein C) having top random cluster R_S = 0.69 and best cluster R_S = -0.49 (discussed in section 2.5.2.1). The results show that selecting the top cluster from 1,000 random ensembles is a good proxy for estimating what is possible to achieve using FRAGFOLD. The best PFClust clusters lay close to the top random cluster results and are highly correlated (Pearson's r = 0.87). This means that the ensemble extraction strategy employed works effectively for the problem at hand.

Overall, the results in this section show that FRAGFOLD-IDP works better than a median cluster from a random set of models. The results also show that the cluster selection criteria employed in FRAGFOLD-IDP are good, however it is likely that the results could be improved, should better cluster selection criteria, or clustering methodology exist. PFClust, in general, performs well and the clusters generated using this approach are close to what could be achieved within FRAGFOLD-IDP framework.

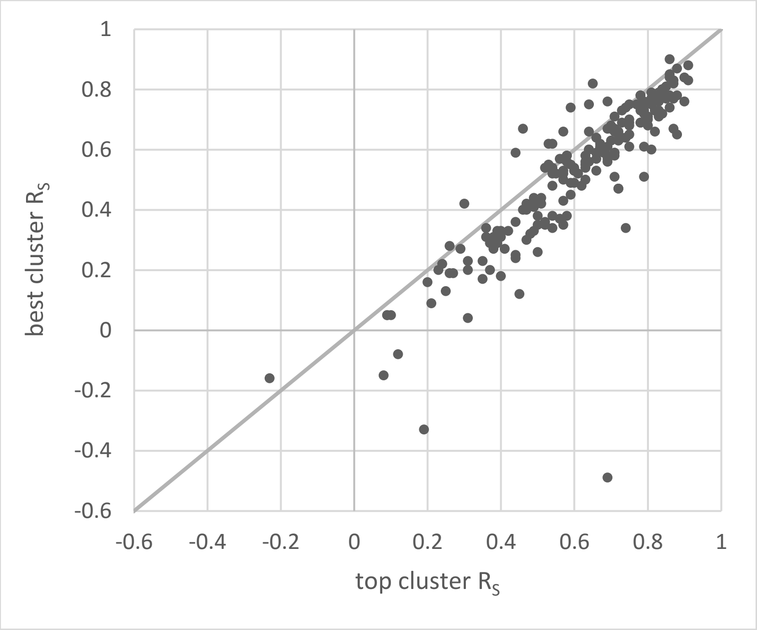


Figure 32. Comparison of the top cluster from 1,000 randomly generated ensembles with the best cluster in terms of R_S extracted using PFClust.

2.5.5. *Impact of protein class on FRAGFOLD-IDP performance*

A crucial aspect of protein structure prediction is the structural class of the protein. Usually, all-alpha proteins are easier to predict than all-beta proteins because of the local hydrogen bonding patterns and the number of local interactions (Kosciolek and Jones, 2014). The same applies to the predictions of protein secondary structures (Cuff and Barton, 2000; Jones, 1999). The aim of FRAGFOLD-IDP is not to predict the structure of the protein, but rather its backbone dynamics. Still, it is an interesting and important aspect to investigate, whether the quality of backbone dynamics predictions also depends on the class of the analysed protein.

For each protein in the dataset, the structural class defined by the CATH database was extracted from the latest CATH version 4.0 (Orengo et al., 1997; Sillitoe et al., 2015). CATH classifies proteins into four categories – mainly-alpha (alpha), mainly-beta (beta), alpha/beta and few secondary structures (few). Another class was constructed from proteins which were not indexed in CATH (none). Then, the top random cluster, median random cluster and FRAGFOLD-IDP results were compared on the basis of their R_S results in each CATH class (Figure 33).

There are 58 proteins in all-alpha class, 30 in all-beta, 60 in alpha/beta class, 7 in few secondary structures class and 45 not classified in CATH (none class).

The results are consistent between different ensemble extraction methods, i.e. all-alpha and few secondary structure protein classes achieve the highest scores when analysing top random cluster, median random cluster and FRAGFOLD-IDP results.

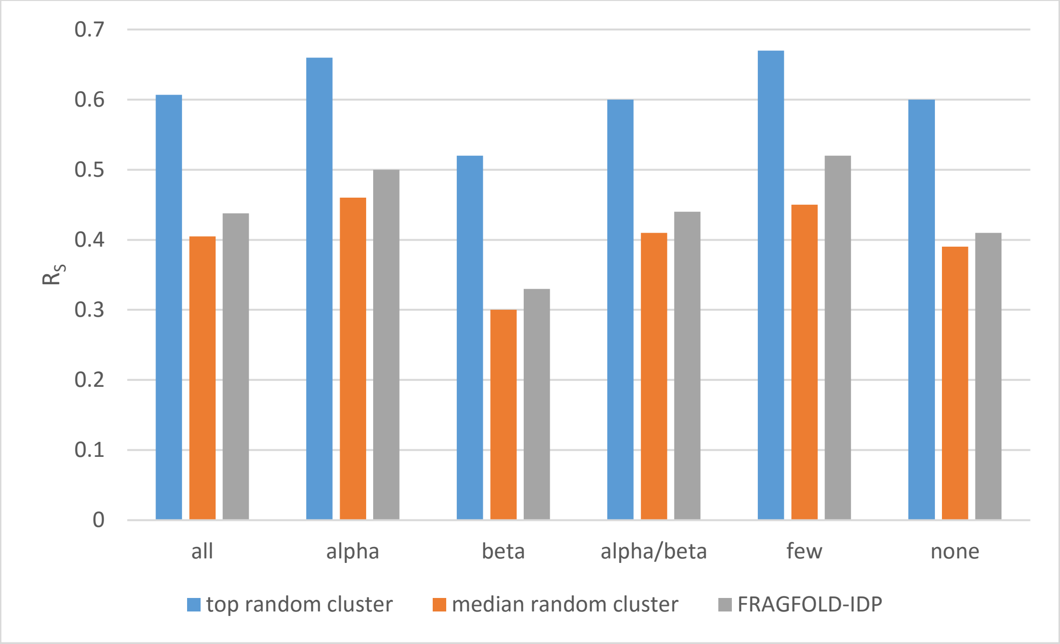


Figure 33. Comparison of full dataset results by CATH class.

Overall, the predictions can be grouped into 3 main categories. All-alpha and few secondary structures proteins perform better than average (all proteins). Alpha/beta proteins perform on par with the average. Beta proteins and none class proteins (not indexed in CATH) perform worse than average. This is a similar behaviour to what could be expected from traditional protein structure predictions (e.g. Kosciolek and Jones, 2014). At the same time, the largest gap between the top R_S and FRAGFOLD-IDP R_S is observed for mainly-beta and none class proteins, while the smallest gap can be observed for all-alpha, alpha/beta and few secondary structures class.

CATH classification does not differentiate proteins on the basis of their disorder content, i.e. proteins with a high disorder content do not necessarily belong to the few secondary structures class or none class. An example could be 2FKX, 30S ribosomal protein S15 from *Thermus thermophilus*. According to MOBI, it is 46.6% disordered (88 residues). Nevertheless, the NMR ensemble of this protein shows it is a helical molten globule, hence its CATH classification as an all-alpha protein. An example of a high disorder all-beta protein is 1TXB, long neurotoxin 2 from *Ophiophagus hannah* (king cobra) (Figure 34). According to MOBI it is 79.4% disordered (73 residues). Again, the protein can be described as a molten globule and although it is highly disordered, a beta-sheet core of the protein is identifiable (Figure 34). FRAGFOLD-IDP produces a poor prediction for this protein (R_S = 0.28).

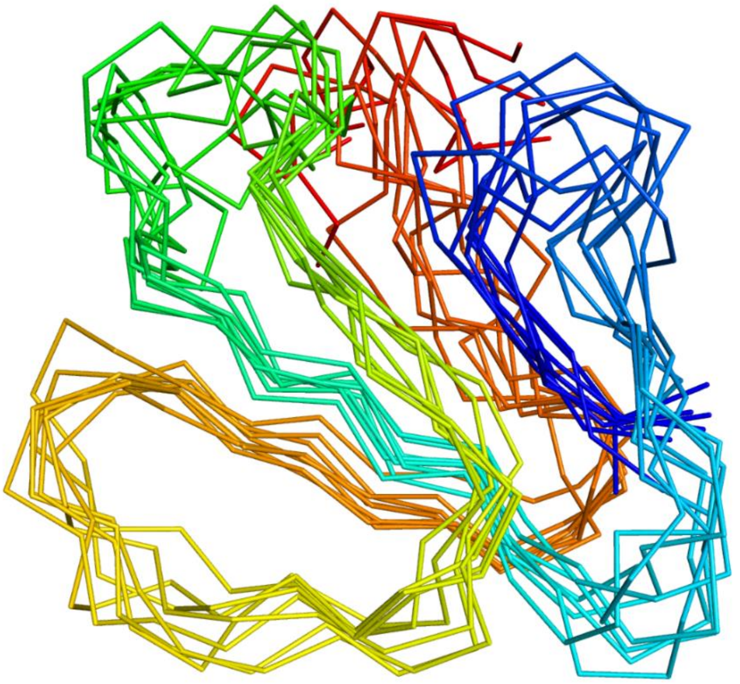


Figure 34. NMR PDB ensemble of 1TXB (long neurotoxin 2). The protein is highly disordered (79.4%), still it is classified by CATH as an all-beta protein. A beta-sheet core of the protein is visible.

These observations show that the class of the protein plays a role in the predictions of protein backbone dynamics. The fact that the classification is robust regardless of the cluster extraction criteria prove that the observations are not biased by the ensemble extraction protocol. However, they might be biased by the model generation methodology, i.e. FRAGFOLD (sections 2.1.3, 2.3.2 and 2.4.1). Even though structure predictions and backbone dynamics predictions are decoupled in

FRAGFOLD-IDP by the use of the sliding window (section 2.4.3), it is possible that the amount of non-local interactions often encountered in all-beta proteins plays a role in the predictions of backbone dynamics and does not only impact structure prediction. This way, the structure prediction quality in (usually more difficult to predict) all-beta proteins may have an impact on the quality of backbone dynamics predictions. The impact of structure prediction quality on the predictions of backbone dynamics is discussed elsewhere (section 2.5.7). Also, the links between structure prediction and per-class performance are discussed in the summary of this chapter (section 2.6).

2.5.6. *Impact of disorder content on FRAGFOLD-IDP performance*

FRAGFOLD was originally developed as a method to predict the structures of globular proteins (described in section 2.1.3). Up to this point in this chapter, I have shown that it is possible to use FRAGFOLD also as a method to predict protein backbone dynamics *de novo* from sequence. One of the reasons why this could be possible is the fact that the proteins included in the dataset are mostly ordered (mean disorder content in the dataset = 33%; section 2.2.2) and FRAGFOLD-IDP performs better on the targets with low disorder content, than those significantly disordered. To verify this hypothesis, FRAGFOLD-IDP results were plotted against disorder content extracted from NMR PDB ensembles using the MOBI method (Figure 35).

Indeed, most of the targets in the dataset contain between 10% and 50% disorder (173 out of 200 cases). Nevertheless, the disorder content in this region does not impact the quality of FRGFOLD-IDP predictions in general. Above 50% disorder content, the predictions are sparser in terms of sampling the amount of disorder (26 cases; 14 cases > 60% disorder). The best result for proteins with > 60% disorder is R_S = 0.6. It is not clear whether for highly disordered proteins FRAGFOLD-IDP is not able to perform any better. Still, the plot does not show any significant correlation (Pearson's r = -0.08; p-value = 0.27) between disorder content and the quality of FRAGFOLD-IDP predictions (Table 11).

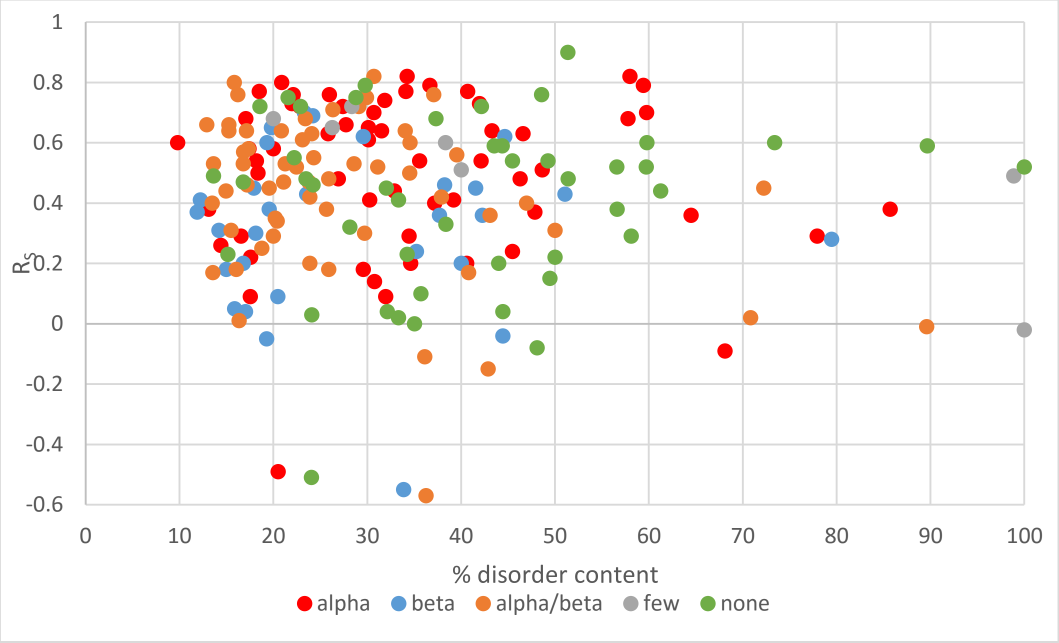


Figure 35. Relationship between the disorder content in NMR ensembles and the quality of FRAGFOLD-IDP predictions.

Analysing the results as in the previous section, on a per-CATH class basis, there are some more remarks that can be made (Figure 35, Table 11). Clearly, for all-alpha and all-beta proteins, there is no significant correlation between the quality of predictions and disorder content. For alpha/beta class however, there is a weak negative correlation which is statistically significant (p-value = 0.01). For not classified (none class) proteins, similarly there is a weak negative, but not significant correlation (p-value = 0.37). Finally, for few secondary structures (few class) there is a strong negative correlation between FRAGFOLD-IDP results and disorder content (Pearson's r = 0.80) which is borderline significant (p-value = 0.03). However, it should be noted that there are only 7 proteins belonging to this class – 5 cases have between 20% and 40% disorder and achieve high results, while 2 cases close to 100% (98.9% and 100%) achieve mixed results (Figure 35; grey dots). So although the results are shown to be borderline significant, the space between 40% and 100% of disorder is not sampled for this class.

Overall, there is no strong evidence that FRAGFOLD-IDP performance is dictated by the disorder content of the input protein on the whole. However, there are some indications that for alpha/beta and few secondary structures proteins, the increase in disorder content negatively impacts the quality of FRAGFOLD-IDP predictions. Possibly, this could be addressed by altering the ensemble generation parameters on the basis of secondary structure predictions.

Table 11. Correlations between FRAGFOLD-IDP R_S value and disorder content by protein CATH class.

	CATH class				
	all	**alpha**	**beta**	**alpha/beta**	**few**
Pearson's r	-0.08	-0.04	-0.02	-0.33	-0.80
no. proteins	200	58	30	60	7

2.5.7. *Structure and backbone dynamics predictions*

Sections 2.3.4 and 2.4.3 introduced the idea of separating protein backbone dynamics from protein structure predictions. Having established FRAGFOLD-IDP and taking advantage of this separation, it can now be compared whether the quality of backbone dynamics predictions depend on the quality of structure predictions. To do this, the TM-score for each protein in the dataset was calculated as described in section 2.3.4.4. The quality of backbone dynamics predictions is described by R_S values. Therefore, each protein can now characterized by 2 values – the TM-score representing the structure prediction quality between the 2 ensembles (FRAGFOLD-IDP and NMR PDB) and R_S reflecting the quality of backbone dynamics predictions. The results are plotted in Figure 36.

The immediate conclusion that is apparent from Figure 36 is that it is not necessary to find the correct fold of the protein to predict its backbone dynamics accurately. Several examples of that were already presented (compare 1K19 or 2K36 in Table 7). Posing an alternative hypothesis is more challenging – does high structure prediction quality (high TM-score) hinder the predictions of backbone dynamics (low R_S values)? The analysed dataset is under-represented in well-folded (TM-score ≥ 0.5) structures. Only 8% of the dataset (16 structures) has TM-score ≥ 0.5. In comparison, another research concentrated on predicting the structures of globular proteins found that FRAGFOLD is able to correctly predict around 14% – 25% of cases, depending on the final model selection criteria (Kosciolek and Jones, 2014). Because of the under-representation of high quality structure predictions, it is difficult to draw robust conclusions based on this dataset.

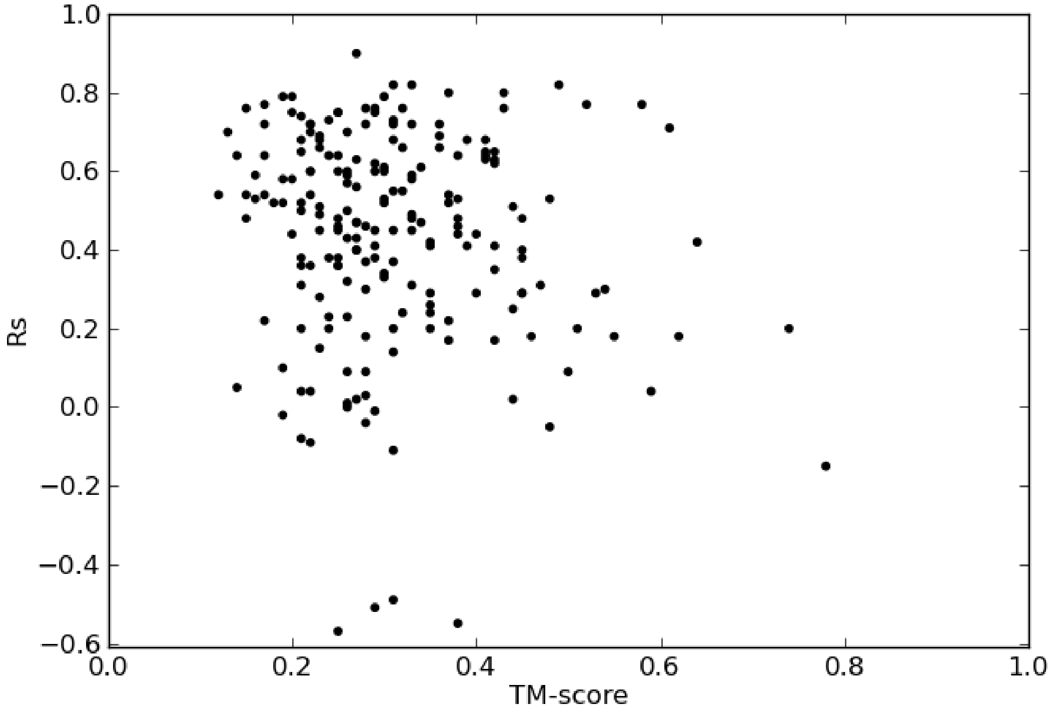


Figure 36. Impact of structure prediction quality (TM-score) on backbone dynamics predictions (R_S).

An example of a high TM-score, low R_S score case is 1APS (Horse acylphosphatase-2) (Figure 37). The TM-score of this ensemble is 0.78 – the highest value achieved on the dataset; R_S of both FRAGFOLD-IDP prediction and the best cluster is -0.15. Figure 37 shows the disorder profile of this target. Except for the N- and C-terminus regions, the disorder profile of FRAGFOLD-IDP shows per-residue RMSD values close to zero. This suggests that structural fragments of 1APS or its close homologue are in the FRAGFOLD fragment library. As a result, this would make the structure more rigid that it is in reality. During conformational sampling FRAGFOLD would be biased to select those homologous fragments and produce an ensemble of near-identical structures. To verify this hypothesis an additional validation was made, similar to the one presented in previous FRAGFOLD predictions validation (Kosciolek & Jones, 2014). FRAGFOLD was run in refinement mode, using a single structure from the PDB ensemble as a reference structure. During this process, FRAGFOLD verifies fragment selection by calculating the RMSD of each selected supersecondary structural fragment and 9-residue (fixed-length) fragment (compare section 2.1.3) to the reference structure. The output mean RMSD for fixed-length fragments was 0.97 Å and for supersecondary fragments 3.30 Å, which is close to the mean values reported previously. It therefore suggests that there are no homologous fragments to 1APS in the library.

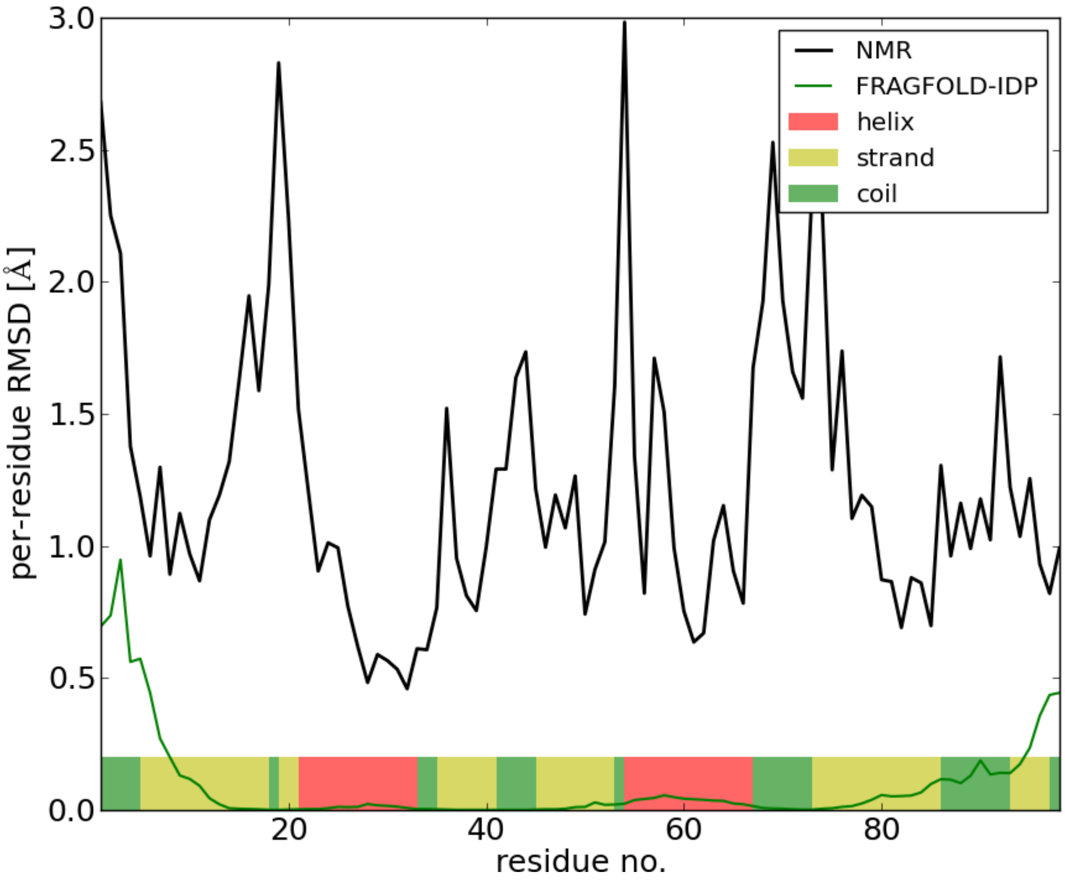


Figure 37. Disorder profile of 1APS (Horse acylphosphatase-2).

Also, 1APS has the lowest RMSD of the entire raw ensemble (0.14 Å) in the whole dataset, this means that the structures generated are near identical. It's similar with the next highest TM-score target 1T8V (TM-score = 0.74; raw ensemble RMSD = 0.19 Å). The mean raw ensemble RMSD over the entire dataset is 1.66 Å.

The overall tendency of how a correct structure impacts the disorder profile remains unclear (Figure 36). The data shows almost no correlation between the TM-score and R_S (Table 12). A very similar behaviour could be observed regardless chosen ensemble extraction method, e.g. analysing the data for the best cluster (instead of the selected largest cluster).

Table 12. Correlations between TM-score and R_S for different cluster selection criteria.

	mean TM-score	Pearson's r
FRAGFOLD-IDP cluster	0.31	-0.13
best cluster	0.31	-0.16
FRAGFOLD-IDP cluster (TM-score bins)	-	-0.13
best cluster (TM-score bins)	-	-0.18

To investigate this further, an alternative approach was also tested. TM-score is a continuous measure, so a TM-score of 0.2 indicates a worse model than one having TM-score of 0.3. But still, both of those models indicate unsatisfactory predictions. Following this rationale, TM-score values were binned and the correlations were recalculated. The bin boundaries were established based on the findings of Xu & Zhang (Xu and Zhang, 2010). The first bin contains TM-score values from 0 to 0.2, corresponding to random non-homologous structures. The second bin includes ensembles with TM-score between 0.2 and 0.4 TM-score – values where the posterior probability of 2 structures belonging to the same CATH or SCOP class is close to zero. The third bin contains TM-scores between 0.4 and 0.6 – the “phase transition” region, where the probability of the two protein belong to the same fold increases drastically and reaches around 90%. The last bin includes ensembles with TM-score above 0.6 and contains cases where the posterior probability of the two proteins/ensembles belonging to the same fold is > 90%. The bin boundaries were also validated to maximize the Pearson’s r correlation value between TM-score and R_S.

Still, applying this binning protocol did not improve the results significantly (Figure 38, Table 12). 162 targets that belong to bin 1 or 2 have FRAGFOLD-IDP ensembles that are unlikely to belong to the same CATH or SCOP class as their NMR PDB counterparts. Mean R_S values in bins 1 and 2 are on average higher than the ones in bins 3 and 4 (Figure 38). There are 33 proteins in bin 3 and 5 proteins in bin 4, with a total of 14 all-alpha proteins, 11 all-beta, 11 alpha/beta and 2 none class proteins. Comparing the enrichment of protein populations in the two top bins (3 and 4), the bins are most enriched in all-beta (1.93) and all-alpha (1.27) proteins. Alpha/beta proteins are proportionally represented and none class proteins have reduced representation (0.23) in the top two bins. As discussed in an earlier section (2.5.5), all-alpha and few secondary structures classes generally give higher than an average R_S values, whereas all-beta and none class proteins perform below average in terms of R_S. Hence, one of the possibilities why bins 3 and 4 show lower R_S predictions, is because they are highly enriched in all-beta proteins. However, as it was mentioned previously, bins 3 and 4 (high TM-score ensembles) are under-represented in the set (38 protein in total) and some of the best scoring targets in terms of TM-score are outliers in terms of their backbone dynamics predictions (e.g. 1G6M discussed in section 2.5.2.1 and 1APS discussed in this section). Hence, the decrease of R_S values in bin 4 is unlikely to be a significant effect.

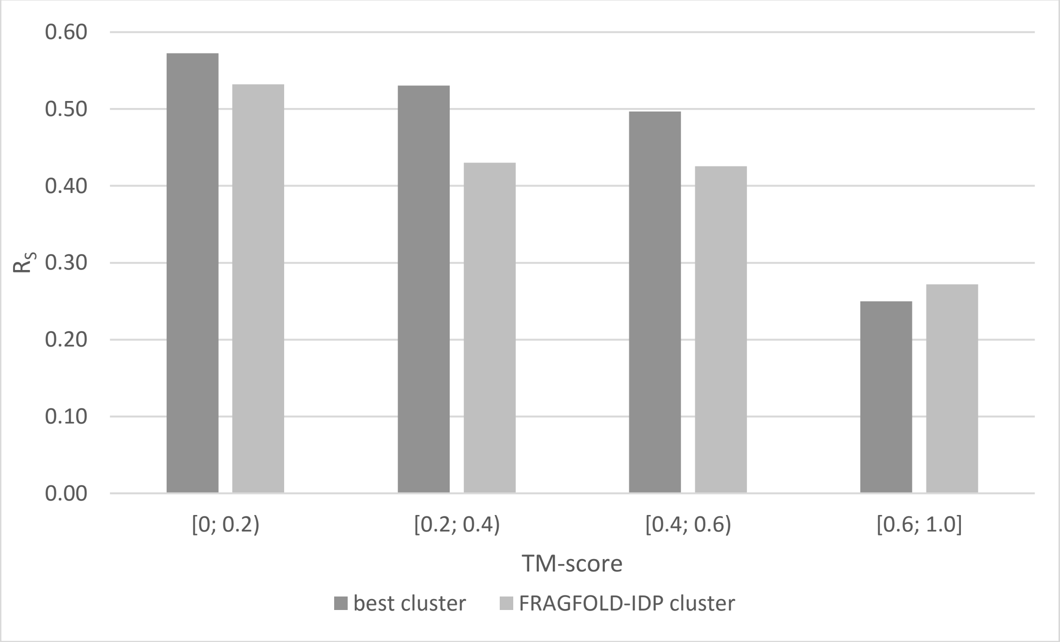


Figure 38. Backbone dynamics predictions quality (R_S) from TM-score binning.

These observations would point to the conclusion that in FRAGFOLD-IDP it is not necessary to find the correct fold of the protein in order to be able to predict its backbone dynamics accurately. From a computational perspective, it can be interpreted that in FRAGFOLD-IDP, during the folding process (i.e. FRAGFOLD simulations), only local conformations play an important role in the outcome of the calculations. Looking at the problem biologically, the results suggest that disordered regions form early in the folding process and the final conformation reached during folding does not significantly impact the disordered regions. Alternatively, it could speculated that disorder is an intrinsic local property of the sequence.

This finding could be related to other studies. During DynaMine optimisation it was found that using a wider sequence window as an input for the predictor increases the correlation between DynaMine predictions and reference experimental data (Cilia et al., 2013). However, the improvements are significant up to a window size of 23 (11 residues on either side of the residue of interest). As the authors themselves point out, the residues in the immediate neighbourhood have the greatest impact on the backbone dynamics. Hence, the conclusions from DynaMine also confirm the notion of the locality of intrinsic protein disorder.

This finding not only serves as an important observation in terms of expanding our understanding of the protein folding process, but it could also help the possible future development of FRAGFOLD-IDP directly. The computational time needed for simulating long sequences is substantial and increases exponentially with the length of the sequence (discussed in section 2.1.3). Since it is not necessary to find a correct fold for the sequence, overlapping sequence fragments could be simulated independently and then the disorder profiles assembled from the fragments. This could make long sequences accessible to FRAGFOLD-IDP simulations and could also reduce the computational time needed to obtain results.

2.5.8. *Comparison with other approaches to predict IDP ensembles*

For every newly developed computational method, it is desirable to make comparisons to other state-of-the-art approaches to determine how this method performs and what are its strengths and weaknesses in the spectrum of all computational techniques. In case of FRAGFOLD-IDP, the comparison is difficult, since as explained in the introduction to this chapter (section 2.1.2), the number of computational methods that attempt to predict protein backbone dynamics is limited, i.e. the only approach to predict protein backbone dynamics from sequence is DynaMine (sections 1.5.2.4 and 2.1.2.1). To increase the variety of computational methods, some other approaches that provide related information were also included, i.e. crystallographic B-factor predictors and disorder/order predictors which were also shown to contain information related to protein backbone dynamics (Daughdrill et al., 2011).

2.5.8.1. B-factor predictions

B-factors, also known as Debye-Waller or temperature factors, indicate the degree of electron density spread (Rupp, 2009). Therefore, B-factor indicates the static or dynamic mobility of an atom and is a function of atom displacement (Rupp, 2009). In an experimental setting, B-factor can also indicate possible X-ray structure errors and depend on the resolution of the crystal structure, crystal contacts and on the structure refinement procedures (Schlessinger and Rost, 2005; Tronrud, 1996). By definition B-factors are expressed as:

$$B = 8\pi^2 \left\langle u^2 \right\rangle ,$$

where u^2 is the unidirectional mean-square displacement.

Predictions of B-factor values from sequence attracted some attention, and several computational methods to perform such predictions have been developed over the years, e.g. PROFbval (Schlessinger and Rost, 2005; Schlessinger et al., 2006) or the work by Yuan and colleagues (Yuan et al., 2005) and Radivojac and colleagues (Radivojac et al., 2004).

For this comparison, PROFbval was used (Schlessinger and Rost, 2005; Schlessinger et al., 2006). It is a sequence-based machine learning method that attempts to predict B-factors from sequence. Starting with a query sequence, a PSI-BLAST sequence profile is generated first and then based on it, a HSSP profile is generated. The profile is used to run PROF predictions of secondary structure (PROFsec) and solvent accessibility (PROFacc) (Rost, 2005). The predictor is a set of 2 feed-forward neural networks that use the sequence profile, PROFsec and PROFacc predictions and global features (global secondary structure, solvent accessibility contents and sequence length) as inputs. One of the networks predicts B-factors for all residues, while the other network runs only the predictions for buried residues (depending on the PROFacc predictions). The method outputs both raw and normalized B-factors:

$$B_{norm} = \frac{B - \langle B \rangle}{\sigma}.$$

PROFbval was trained on a set of 1,513 non-redundant X-ray structures with resolution ≤ 2.5 Å (3-fold cross validation). It achieved a Pearson's correlation coefficient value of 0.44 between the experimental and predicted normalized B-factors on the entire dataset.

In the PROFbval evaluation, the results were compared to solvent accessibility predictions (PROFacc), which served as a baseline method. It was shown that PROFbval significantly improves over those predictions (Schlessinger and Rost, 2005).

The authors also compared their B-factor predictions to NMR order parameter values (S^2) for a single case (Schlessinger and Rost, 2005). The comparison concerned only a subset of (81 out of 149) residues for RNase H. PROFbval predictions were accurate for some regions of the protein, but overall the correlation between order parameters and predicted B-factors was lower than the correlation between experimental and predicted B-factors.

2.5.8.2. Disorder/order predictors

In one study, Daughdrill and colleagues showed that for a p53TAD domain disorder predictor (IUpred (Dosztányi et al., 2005b), VL-XT (Romero et al., 1997), VSL2B (Peng et al., 2006)) output values correlate with amide nitrogen and normalized hydrogen nuclear Overhauser effect (NHNOE) values (Daughdrill et al., 2011). The predictors achieved Pearson's r correlation values between 0.42 and 0.71 on a dataset of 6 homologs of p53TAD from different organisms. Overall, none of the methods provided outstanding results that would be consistently better than any other method. It is not surprising as the problem formulated by the authors is rather difficult. They compared 6 closely related proteins (between 42% and 91% pair-wise sequence identity) with different disorder profiles. Some targets proved to be more difficult than others, i.e. rabbit p53TAD produced an average r = 0.45, while cow and guinea pig r = 0.61. Overall, different disorder predictors performed similarly, achieving Pearson's r between 0.54 and 0.57.

Even though the study was limited and did not show robust results for any single disorder predictor, it strongly suggests that relating the values from disorder predictors to protein backbone dynamics is a viable option and should be explored.

2.5.8.3. Overall comparison

The comparison between FRAGFOLD-IDP, DynaMine, PROFbval, DISOPRED3 and IUpred was carried out on the 200 protein dataset introduced previously (section 2.2). DISOPRED3 and IUpred were selected to represent the disorder predictors as they reflect the two main approaches to disorder/order classification – machine learning-based (DISOPRED3) and statistical energy-based (IUpred; compare section 1.5.1).

For DynaMine, the predictions were run locally using the August 2014 version of DynaMine downloaded from the authors' website (`http://dynamine.ibsquare.be`). PROFbval was downloaded from the Debian repository. DISOPRED3.16 and IUpred were also downloaded and run locally.

The results of the predictions are presented below (Figure 39). The chart uses median R_S values for comparison, because it is a more robust metric than the average, especially in the presence of outliers. Overall, FRAGFOLD-IDP and DynaMine clearly perform best and significantly better than the naïve method (discussed in section 2.5.3). PROFbval predictions and IUpred achieve performance on par with the naïve approach. DISOPRED3 achieves higher median R_S than the naïve, but the result is not statistically significant (Wilcoxon signed-rank test p-value = 0.73).

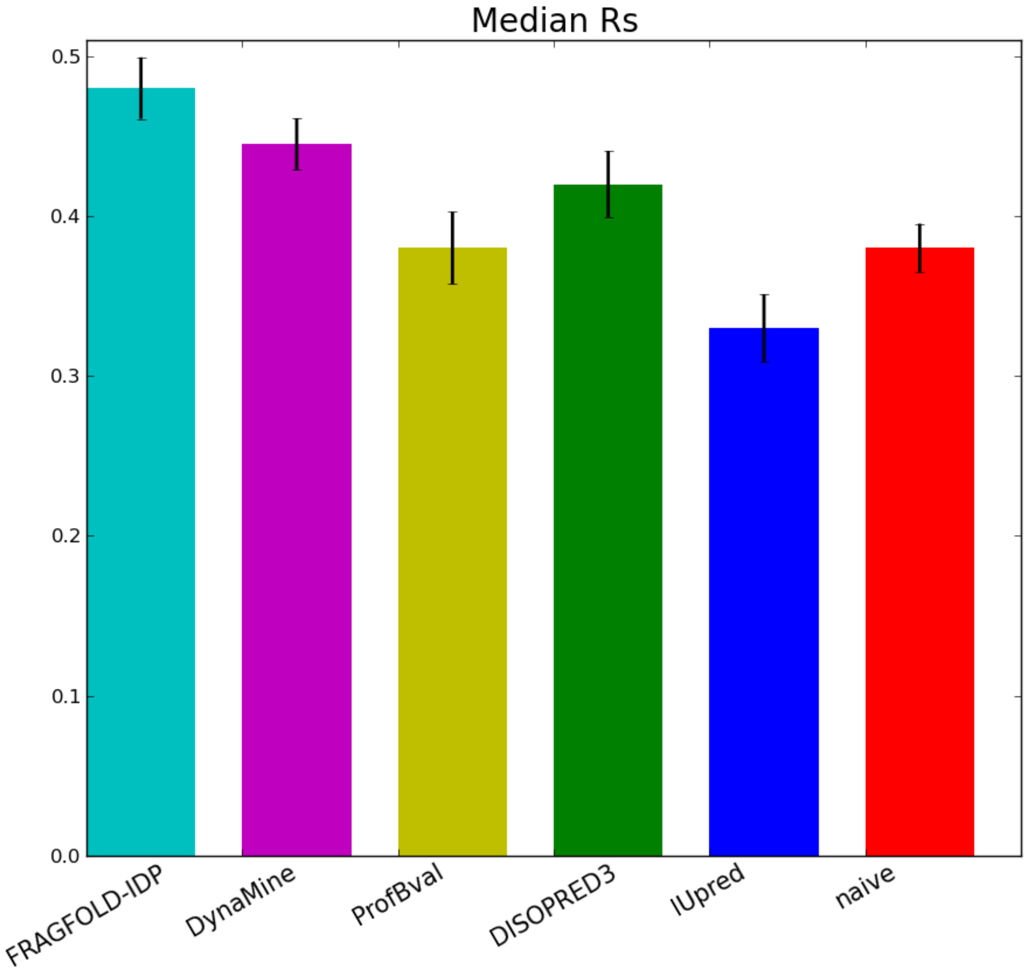


Figure 39. Median R_S values between FRAGFOLD-IDP and other computational techniques.

2.5.8.4. FRAGFOLD-IFP and DynaMine

Because the only computational techniques that achieve results significantly higher than the naïve approach are FRAGFOLD-IDP and DynaMine, let us compare the results of those methods in more detail.

The results of FRAGFOLD-IDP and DynaMine are comparable in terms of their overall performance. Median FRAGFOLD-IDP R_S is 0.48 (mean R_S = 0.44), whereas median DynaMine R_S is 0.45 (mean R_S = 0.44). Analysing the results on a per case basis, FRAGFOLD-IDP achieves higher R_S for 109 out of 200 cases (Figure 40). But more interestingly, the results of the two methods are very weakly correlated (r = 0.17, p-value = 0.013), even when FRAGFOLD-IDP outliers are removed (p-value goes up to 0.015).

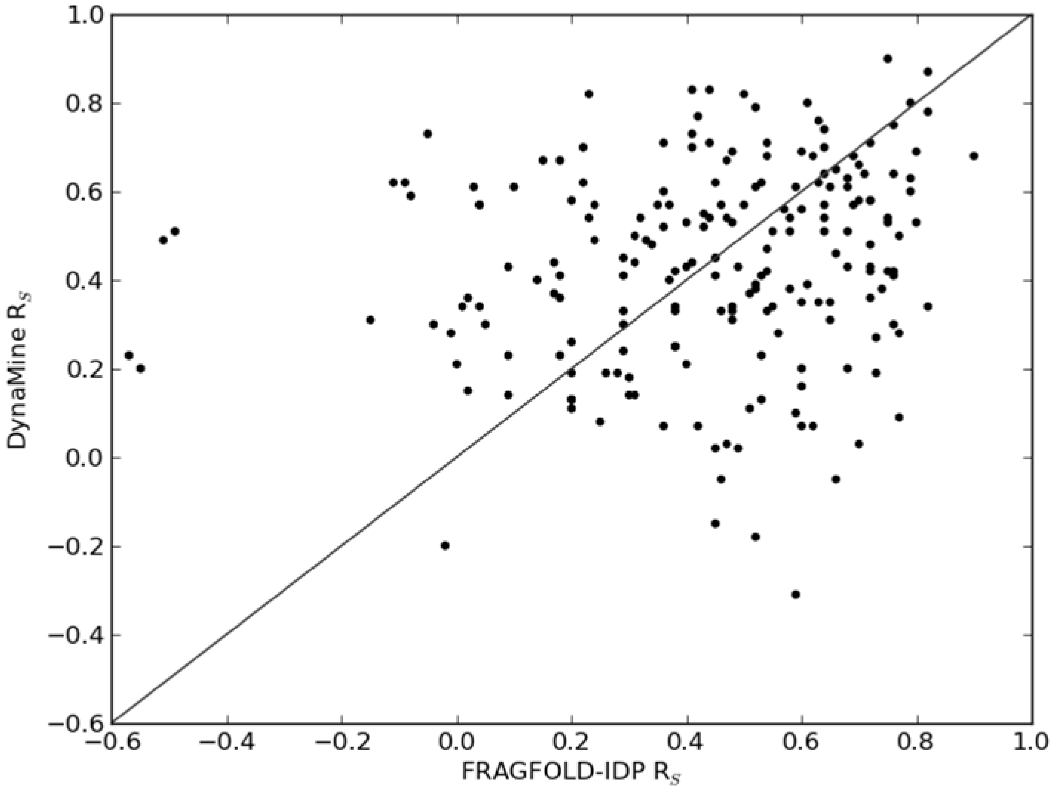


Figure 40. Per target comparison of FRAGFOLD-IDP and DynaMine results.

The lack of correlation (or very weak correlation) between the FRAGFOLD-IDP and DynaMine results suggests that the methods in a practical setting could complement one another. The results also suggest that for the most part poor results achieved by FRAGFOLD-IDP are not a cause of some experimental bias (apart from cases highlighted in section 2.5.2.1), but rather that FRAGFOLD-IDP is unable to cope with them effectively.

Orthogonality of FRAGFOLD-IDP and DynaMine is leveraged in Chapter 4, where a consensus protein backbone dynamics predictor is constructed and discussed. The predictor uses the outputs of these two methods as inputs to further improve the backbone dynamics predictions and take advantage of the strengths of both of FRAGFOLD-IDP and DynaMine.

2.6. Summary

This chapter introduced FRAGFOLD-IDP, a fragment-based method for *de novo* predictions of protein backbone dynamics from sequence. It addresses the problem of intrinsic protein disorder predictions by going beyond the binary order/disorder classification. Most of current computational techniques treat intrinsic disorder as a binary property (disorder classification), while the methods that go beyond that are computationally expensive and require a starting information (section 2.1). FRAGFOLD-IDP addresses this issues first by relying on a *de novo* method (FRAGFOLD) to be able to generate the ensembles of proteins of unknown structures and secondly, it predicts per-residue RMSD profiles which provide the information about backbone dynamics.

The method was benchmarked and tested on an exhaustive dataset of 200 protein structures solved by NMR and deposited in the PDB (section 2.2). FRAGFOLD-IDP optimisation included finding suitable parameters and algorithms for each step of the method (sections 2.3 and 2.4). During this process, it was found that the optimal parameters include using all FRAGFOLD potential terms to generate the raw ensemble of structures. Then, to extract the final ensemble, PFClust, a parameter-free clustering method was used. As a cluster selection criterion, cluster size proved to work best. Finally, FRAGFOLD-IDP and NMR ensembles were compared by first generating the disorder profiles using a sliding window superposition method and then quantifying the agreement between those profiles using Spearman's rank correlation (R_S).

FRAGFOLD-IDP was then evaluated on the 200 protein dataset (section 2.5). It was found that R_S values of ≥ 0.6 indicate good predictions (67 cases) and $R_S \geq 0.7$ indicate excellent predictions (35 cases). The method also performs significantly better than a naïve approach, which bases on secondary structure predictions and assumes that all loops are disordered, all sheets allow for some degree of flexibility, while all helices are rigid (section 2.5.3). Nevertheless, FRAGFOLD-IDP produced some outliers (section 2.5.2.1). Those targets were found to either have some factors affecting their backbone dynamics (e.g. disulphide bridges constraining the structure), or for some other targets the ensemble extraction criteria failed to extract the correct ensemble.

Comparing FRAGFOLD-IDP to other ensemble extraction methods, it was found that the method performs better than selecting median random cluster and comparably to selecting the best cluster from the PFClust pool, which proves that the ensemble selection criteria work well (section 2.5.4).

FRAGFOLD was originally designed to predict the structures of globular proteins. Hence it was interesting to check whether disorder content within studied proteins impacted the quality of protein backbone dynamics predictions (section 2.5.6). The majority (86%) of proteins within the dataset contain between 10% and 50% disordered residues, so this evaluation did not sample the whole spectrum of disorder. The results showed that for all-alpha and all-beta proteins the disorder content does not impact the predictions, but for alpha/beta and few secondary structures proteins an increase in disorder content seems to negatively impact the quality of FRAGFOLD-IDP predictions.

The quality of FRAGFOLD-IDP predictions appears to be independent of the quality of structure predictions (section 2.5.7), but CATH class of the protein does impact FRAGFOLD-IDP performance (section 2.5.5). Proteins belonging to all-alpha or few secondary structures classes perform better than an average, whereas all-beta proteins are more difficult to predict. These observations suggest that FRAGFOLD-IDP samples local protein conformations and that disorder is a local property of the protein chain. In all-beta proteins, local conformations are more dependent on the overall fold of the protein, since beta-sheets are non-local secondary structures. Hence, it is more difficult to predict protein backbone dynamics for this class of proteins.

Practically, the independence of structure and backbone dynamics prediction suggests that it could be feasible to run simultaneous FRAGFOLD-IDP simulations of fragments of proteins, instead of the entire protein chains. This could speed up the calculations and enable simulations of longer proteins than considered in this study (i.e. longer than 150 residues).

FRAGFOLD-IDP was also compared against a range of other computational techniques which use sequence information to predict protein backbone dynamics, or other qualities that could be related to it (section 2.5.8). The comparison included DynaMine – a method to predict NMR order parameters from sequence (sections 1.5.2.4 and 2.1.2.1); PROFbval – a sequence-based crystallographic B-factor predictor (section 2.5.8.1); IUpred and DISOPRED3 – disorder/order predictors (section 2.5.8.2). The assessment of those methods showed that only FRAGFOLD-IDP and DynaMine produce results that are significantly better than the naïve method.

Head-to-head comparison of FRAGFOLD-IDP and DynaMine showed that both methods achieve comparable results that slightly favour FAGFOLD-IDP (DynaMine median R_S = 0.45; FRAGFOLD-IDP median R_S = 0.48; FRAGFOLD-IDP produced higher R_S for 109 out of 200 proteins; section 2.5.8.4). More interestingly, the results showed no correlation between FRAGFOLD-IDP and DynaMine results. The methods seem to be mostly orthogonal in their predictions and could supplement each other well. This observation led to the idea of combining both methods into a consensus machine learning-based predictor. This approach is described in Chapter 4.

Overall, FRAGFOLD-IDP proved to be a state-of-the-art method that is able to accurately predict protein backbone dynamics from sequence. The only method on par with FRAGFOLD-IDP is DynaMine and the predictions of both of those methods are significantly better than the naïve approach.

This chapter introduced FRAGFOLD-IDP, described its optimisation and showed that is performs optimally within the designed workflow. It also described how well the method performs with regard to proteins of different classes, disorder content and FRAGFOLD-IDP predictions are independent of structure predictions quality.

Chapter 3.
DISORDER/ORDER CLASSIFICATION WITH FRAGFOLD-IDP

3.1. Background

Binary disorder/order classification is the most widespread computational technique for the analysis of intrinsically disordered proteins (described in section 1.5.1). The methods that perform sequence-based predictions are quick, can be routinely performed on many proteins and modern algorithms achieve high precision values (Monastyrskyy et al., 2014; Walsh et al., 2015).

The CASP experiment (Critical Assessment of the techniques for protein Structure Prediction) played an important role in the advancement of disorder classification methods. Predictions of disordered regions were included for the first time in CASP5 in 2002 (Moult et al., 2003). Considering that the seminal papers elucidating the role of protein intrinsic disorder appeared between 1999 and 2001 (Dunker et al., 2001; Uversky et al., 2000; Wright and Dyson, 1999) and the first disorder prediction algorithms were developed between 1997 and 1999 (Li et al., 1999; Romero et al., 1998, 1997), the structure prediction community quite quickly realized the importance of computational methods in the area of intrinsic protein disorder.

So far, the last disorder assessment was carried out in CASP10 in 2012 (Monastyrskyy et al., 2014). The main reason behind the halt in the CASP evaluation of disorder classification was the way that the experiment runs, accumulating the targets from experimental groups and structural genomics centres during the predictions season (approx. 6 months in which the CASP experiments is carried out). Most of the experimental groups either concentrate on ordered targets, leaving out targets that might pose experimental problems due to intrinsic disorder (ironically, often aided by disorder classification methods), or removing known disorder regions to help crystallization. Hence, the number of disordered targets available for assessment is low. Among disordered proteins, the disorder content is also low leaving relatively little space for objective assessment. The majority of the disordered regions in CASP are shorter than 10 residues and were mostly determined by X-ray crystallography (Monastyrskyy et al., 2014, 2011).

In CASP9 there were 26,075 residues (2,417 disordered) from 117 sequences in the disorder predictions assessment. In CASP10 there were 24,470 residues (1,664 disordered residues) from 94 sequences in the disorder assessment category.

Most of top disorder predictors were trained on X-ray data (Ishida and Kinoshita, 2007; Jones and Cozzetto, 2015; Walsh et al., 2012), or data derived from DisProt (Obradovic et al., 2003; Walsh et al., 2012). The physical characteristics of disordered regions solved by NMR are similar, but the sources of NMR-resolved proteins are usually different, i.e. more nuclear proteins, shorter, etc. (Vladimir N Uversky, 2013).

The first disorder/order classification method were rule-based and neural network predictors (Romero et al., 1997). The rules were based on the identification of stretches of order-promoting aromatic residues. Later disorder classification methods were mostly based on machine learning approaches, either as individual predictors or consensus methods (reviewed in section 1.5.1).

Although the disorder/order classification problem has been studied since 1997 and assessed in 5 CASP experiments, it is still an open problem in many respects. First, there is no universal definition of what are the experimental characteristics of disordered proteins (see section 1.3). Second, the experimental studies of intrinsically disordered proteins show bias to short regions of disorder, i.e. there is a shortage of information on long disordered regions (section 1.5.1). Finally, intrinsic protein disorder is an unstable phenomenon and various experimental factors (e.g. binding partners, inorganic ligands, experimental conditions) impact the behaviour of intrinsically disordered proteins observed experimentally (see section 1.2). All those arguments are true for both disorder/order classification and for the predictions of protein backbone dynamics.

3.1.1. *Problem formulation*

In the previous chapter, FRAGFOLD-IDP was introduced, a method to cope with the problem of predicting protein backbone dynamics *de novo* from sequence. This is a novel approach to computational studies of intrinsic disorder in proteins and there are not many methods available to make an exhaustive assessment of the approach (section 2.5.8).

Alternatively, the problem of predicting backbone dynamics can also be simplified to a binning problem. A borderline example of this procedure would be binning the predictions into 2 categories – ordered and disordered.

Binning FRAGFOLD-IDP predictions into 'ordered' and 'disordered' bins limits the amount of information that is produced using this method, since all of the dynamic information predicted by FRAGFOLD-IDP is lost. Nevertheless, it enables the comparison of the predictions to a wide variety of disorder predictors which have been studied and developed over the years (reviewed in 1.5.1).

Performing such comparisons can not only provide a broader outlook of FRAGFOLD-IDP performance, but also highlights its strengths and weaknesses. FRAGFOLD-IDP predictions are clearly more resource expensive than sequence-based disorder classifiers, but should FRAGFOLD-IDP prove to perform well in disorder/order classification, it can also be envisaged that the FRAGFOLD approach could be useful in cases where one is performing structure prediction, obtaining information about disordered regions alongside predicted protein structure. Therefore, despite a large computational cost that needs to be paid for in FRAGFOLD simulations, in comparison to classical machine learning-based disorder predictors, it can be advantageous to carry out an assessment of FRAGFOLD-IDP performance in binary disorder/order classification.

3.2. Methods

3.2.1. *Dataset*

The dataset used in this evaluation is identical to the one used and described in the previous chapter (section 2.2). It consists of 200 NMR PDB ensembles. Disorder/order annotations were also extracted from mobiDB and rely on the MOBI method, which uses a set of conditions to assign each residue from an NMR ensemble as ordered or disordered (compare section 1.3.4.6 and the discussion in sections 2.2.1 and 3.3.1).

3.2.2. *NMR ensemble disorder boundaries*

Annotations of NMR ensembles come from mobiDB and were determined using the MOBI method (Martin et al., 2010). Nevertheless, it is useful to verify how a single per-residue RMSD boundary from a sliding window superposition works on NMR PDB ensembles.

To determine the per-residue RMSD value boundaries between order and disorder, first each per-residue RMSD value for each protein in the dataset was separated into disorder and order groups according to the MOBI classification. Then the values were binned into percentile bins according to the distribution of values in the whole population. For each bin, the precision values of selecting disordered residues (ratio of disordered residues to all residues in the bin) were calculated and a logistic curve was fitted to the data. The value at 0.5 probability of the fitted curve corresponds to an optimal partitioning of the data into order and disorder sets on the basis of their per-residue RMSD (order/disorder threshold).

To minimize the impact of borderline cases (i.e. ordered residues between 2 disordered regions, or vice versa) several variations were also taken into consideration (Table 13). For example, only residues neighbouring at least one or more residue of the same class were considered (e.g. an ordered residue needs to be adjacent to at least one other ordered residue to be considered). The results proved to be robust regardless of the neighbourhood considered (Table 13). The results show a stable accuracy of 86% (0.87 F-score), hence the partitioning is not in perfect agreement with the MOBI annotations. In contrast with MOBI, no heuristics are used to augment the assignments and also MOBI method itself is not in perfect agreement with other disorder assignments (0.94 F-score compared to CASP8 annotations on 18 NMR structures) (Martin et al., 2010; Noivirt-Brik et al., 2009). Also, no single definition of intrinsic disorder in proteins exists (compare section 1.1.2), while the results here show a behaviour one would expect – with higher RMSD value, the probability of correctly annotating a disordered region increases (Table 13). Therefore, the partitioning should be considered as reliable.

Table 13. Order-disorder boundary of NMR ensembles.

RMSD at 0.5 probability	**order RMSD**		**disorder RMSD**						**min. neighbours**	
	mean	**median**	**mean**	**median**	**prec.**	**recall**	**acc.**	**F1-score**	**disorder**	**order**
1.25	0.68	0.48	1.99	2.03	0.81	0.76	0.86	0.78	0	0
0.94	0.58	0.46	1.98	2.03	0.90	0.84	0.86	0.87	1	1
0.95	0.58	0.46	1.99	2.04	0.90	0.84	0.86	0.87	2	1
0.94	0.58	0.46	1.98	2.03	0.90	0.84	0.86	0.87	1	2
0.94	0.58	0.46	1.99	2.04	0.90	0.84	0.86	0.87	2	2
0.97	0.58	0.46	1.97	2.01	0.89	0.82	0.85	0.85	no N- and C-term. (5 res.)	

RMSD at 0.5 probability indicate the RMSD threshold value achieved at 0.5 probability calculated from binning annotated ordered and disordered residues.

3.2.3. *FRAGFOLD-IDP as a disorder/order classifier*

To achieve the correct classification of FRAGFOLD-IDP results, a threshold needs to be established separating the residues annotated as disordered and ordered. The output of FRAGFOLD-IDP are per-residue RMSD values. The binning of FRAGFOLD-IDP output values was done using several approaches:

(1) raw FRAGFOLD-IDP results,

(2) a Savitzky-Golay filter,

(3) a median filter on the output results,

(4) a Gaussian filter.

Data filters were used to remove potential outliers in FRAGFOLD-IDP per-residue RMSD profiles.

Savitzky-Golay filter processes the data by fitting subsets of adjacent data points with a low-degree polynomial (2nd and 3rd order) using linear least squares.

3.2.4. ***Sequence-based disorder predictors***

The state-of-the-art in disorder classification can be established by considering previous CASP experiment results (Monastyrskyy et al., 2014). Some issues with the CASP assessment have already been mentioned (section 3.1), the main ones being small sample size, relatively short stretches of disordered residues and bias towards X-ray-solved structures. To complement the CASP10 assessment of disorder, the results from a systematic assessment of disorder classification by Walsh and colleagues were also included (Walsh et al., 2015). The paper is a part of the Mobility Continuous Assessment (MoCA; `http://moca.bio.unipd.it/`). In this study, the authors investigated the performance of disorder predictors on 25,833 sequences with disorder annotations from X-ray structures. Therefore, this study also is biased towards structures solved by crystallography, but the authors take a conservative approach and perform a majority vote if there are multiple PDB structures for a given sequence available. Another issue with the assessment is that the authors omitted some top performing predictors from CASP10 in favour of computational speed. Nevertheless, basing on both studies, a representation of high quality disorder predictors can be extracted.

For the evaluation of disorder/order classification, 9 methods (represented by 5 different algorithms) that are publicly available for download were selected. The methods are:

DISOPRED3 (Jones and Cozzetto, 2015) – the method ranked second in CASP10 assessment according to most metrics (precision, MCC, AUROC, the area under precision-recall curve (AUC (PR)), but was not included in the MoCA evaluation. DISOPRED3 uses sequence profiles as an input to a hybrid machine learning approach combining SVM, neural network and a nearest neighbour classifier within a neural network framework (described in detail in section 1.5.1.4.a). The method was trained on a concatenated dataset from DisProt and high-resolution non-redundant X-ray PDB structures with missing density or zero occupancy.

ESpritz (Walsh et al., 2012) – three flavours of the method were included: ESpritz-Xray, ESpritz-NMR and ESpritz-DisProt. The flavours indicate sources of training information. ESpritz is a bidirectional recurrent neural network method. In CASP10 ESpritz (consensus single method, without flavour differentiation) ranked in the top 15. In the MoCA assessment, none of the ESpritz flavours ranked at the top of the list, but performed very well on a per-residue basis in terms of accuracy and specificity. The arguments for including ESpritz in this evaluation are that ESpritz-NMR is one of the few methods trained exclusively on NMR data. Also, none of ESpritz flavours cluster closely with any other disorder predictors basing on segment overlap (SOV) scores in the MoCA evaluation, so including ESpritz adds extra information that is not closely related to any other disorder predictors.

IUpred (Dosztányi et al., 2005a) – two flavours of the method were included: IUpred-short and IUpred-long. It is a pair-wise energy-based method (see section 1.5.1.3.a). IUpred was not assessed in CASP10, but performed very well in MoCA evaluation. IUpred-short ranked highly on each measure and provided consistent results. In MoCA, IUpred-short results clustered closely with DisEMBL-465 predictions (Linding et al., 2003), hence only one of those methods is included in this evaluation. IUpred-long achieved worse performance than IUpred-short, but its results cluster with other methods not included in the evaluation (e.g. FoldIndex (Prilusky et al., 2005)).

VSL2 (Obradovic et al., 2005; Peng et al., 2006) – two flavours of the method were included: VSL2B and VSL2P. Both VSL2 predictors are SVM meta predictors, which combine predictors of short and long disordered regions into a single output. VSL2B uses 26 sequence-based features and single sequence as an input. It was assessed in Walsh evaluation. VSL2P on the other hand uses VSL2B features, but supplements them with additional features coming from a PSI-BLAST profile. Both methods were not assessed in CASP10, but VSL2B performed very well in the MoCA evaluation both on per-residue and per-protein levels. It does not cluster closely with any of the methods mentioned previously, but produces predictions similar to the RONN method (Yang et al., 2005).

DynaMine (Cilia et al., 2013) – the method is not a disorder/order classification method *per se*. As explained in the previous chapter and in the Introduction (sections 1.5.2.4 and 2.1.2.1), it attempts to predict NMR order parameter values from sequence. Nevertheless, section 2.5.8.4 showed that DynaMine achieves comparable performance to FRAGFOLD-IDP for the predictions of protein backbone dynamics. So it is reasonable to extend the comparison between those methods to the problem of disorder/order classification. Besides, in the original DynaMine paper, the method was also compared to a range of disorder/order classification methods and achieved high performance (more details in section 2.1.2.1). It wasn't assessed in the CASP or MoCA experiments.

The methodological details of the predictors included in this evaluation were described previously, in the Introduction in section 1.5.1. All methods were run locally using sequence data extracted from the PDB files. Default parameters were used for each method, unless README or instructions suggested otherwise.

Some other top performing methods were not included in the evaluation for several reasons. Some of the top methods are not publicly available, e.g. PrDOS-CNF, which ranked 1st according to AUROC and AUC (PR) in CASP10, or metaPrDOS2 from the same group which ranked in the top 10 in CASP10 according to MCC, AUROC and AUC (PR) – the method combines predictions of 5 other disorder classification servers. Some other methods were also excluded from the comparison, because they were shown to be produce predictions closely overlapping with other methods already included in the assessment (Walsh et al., 2015).

3.3. Results

3.3.1. *Impact of RMSD thresholding on disorder/order classification*

Using a single threshold to separate disorder from order may not be a perfect solution for disorder/order classification using FRAGFOLD-IDP. To account for potential outliers, some pre-processing filters were used (compare section 3.2.3). But to verify whether simple thresholding is sufficient, an evaluation on NMR data was made (as described in section 3.2.1).

Indeed, the results do not show a perfect agreement between MOBI annotations and disorder/order threshold (Table 13). Nevertheless, the accuracy and F1-score are high (above 85%). This shows that applying a single threshold to separate annotated disorder from order is a robust approach. Moreover, as discussed in section 3.2.1, the MOBI method in itself does not achieve 100% accuracy when compared to e.g. CASP annotations.

Therefore, for the purpose of the evaluation of disorder/order classification using FRAGFOLD-IDP, it may be concluded that using a single threshold as a boundary between disorder and order should be an appropriate and robust approach.

3.3.2. *Disorder/order classification on FRAGFOLD-IDP dataset*

Disorder/order classification was evaluated on the 200 NMR PDB proteins set (described previously in section 2.2). Disorder annotations were extracted from mobiDB and rely on the MOBI method (section 1.3.4.6; (Martin et al., 2010)). The results were assessed on the basis of a ROC curve, plotting false-positive rate (FPR = FP/(FP+TN); FP – false positive, TN – true negative) against true-positive rate (TPR = TP/(TP+FN); TP – true positive, FN – false negative) (Figure 41).

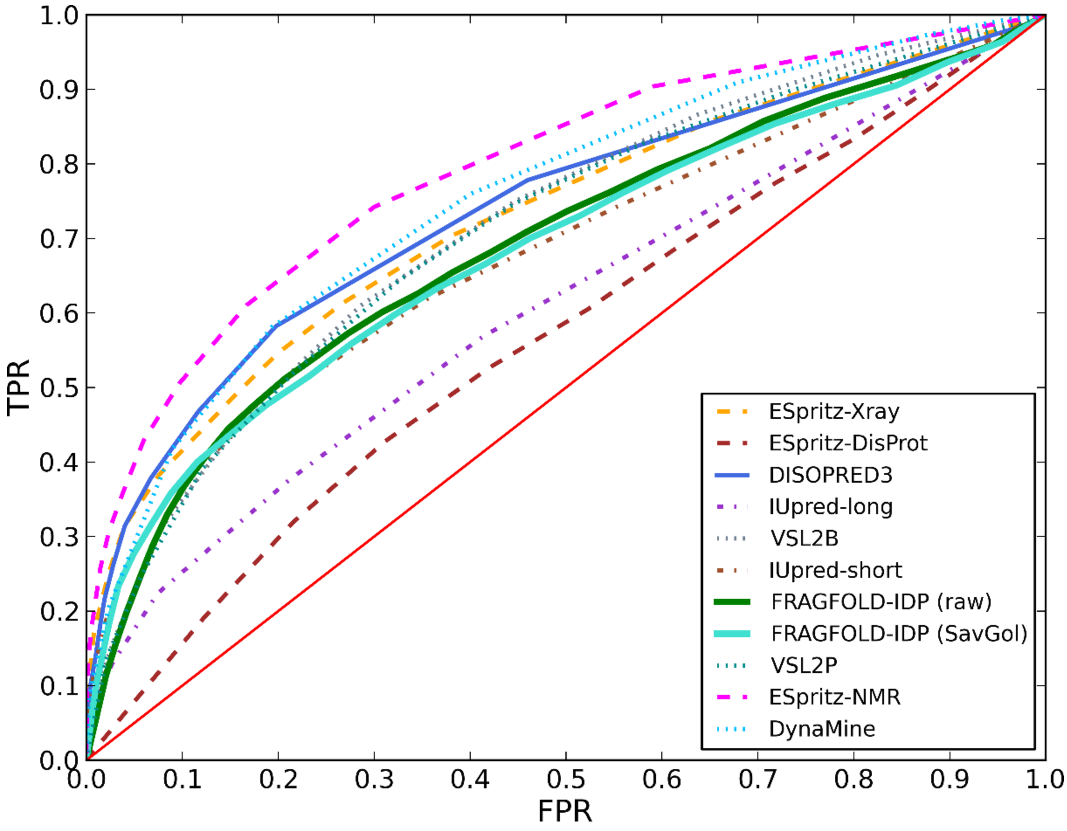


Figure 41. Disorder/order classification ROC curve.

ESpritz-NMR is clearly the top performer on the dataset with regard to AUROC and MCC results, but this should be expected since it is an advanced machine learning approach trained on NMR data (Figure 41, Table 14). The overlap of the evaluation dataset with ESpritz-NMR training data is substantial – 60 targets in the 200 NMR PDB dataset overlap with 2.187 targets in ESprtitz-NMR training set (data obtained from `http://protein.bio.unipd.it/espritz/`). This overlap is likely to increase the performance of the method, but it should not me a major concern, since the focus of this chapter is to evaluate how FRAGFOLD-IDP performs as a disorder/order classification method in the spectrum of other state-of-the art approaches.

The best method in terms of precision and specificity is DISOPRED3 (Figure 41, Table 14). The ROC curve for DISOPRED3 reached only up to about 46% FPR and had to be extrapolated from that to 100% TPR, 100% FPR. Notably, DISOPRED3 was trained on X-ray data and information extracted from DisProt (NMR and biophysical methods) (Jones and Cozzetto, 2015). So unlike ESpritz-NMR, or DynaMine it is less likely that the method was overtrained on NMR data, or that there is an overlap between NMR training data and the evaluation dataset here.

Table 14. Performance of the disorder/order classification methods and FRAGFOLD-IDP.

								rank			
method	**precision**	**recall**	**specificity**	**accuracy**	**F1-score**	**MCC**	**AUROC**	**precision**	**specificity**	**MCC**	**AUROC**
DISOPRED3	0.85	0.21	0.98	0.73	0.34	0.33	0.738	1	1	4	3
VSL2B	0.51	0.61	0.71	0.68	0.55	0.31	0.717	8	8	8	5
VSL2P	0.49	0.64	0.67	0.66	0.55	0.30	0.710	9	9	9	6
IUpred short	0.66	0.35	0.91	0.73	0.46	0.33	0.679	3	3	4	9
IUpred long	0.59	0.22	0.93	0.70	0.32	0.21	0.606	5	2	10	10
ESpritz Xray	0.62	0.47	0.86	0.74	0.54	0.36	0.728	4	5	2	4
ESpritz NMR	0.68	0.55	0.87	0.77	0.61	0.45	0.794	2	4	1	1
ESpritz DisProt	0.33	0.98	0.04	0.34	0.49	0.04	0.566	11	11	11	11
DynaMine	0.49	0.74	0.63	0.66	0.59	0.35	0.755	9	10	3	2
FF-IDP (raw)	0.56	0.48	0.82	0.71	0.52	0.32	0.689	7	7	6	8
FF-IDP (SavGol)	0.57	0.47	0.83	0.71	0.52	0.32	0.692	6	6	6	7

The highest and lowest result in each of the metrics were highlighted in green and red, respectively.

DynaMine performs very well in disorder classification. In the DynaMine paper (Cilia et al., 2013), 2 methods performed better than DynaMine on that task – PrDOS2 (not assessed here) and ESpritz-NMR (marginally better). Here, ESpritz-NMR is significantly better than DynaMine, but DISOPRED3 (not assessed in the DynaMine paper) also achieves comparable performance. However, the initial concern raised while discussing DynaMine results still stands (compare section 2.1.2.1). It is likely that DynaMine was overtrained on NMR data. DynaMine is also a non-specific method (Table 14). It achieves specificity only higher than ESpritz-DisProt.

ESpritz-Xray achieves high results according to most metrics (Table 14). However, it does not excel in any single one. Similarly to DISOPRED3, it is an example of a machine learning-based method that is unlikely to be overtrained on the data assessed here. In fact, ESpritz-Xray was trained exclusively on X-ray derived data (Walsh et al., 2012).

VSL2B and VSL2P go head in head with both FRAGFOLD-IDP approaches up to until 20% false positive rate (Figure 41). In MoCA evaluation VSL2B was the best method in terms of accuracy and recall (VSL2P was not a part of that evaluation (Walsh et al., 2015)). Here, VSL2B achieves medium level values for both recall and accuracy.

Overall, both methods perform similarly and rank in the middle of the list of evaluated methods.

IUpred-short achieves similar performance to FRAGFOLD-IDP across the whole FPR range (Figure 41). In the MoCA evaluation, IUpred-short was indicated as one of the top methods. Here, it also ranks highly in terms of precision, specificity and MCC (Table 14). Given that it is an energy-based method that does not use sophisticated machine learning machinery (as e.g. DISOPRED3 or ESpritz), the results achieved by IUpred-short are impressive.

Finally, IUpred-long and ESpritz-DisProt are significantly worse than all other assessed methods (Figure 41). IUpred-long is the most specific of the assessed methods, but it comes at a great cost in terms of recall. Hence, low F1-score, MCC and AUROC results (Table 14). ESpritz-DisProt, on the other hand, is greatly overpredicting the amount of disorder within proteins. For 176 proteins in the dataset ESpritz-DisProt predictions indicate 100% disorder. This results in the highest recall among all methods, but gives only 0.04 specificity.

In this comparison, FRAGFOLD-IDP does not stand out in any of the metrics (Table 14). FRAGFOLD-IDP data pre-processed using Savitzky-Golay filter (SavGol) perform slightly better than the raw results (median and Gaussian filter achieve identical results as Savitzky-Golay filter; not shown here). In terms of the ROC curve, FRAGFOLD-IDP results lie close to the IUpred-short results (Figure 41). Using other metrics, FRAGFOLD-IDP (both raw and SavGol) competes closely with both flavours of the VSL2 predictor. Overall, FRAGFOLD-IDP shows that it works well as a disorder/order classification method, and it achieves results on par with some state-of-the art disorder classification methods. The raw results of FRAGFOLD-IDP are also comparable to FRAGFOLD-IDP with Savitzky-Golay filter and data pre-processing does not improve the outcome of the classification significantly. This shows that FRAGFOLD-IDP on its own, using a single disorder/order separation threshold, can serve as an effective disorder classification method.

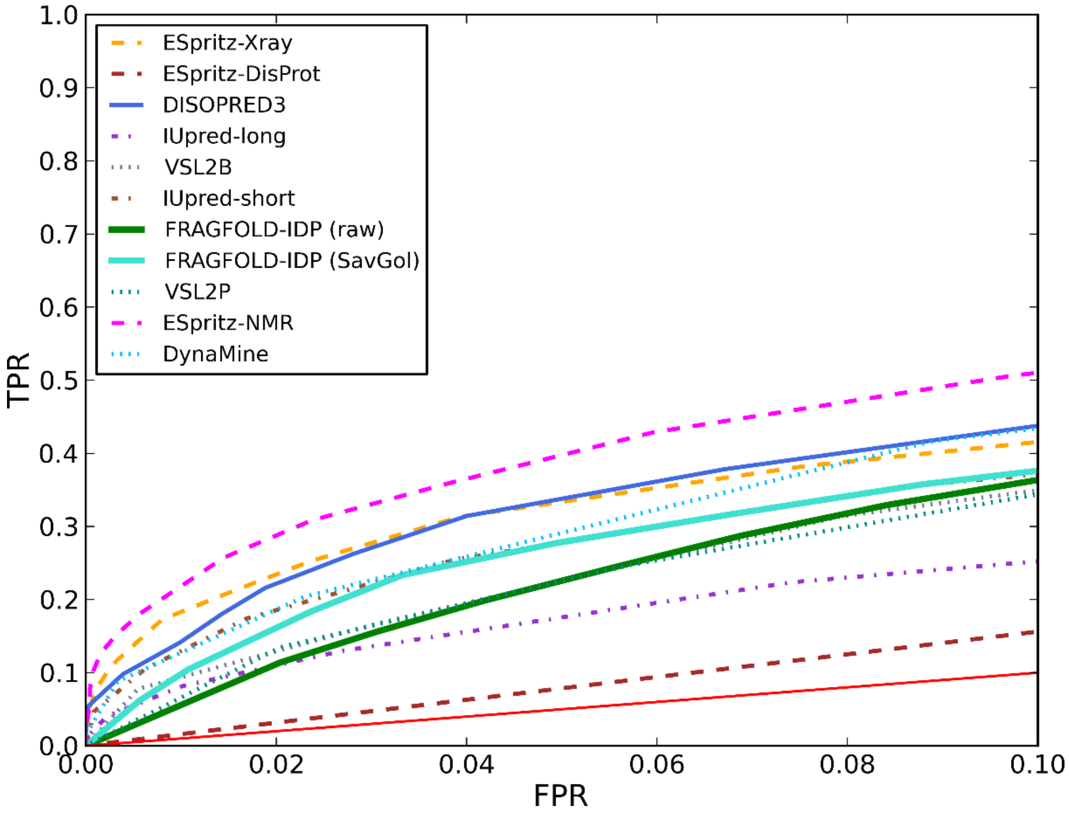


Figure 42. Disorder/order classification ROC curve up to 10% false positive rate.

To gain more insight into the results, studying the ROC curve up to 0.10 FPR is helpful (Figure 42). In this region, the differences between FRAGFOLD-IDP raw and SavGol are more apparent. The use of a Savitzky-Golay filter improves the predictions quite visibly and at around 0.03 FPR the performance of FRAGFOLD-IDP SavGol is higher than ESpritz-DisProt, IUpred-long, VSL2B and VSL2P, and on par with DynaMine (around 0.20 TPR). At around 0.10 FPR the two versions of FRAGFOLD-IDP converge and achieve approx. 0.37 TPR. In this region, FRAGFOLD-IDP is a top 5 method in the evaluation.

Also, two ESpritz flavours (NMR and Xray) perform very well at a low FPR range up to 0.02 FPR. In the region up to 0.10 FPR, DISOPRED3 also performs better than DynaMine. This confirms the high specificity of DISOPRED3. In contrast, DynaMine overtakes DISOPRED3 from 0.20 FPR onwards (compare Figure 41).

3.3.3. *Disorder/order classification of long disordered regions*

An outstanding problem in disorder/order classification is the prediction of long disordered regions. The typical definition of a long disordered region, is a continuous stretch of at least 20 residues annotated as disordered. One of the reasons why this is a problem, is the shortage of data on long disordered regions, but also some algorithmic challenges arise for methods that attempt to be universal, e.g. DISOPRED3 (compare the discussion in Jones and Cozzetto, 2015). Conversely, some algorithms were developed to treat specifically short or long disordered regions, i.e. IUpred-short and IUpred-long, or ESpritz-DisProt.

The MoCA evaluation (Walsh et al., 2015) also showed that some predictors perform better on short, rather than long disordered regions (e.g. ESpritz-Xray, ESpritz-NMR, or IUpred-short), while other predictors excel when predicting long disorder (e.g. VSL2B, RONN, IUPred-long).

The CASP10 evaluation confirmed the dependency of the quality of predictions on the length of disordered regions (Monastyrskyy et al., 2014). On average, CASP10 disorder predictions systematically decreased in terms of MCC with an increase in the length of disorder. The only method that had performed equally well on short and long disordered regions (up to 30 residues) using both MCC and AUROC measures was DISOPRED3.

Using the 200 NMR PDB dataset, an initial evaluation was carried out on proteins with an average disorder content of 33% (section 2.2). There are 60 proteins in this dataset which have long disordered regions (≥ 20 disordered residue segments), with 66 long disordered regions in total. It is a relatively small dataset, but comparing it to the CASP10 set of long disordered targets (20 cases; 2 from NMR), it makes sense to attempt a comparison and identify trends in the data. The results can indicate if FRAGFOLD-IDP performs better than other methods classifying regions of long disorder.

The assessment of long disordered regions was carried out as previously – on the basis of ROC curve analysis (Figure 43). Annotated regions shorter than 20 residues were ignored in the analysis, i.e. only long disordered and ordered regions were considered.

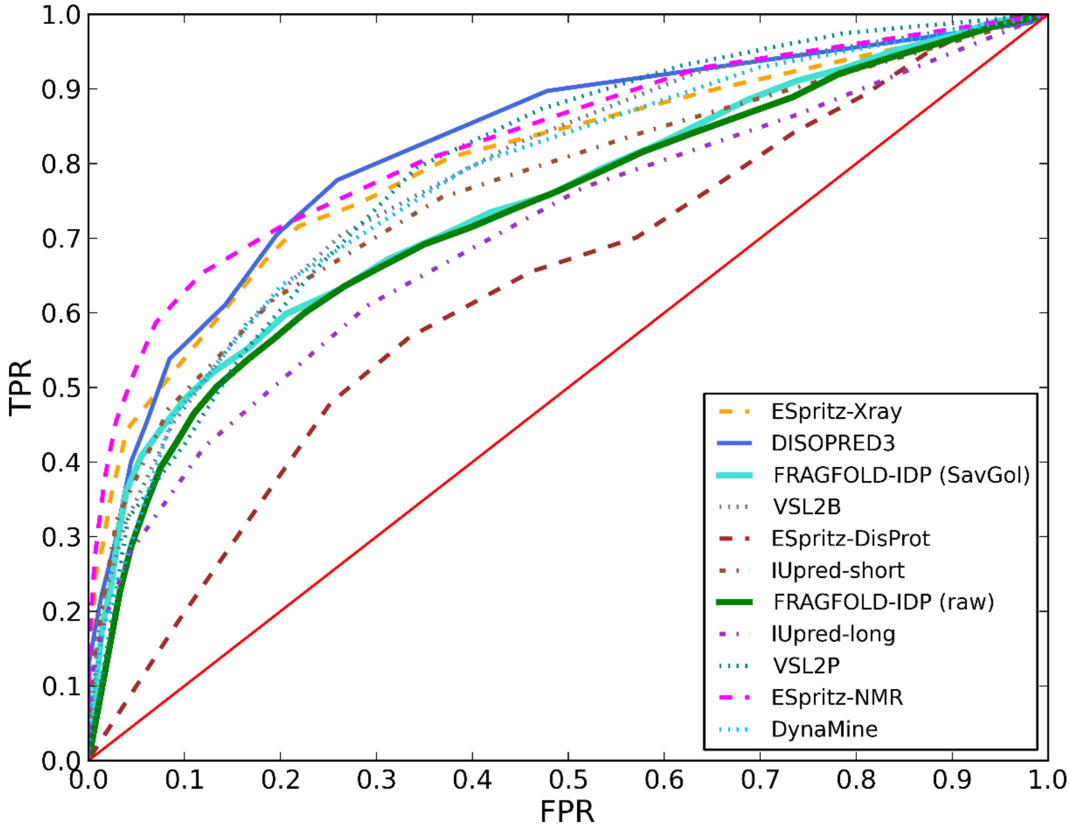


Figure 43. ROC curve of disorder/order classification of long disordered regions (≥ 20 residues).

In this analysis, up to approximately 0.30 FPR ESpritz-NMR and DISOPRED3 are the best predictors. At higher FPR values, VSL2P joins the former methods and also performs exceptionally. On average, IUpred-long achieves the greatest improvement in prediction quality, but it still performs worse than IUpred-short and is only better than ESpritz-DisProt. On the other hand, DynaMine achieves the smallest improvement in the quality of predictions compared to all disorder predictions. Similarly, both flavours of FRAGFOLD-IDP prediction do not improve greatly and overall the method loses some of its initial performance with respect to other methods.

However, notably, FRAGFOLD-IDP (Savitzky-Golay) performs very well in low FPR region (Figure 44). Using the data filter improves the predictions significantly over the raw results. Besides, at 0.10 FPR only ESpritz-NMR, DISOPRED3 and ESpritz-Xray achieve significantly higher TPR values than FRAGFOLD-IDP (SavGol), while IUpred-short, DynaMine and VSL2B achieve similar results.

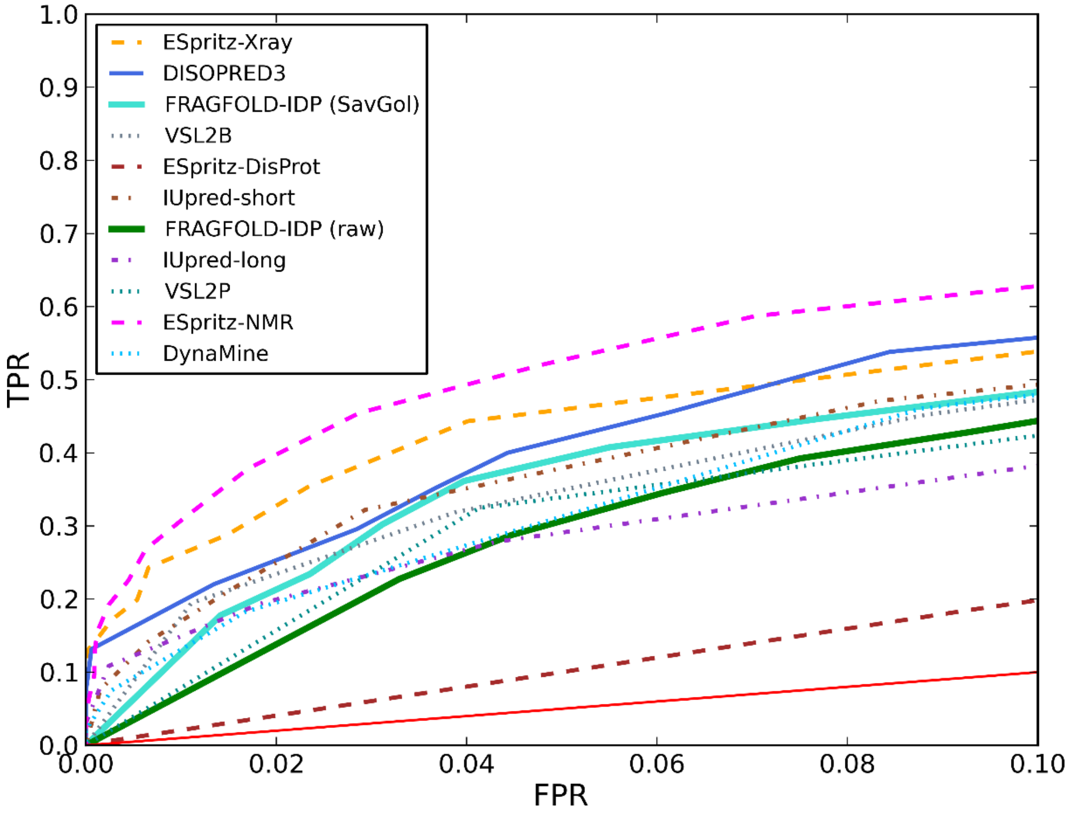


Figure 44. ROC curve of disorder/order classification of long disordered regions up to 10% FPR.

Overall, both flavours of FRAGFOLD-IDP show that the method is able to capture the majority of long disordered regions accurately. From around 0.25 FPR the differences between the two flavours of FRAGFOLD-IDP diminish and pre-filtering the data does not give any significant advantage (Figure 43).

3.3.4. *Relationship between disorder/order classification and backbone dynamics predictions*

The evaluation of disorder/order classification showed that FRAGFOLD-IDP is not the top performing protein classification method (sections 3.3.2 and 3.3.3). Nevertheless, Chapter 2 (section 2.5.8) showed that comparatively only FRAGFOLD-IDP and DynaMine are able to accurately predict protein backbone dynamics. The question is then, what does the evaluation in this chapter say about FRAGFOLD-IDP capability to accurately predict protein backbone dynamics?

First of all, backbone dynamics were assessed on the basis of R_S which is a relative metric, i.e. it assesses the relative signal along the protein backbone, not absolute values. In disorder/order classification an absolute threshold identical for all protein was applied. Even though FRAGFOLD-IDP uses a sliding window approach to decouple structure prediction quality from the predictions of backbone dynamics (section 2.3.4), some targets may have a relatively high background signal.

Also, in this evaluation FRAGFOLD-IDP attempted to reproduce the MOBI classification rather than the data coming from NMR PDB ensembles (compare discussion in section 3.3.1). The implication of this is that MOBI classification does not reproduce the actual disorder profile fully (Table 13).

Finally, disorder/order classification methods themselves were not designed to tackle the problem of protein backbone dynamics predictions. Their outputs do not correspond to per-residue RMSD values, but the thresholds (e.g. < 0.5 ordered, ≥ 0.5 disordered) were set to distinguish ordered residues from disordered. DynaMine showed that it is possible for a machine learning-based method to predict protein backbone dynamics well, but the method itself was trained on data related to protein backbone dynamics (sections 1.5.2.4, 2.1.2.1 and 2.5.8.4).

3.4. Summary

This chapter describes an evaluation of disorder/order classification using FRAGFOLD-IDP and a set of state-of-the-art disorder predictors, including DynaMine.

Disorder/order classification is a widely studied computational problem, which is still relevant today (section 3.1). The purpose behind FRAGFOLD-IDP is not to perform disorder/order classification, but to predict protein backbone dynamics. Still, it is a good test case to evaluate how well FRAGFOLD-IDP performs compared to a wide spectrum of computational methods. Especially, since there are far more methods available to tackle the task of disorder classification, rather than backbone dynamics predictions.

FRAGFOLD-IDP performs well on this task, achieving results on par with other state-of-the-art disorder classification methods (section 3.3.2). Performance-wise it ranks closely to IUpred-short and outperforms both VSL2B and VSL2P according to precision, specificity and MCC. All those approaches were among the top methods in recent MoCA evaluation (Walsh et al., 2015). Data pre-filtering (using either median, Gaussian, or Savitzky-Golay filters) improves the raw results of FRAGFOLD-IDP slightly, mostly in low FPR regions.

Looking at long disordered regions, FRAGFOLD-IDP on its own does not solve this problem to a greater extent than previously available methods (section 3.3.3). It shows good results in low FPR regions, but overall, the changes in predictions are weaker than of other methods included in the assessment here.

Overall, FRAGFOLD-IDP is an effective method for disorder/order classification, although it was designed to tackle the problem of predicting protein backbone dynamics. Computationally, it is not feasible to use FRAGFOLD-IDP as a routine disorder/order classification method, as it requires generating hundreds of protein models using FRAGFOLD. The procedure can take up to several hours on a computer cluster, while the use of standard disorder/order classifiers is limited by the time required to generate a sequence profile, for methods such as ESpritz or DISOPRED3, which usually take in the order of minutes on a desktop computer. Nevertheless, it may be feasible to use FRAGFOLD-IDP for disorder/order classification alongside protein structure predictions using FRAGFOLD, to provide an additional source of information of the disorder content within simulated sequences.

Chapter 4. CONSENSUS MACHINE LEARNING-BASED PREDICTIONS OF PROTEIN BACKBONE DYNAMICS

4.1. Background

4.1.1. *Summary of FRAGFOLD-IDP and DynaMine results*

Chapter 2 introduced FRAGFOLD-IDP and showed that it is an effective method for the predictions of protein backbone dynamics. The results achieved by FRAGFOLD-IDP are significantly better than those of a naïve approach (sections 2.5.3 and 2.5.8.3) and the only other method that achieves comparable performance in an assessment of protein backbone dynamics predictions is DynaMine (section 2.5.8.4).

Both mean and median R_S values achieved by FRAGFOLD-IDP and DynaMine across the 200 NMR PDB dataset are comparable – FRAGFOLD-IDP mean R_S = 0.44 (median R_S = 0.48); DynaMine mean R_S = 0.43 (median R_S = 0.44). Also, the predictions of both methods are largely orthogonal – correlation between FRAGFOLD-IDP and DynaMine predictions is low, r = 0.17 (section 2.5.8.4).

These observations suggest that it might be useful to combine the predictions of both methods to construct a consensus predictor which would elevate the strengths and decrease the weaknesses of FRAGFOLD-IDP and DynaMine.

4.1.2. *Meta predictors in other bioinformatics approaches*

Meta predictors, or consensus methods are one of the most essential and widely used tools in bioinformatics. They are especially popular for problems where a simple algorithmic solution is impossible or unknown.

Consensus methods are used in almost all domains of structural bioinformatics. One of the first applications of consensus prediction made its way to secondary structure predictions (e.g. JPred using the JNet algorithm (Cuff and Barton, 2000; Drozdetskiy et al., 2015)). Sometime later, protein structure prediction meta servers appeared. Those methods attempted to combine the wealth of publicly available structure prediction methods (fold recognition, homology modelling, sequence alignment and other) to obtain consensus predictions from amino acid sequence alone (e.g. Bioinfo.pl metaserver (now defunct; (Bujnicki et al., 2001)), the Genesilico metaserver (Kurowski, 2003), Pcons (Lundström et al., 2001), Pcons.net (Wallner et al., 2007)). Importantly, structure prediction meta servers not only attempt to produce consensus predictions from alternative methods, but they actually implement their own structure prediction pipelines that rely on several sequence search, fold recognition, etc. methods and unify their output formats to communicate effectively within the server infrastructure.

In a similar spirit, the most efficient modern residue-residue contact prediction methods rely on meta approaches, combining several sources of information within neural network or random forest frameworks (MetaPSICOV (Jones et al., 2015) and PConsC2 (Skwark et al., 2014)). For transmembrane proteins, some effective transmembrane topology predictors have also been developed (TOPCONS (Tsirigos et al., 2015)).

Finally, in disorder predictions consensus predictors are also popular (e.g. MetaDisorder (Kozlowski and Bujnicki, 2012), PrDOS-meta (Ishida and Kinoshita, 2007), MFDp (Mizianty et al., 2010); section 1.5.1.5). These predictors are one of the most effective approaches for disorder prediction (Monastyrskyy et al., 2014). In fact, even DISOPRED3 (described in sections 1.5.1.4.a and 3.2.4; Jones and Cozzetto, 2015), which was included in this work to evaluate its capability to predict protein backbone dynamics (in section 2.5.8) and compared to FRAGFOLD-IDP in disorder/order classification (in Chapter 3) is also a consensus predictor. It does not combine alternative external methods (i.e. developed by other groups, or standalone methods), but it does include 3 different predictors – a neural network, a support vector machine and a nearest neighbour classifier – that are combined together using another neural network.

4.1.3. *Consensus backbone dynamics predictor*

Protein backbone dynamics prediction is clearly one of the difficult and unresolved problems in bioinformatics (compare results in section 2.5.8). Therefore, it is desirable to try and combine the predictions of known methods to improve the quality of predictions. Since the known backbone dynamics predictors provide largely orthogonal results (as outlined in section 4.1.1), a machine learning framework should make it possible to achieve this goal.

This chapter introduces a novel consensus predictor, which combines the results of FRAGFOLD-IDP and DynaMine to produce improved protein backbone dynamics predictions. The predictor is based on a neural network architecture, which, as discussed above, have previously been utilized to successfully construct consensus predictors, significantly improving the predictions in other problems in bioinformatics.

4.2. Methods

The consensus protein backbone dynamics predictor was built using a neural network (also known as artificial neural network; ANN). It is a statistical learning model inspired by the nervous system (Figure 45). Neural networks are supervised learning methods, i.e. a network has to learn its parameters on a training set of known data before it can be used to carry out predictions on unknown data. They can be used to solve both regression and classification problems.

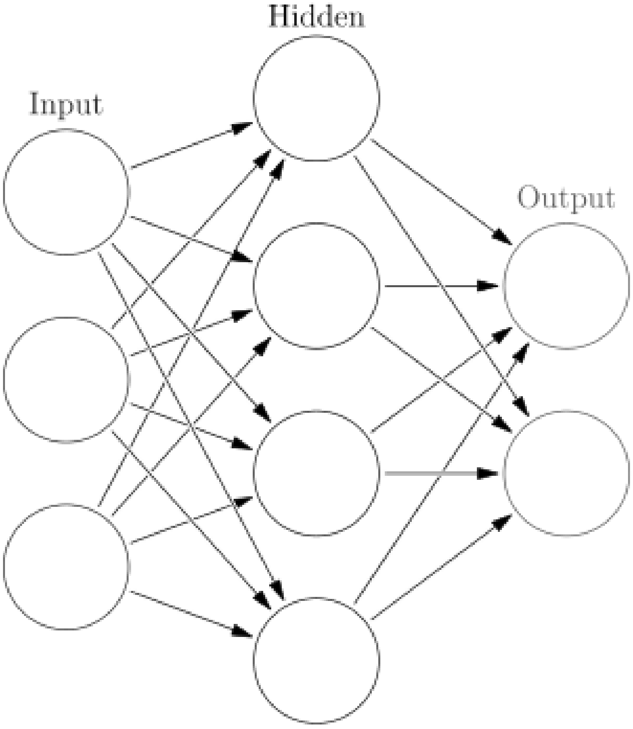


Figure 45. Sample neural network architecture. The network is composed of 1 input layer of 3 units, 1 hidden layer of 4 units and an output layer of 2 units. Source: en.wikipedia.org/wiki/Artificial_neural_network

Neural networks are common machine learning techniques used in bioinformatics (e.g. in secondary structure predictions, disorder predictions; compare sections 1.5.1.4 and 4.1.2). They are very popular in other areas as well, e.g. handwriting recognition, automated stock trading. Recently, neural networks have had their renaissance due to the use of deep neural networks (also known as deep learning), that show promise in solving difficult cognitive problems (e.g. self-driving cars, computer vision and speech recognition) and due to their ability to perform unsupervised learning (LeCun et al., 2015).

4.2.1. *Consensus predictor input features*

The consensus predictor combines the input methods (FRAGFOLD-IDP and DynaMine predictions) using a neural network architecture (Figure 46). The results of backbone dynamics prediction methods are not sufficient on their own to provide robust information for the network on how to combine the results. So aside the features from the two methods, some additional features were introduced (Table 15).

A good source of additional features is physicochemical information, represented by the amino acid composition of the query sequence. From the amino acid composition the network can infer information about hydrophobicity, disorder promoting residues and related information which needs not to be added separately to the network. Amino acid composition information is represented by frequencies of each residue or gap (21 features per residue) from the input multiple sequence alignment used as an input for FRAGFOLD (compare section 2.3.2).

Another source of information that should improve the predictions are secondary structure predictions (from PSIPRED (Jones, 1999)). It was shown before, that FRAGFOLD-IDP and DynaMine go beyond predicting only flexible loops within the proteins (compare sections 2.5.3 and 2.5.8.3), still the majority of disordered regions occur within loop regions (section 2.3.5.1). Secondary structure predictions are represented by 3 features corresponding to the probabilities of helix, strand and coil.

Also, because the network uses a sliding window method (see section 4.2.2), there is one extra feature per residue, indicating if the window reaches beyond the sequence (e.g. for a sliding window of size 9, the first residue in the sequence would have 4 missing residues).

Finally, features relating to global sequence parameters were introduced – sequence length and the distance of a predicted residue from the protein termini. These features should help the network to locate where in sequence the predictions are carried out and vary the predictions depending on the location, e.g. sequence termini are more often disordered (Monastyrskyy et al., 2014).

Table 15. Summary of consensus predictor per-residue features.

Input	Number of features
FRAGFOLD-IDP result	1
DynaMine result	1
Amino acid composition frequency	21
Secondary structure (PSIPRED)	3
Missing residue	1
Log sequence length	*1*
Log distance from termini	*2*

Window features (normal font); *global features (in italics)*

Multiple-sequence alignments and secondary structure predictions were used as inputs for FRAGFOLD to guide the selection of fragments (compare section 2.1.3). Otherwise, none of the consensus predictor input features were used in FRAGFOLD-IDP or DynaMine methods.

All of the input features were normalized to values of comparable order. DynaMine results are predicted order parameters, so they do not require normalization (i.e. are always between 0 and 1). FRAGFOLD-IDP values were normalized using a logistic squashing function. PSIPRED secondary structure predictions are probabilities of forming helix, strand and coil at a given position, hence they do not require normalization. Sequence length and the distance from termini were represented as a logarithm of the input values. The features are summarized in Table 15. Training output values (NMR per-residue RMSD values) were also subject to the logistic squashing function.

4.2.2. ***Consensus predictor architecture***

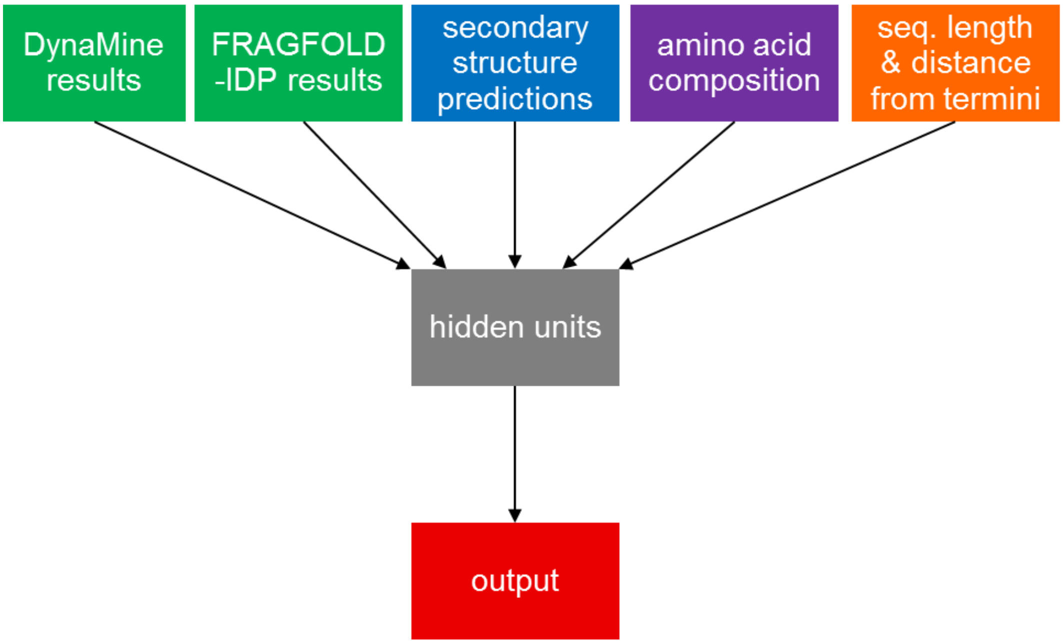


Figure 46. A schematic representation of the consensus backbone dynamics predictor.

The consensus predictor is a classical feed-forward neural network with a bias unit in input and hidden layers. A sliding window on input features is used. There are 27 window features and 3 global features (Table 15). Using a sliding window of 9 residues, there are 246 input features per residue. One hidden layer and a single output unit was used. A set of alternative numbers of hidden units were tested: between 10 and 200 hidden units.

Predictions are carried out for each residue in the input sequence. In the case of a 9 residue window, for each residue, 4 neighbouring residues to each side are also considered. Larger sliding windows were also tested (11 and 15 residues). The use of a sliding window is a common practice in bioinformatics machine learning methods (e.g. PSIPRED (Jones, 1999), MetaPSICOV (Jones et al., 2015)). The sliding window enables the network to consider not only the single position in question, but also the immediate environment of the residue.

4.2.3. *Training procedure*

The network was constructed and trained using the PyBrain Python library (`http://pybrain.org/`).

Because of the relatively small dataset size (200 proteins; section 2.2), the method was cross-validated, instead of creating separate training and test sets. To avoid overtaining, the cross-validation was performed on the basis of CATH classification (Orengo et al., 1997; Sillitoe et al., 2015), separating the proteins at the fold level. It is a rigorous criterion that ensures the proteins share no significant structural similarity, regardless of their disorder content (compare section 2.5.5). Some proteins in the dataset were not classified in CATH (45 cases). Those examples, for the purpose of cross-validation, were assigned to the CATH fold with which they share the highest similarity (lowest RMSD). All singletons were clustered together to form a separate class for cross-validation. The procedure resulted in 33 sets.

The training was performed on each class to minimize the mean squared error value. It was carried out until convergence with 20% of input data used for validation.

4.3. Results

4.3.1. *Network training*

The consensus predictor was optimised by training it on different window sizes and using a set of different architectures (number of hidden units).

First, the window size was optimised. Four window sizes were tested – 9, 11, 15 and 21 (Figure 47). All window sizes are odd numbers, because they consider equal number of residues on each side of the main residue. In this step of the optimization, two variations on the number of hidden units were considered – number of input features divided by two (Figure 47A and B) and a geometric mean between the number of input and output features (Figure 47C and D). This was done to ensure that for varying window sizes the network has similar properties. The window size was evaluated using MSE (mean squared error) and R_S values.

The network shows no significant window size dependency on the quality of predictions, regardless of the number of hidden units, or scoring. Hence, the behaviour of the consensus predictor is substantially different to that of DynaMine (compare sections 1.5.2.4 and 2.1.2.1; Cilia *et al.*, 2013), where the authors observed a significant dependency of the predictions on the window size used (up to around 23-residue window). This behaviour of the consensus predictor is likely caused by the fact that the most important sources of information, i.e. DynaMine and FRAGFOLD-IDP results, were already extracted using a sliding window approach. Here, only a small window is necessary to account for the immediate sequence and physicochemical environment.

Because there is no strong dependency of the performance of the consensus predictor on the size of the sliding window, the smallest sliding window, 9 residues, was selected.

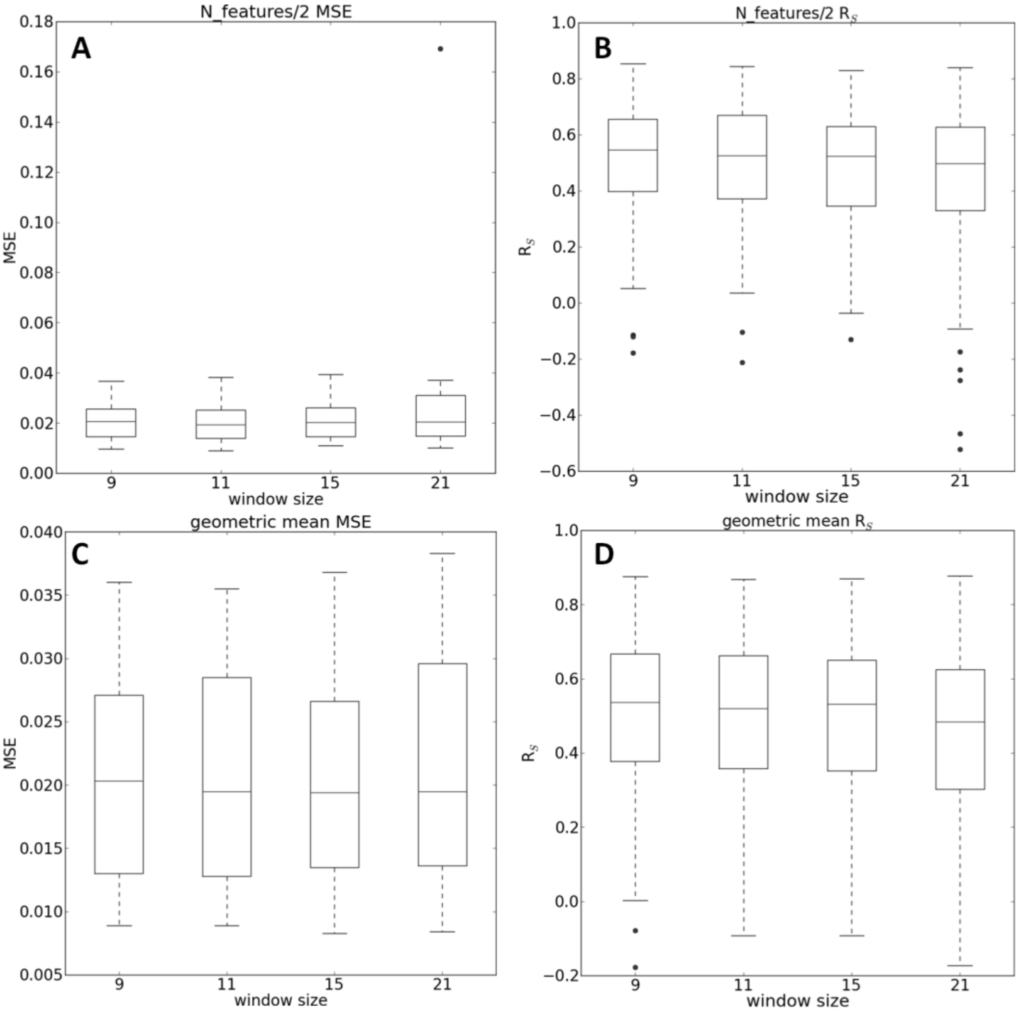


Figure 47. Optimisation of the consensus predictor. (A and B) optimisation of the window size using number of features/2 as the number of hidden units. (C and D) optimisation of the window size using geometric mean of the number of input and output units as the number of hidden units. Outliers are shown as red dots.

Having optimised the sliding window size, optimisation of the network architecture was performed (Figure 48). The assessment here concentrates on R$_S$ value distribution, as it is the metric which directly relates to the quality of results, which are assessed in this work. The network was trained on a number of different hidden units, ranging from 10 to 200. As in the case of optimising the window size, there is no significant dependency of the results on the number of hidden units. Hence, the criterion by which the final network was selected was to minimize the probability of over-training the network and 10 hidden units were selected as the optimal network size.

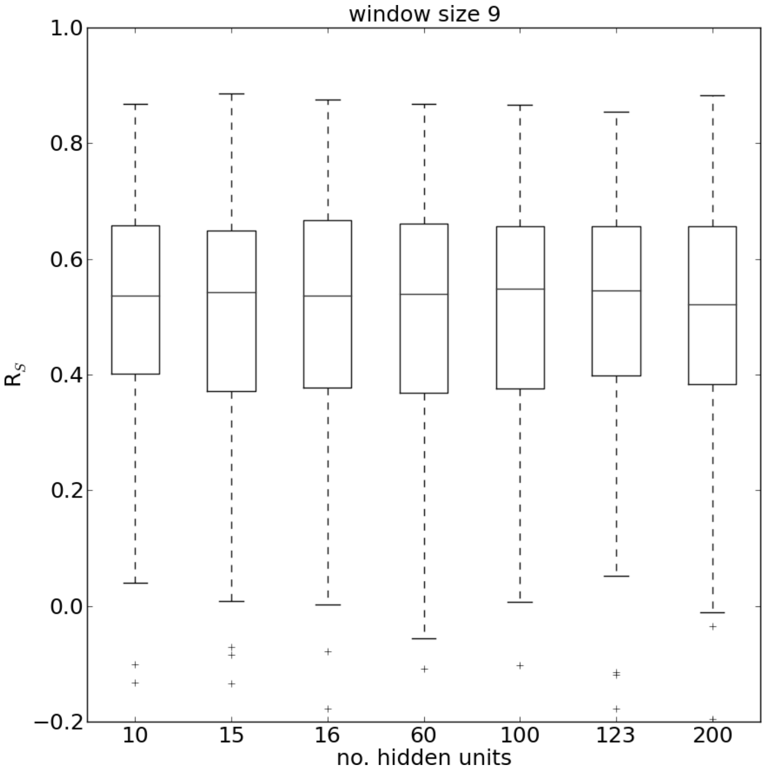


Figure 48. Optimisation of the number of hidden units in the consensus predictor.

4.3.2. *Consensus predictor results*

The results of the consensus predictor come from cross-validation performed as described in section 4.2.3. Comparing median R$_S$ values obtained on the 200 NMR PDB dataset, the consensus predictor quite clearly improves over both FRAGFOLD-IDP and DynaMine (Figure 49).

Results obtained by FRAGFOLD-IDP and DynaMine are similar (Figure 49). Although FRAGFOLD-IDP achieves higher median R$_S$ (0.48) than DynaMine (R$_S$ = 0.44), the differences are not significant (Wilcoxon signed-rank test p-value = 0.63). In comparison, the consensus predictor achieves median R$_S$ = 0.54 and those results are significantly better than both DynaMine and FRAGFOLD-IDP (Wilcoxon signed-rank test p-value < 0.001 for both methods).

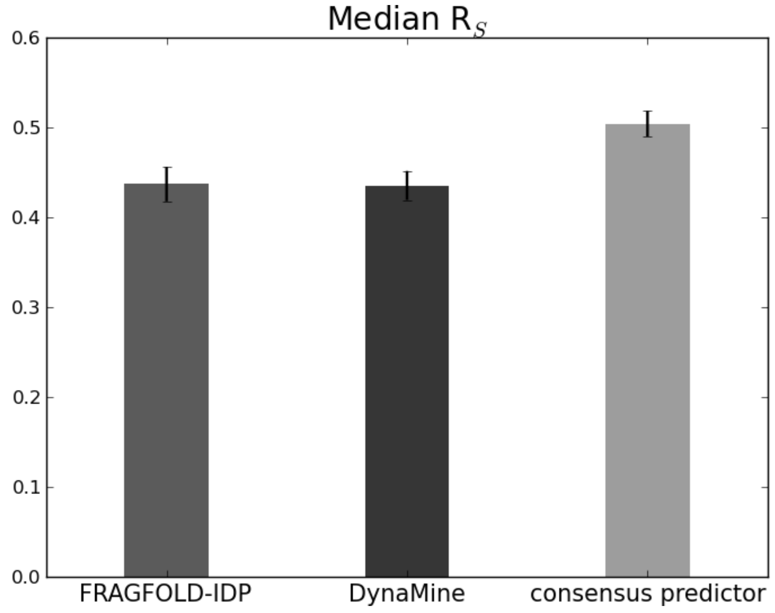


Figure 49. Comparison of FRAGFOLD-IDP, DynaMine and consensus predictor median R_S values.

Interestingly, the results of both input methods were not correlated (r = 0.17), but the results of the consensus predictor are correlated with both FRAGFOLD-IDP (r = 0.57) and DynaMine (r = 0.65). This shows that the consensus predictor was able to extract top results from both approaches, still significantly improving over any of them.

Also, looking at the number of ‘good’ ($R_S \geq 0.6$) and ‘excellent’ ($R_S \geq 0.7$) predictions (as in section 2.5.2), the consensus predictor performs well (Table 16). It significantly improves over both input methods in terms of the number of very good predictions ($R_S \geq 0.6$), achieving 77 such results. But in terms of excellent predictions ($R_S \geq 0.7$) it performs slightly worse than FRAGFOLD-IDP alone (30 in the consensus predictor and 35 in FRAGFOLD-IDP). The likely cause of the drop in the number of excellent predictions is the relatively large discrepancy in the number of FRAGFOLD-IDP and DynaMine predictions in this class (Table 16). Although the consensus predictor improves over both input methods, it is still constrained by the results provided by FRAGFOLD-IDP and DynaMine as inputs.

Table 16. Good and excellent predictions produced by the algorithms.

R_S	FRAGFOLD-IDP	DynaMine	consensus predictor
< 0	12	6	2
≥ 0.6	67	54	77
≥ 0.7	35	22	30

The predictions produced by the consensus predictor are also more conservative and there are only 2 cases with R_S below 0 (Table 16). Notably, the consensus predictor is able to remove all of the outliers produced by FRAGFOLD-IDP (compare section 2.5.2.1). However, the results obtained for some, i.e. 1G6M consensus R_S = 0.18 (Figure 25) and 1K0T consensus R_S = -0.10 (Figure 26) are still below the acceptable level for the reasons described previously.

A good example of a target where the consensus predictor works well, improving over both input methods and taking advantage of the strengths of both approaches is 1P94 (Figure 50). This target was already discussed in section 2.5.2, as an example of a medium quality FRAGFOLD-IDP prediction (Figure 23). The consensus predictor achieves an excellent result on this target (R_S = 0.87). FRAGFOLD-IDP (R_S = 0.54) correctly identifies part of the highly disordered N-terminal region (up to residue 15) and the ordered part of the protein between residues 48 and 76. DynaMine performs better (R_S = 0.71), but also fails to identify the behaviour of the protein in the highly disordered regions between residues 1 and 35. Also, the region between resides 25 and 35 is predicted to exhibit similar behaviour as the region between residues 50 and 60. Similarly, residues 1-5 and 70-76 show near identical behaviour, while the NMR ensemble shows that the N-terminus is highly disordered, and the C-terminus is ordered. The concerns about the behaviour of DynaMine partly stem from the fact that DynaMine predicts order parameters, not per-residue RMSD. The results shown in Figure 50 are scaled results of $1\text{-}S^2$ (DynaMine predictions). Due to scaling issues they might be more difficult to interpret visually, although the predictions achieve a high R_S score and should behave as described in section 2.3.6.

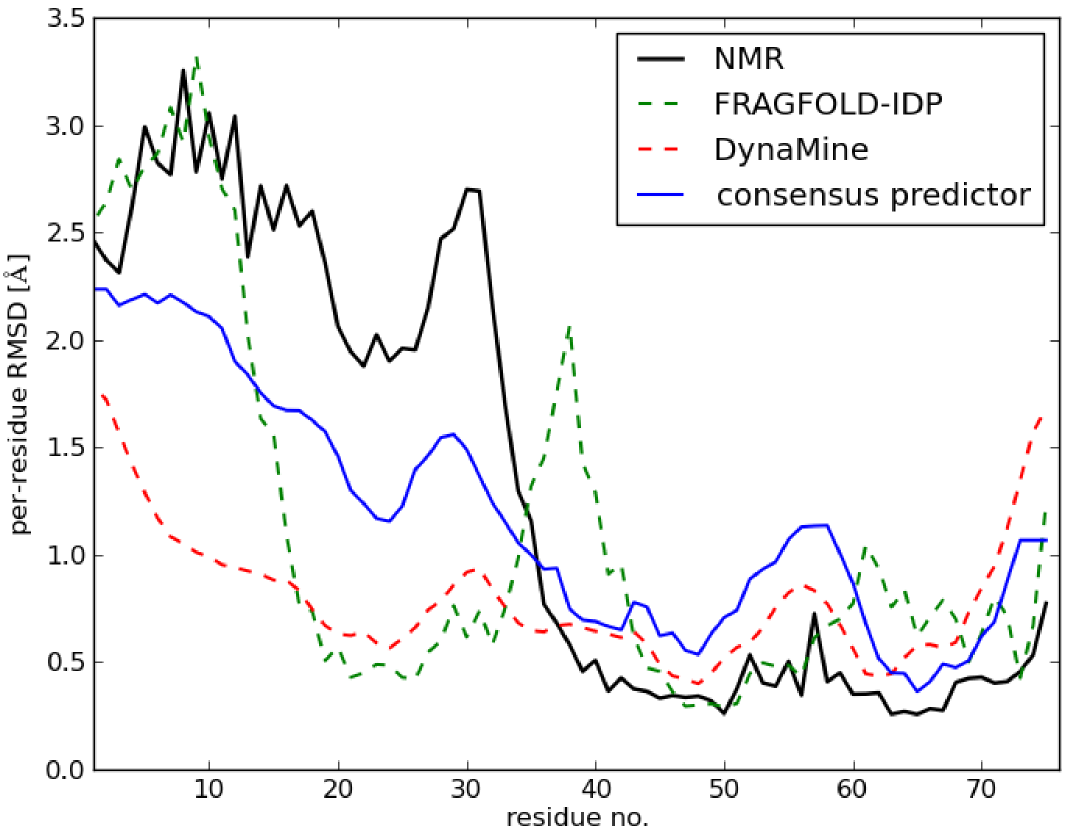


Figure 50. Example of an excellent consensus predictor result – 1P94. DynaMine (R_S = 0.71) and FRAGFOLD-IDP (R_S = 0.54) produce good predictions for this target. Consensus predictor performs remarkably well (R_S = 0.87).

The consensus predictor performs remarkably well on this target. Although the per-residue RMSD values do not match exactly (they were back-calculated from 0 to 1 values using an inverse logistic function), all of the features of the NMR ensemble are captured (Figure 50). The long disordered region between residues 1 and 35 is reproduced well – the consensus predictor values are highest in this region (i.e. higher than between residues 50 and 60, or in the C-terminus region). This includes the trough between residues 20 and 30 and the per-residue RMSD values which are higher between residues 1 and 10 than between residues 25 and 30. Also, the short region of elevated per-residue RMSD between residues 50 and 60 is reproduced accurately. Contrasting the predictions of the consensus predictor with those of FRAGFOLD-IDP and DynaMine it is clear that the predictor goes beyond simply combining the results of the input methods (Figure 50). For example, let us consider the region around residue 20, including the trough around residue 25. Both FRAGFOLD-IDP and DynaMine predict that the region around residue 25 has relatively low per-residue RMSD. But considering the immediate environment around this trough, both input methods over-predict its breadth, while the consensus predictor is able to correctly find the behaviour of the disorder profile between residues 20 and 25. Also, according to both input methods, the trough at residue 25 shows per-residue RMSD values lower than the region between residues 50 and 60. The consensus predictor is also able to rectify this mistake and correctly assign per-residue RMSD values as higher than between residues 50 and 60 (and above 70, where DynaMine and FRAGFOLD-IDP also fail).

4.4. Summary

This chapter introduced a consensus predictor which predicts protein backbone dynamics from sequence, combining the predictions of FRAGFOLD-IDP and DynaMine using a neural network.

FRAGFOLD-IDP and DynaMine produce predictions of comparable quality, but the results of those methods are weakly correlated (sections 4.1.1 and 4.3.2). This situation created a good opportunity to design a consensus predictor which would attempt to combine the predictions of those two input methods and try to maximize the strengths of each approach, at the same time minimizing their weaknesses.

Consensus predictors are common in almost every branch of bioinformatics (section 4.1.2). Their strength and popularity come from the fact that using machine learning techniques it is possible to effectively combine alternative methods that were designed to tackle distinct cases or trained on some specific datasets (e.g. consensus disorder classification methods trained on missing electron density and on NMR data, or trained to predict short and long disordered regions).

The consensus predictor developed here uses a feed-forward neural network to predict protein backbone dynamics from FRAGFOLD-IDP and DynaMine predictions (section 4.2). The predictor uses a 9-residue sliding window and 246 features per residue. The window features include: DynaMine predictions, FRAGFOLD-IDP predictions, amino acid composition and secondary structure probabilities. There are also some global features – sequence length and the distance of a residue from N- and C-terminus. All of the features were normalized to produce values of the same order of magnitude. The method was rigorously cross-validated using the 200 NMR PDB dataset introduced and evaluated previously (section 2.2).

The results of the consensus predictor are significantly better than both of the input methods (section 4.3). The predictor achieves a median $R_S = 0.54$ and is a more reliable approach than either of the input methods. There are only 2 predictions with $R_S < 0$ (the initial FRAGFOLD-IDP outliers (section 2.5.2.1) were removed by the consensus predictor) and the method is able to make very good predictions ($R_S \geq 0.6$) for 77 proteins (39% of cases).

In terms of the computational cost, the bottleneck of the consensus predictor are still FRAGFOLD-IDP predictions. Once trained, the neural network calculations are quick, but they require inputs coming from both methods, i.e. FRAGFOLD-IDP and DynaMine. Nevertheless, DynaMine alone is not able to produce predictions on par with the consensus predictor.

Finally, there is an issue about whether the consensus predictor developed in this chapter could be improved further. Including more features, even from methods marginally better than the naïve approach (discussed in section 2.5.8), could give some extra predictive power to the predictor. Also, only a limited subset of network architectures were tested. There are also more machine learning algorithms that could work well on this problem, e.g. Support Vector Regression, or a Random Forest. In machine learning there is no robust way to *a priori* determine which algorithm would work best for a given problem, e.g. the wealth of algorithms and approaches used for disorder/order classification (compare sections 1.5.1 and 3.2.4). Regardless, there is no other currently available computational method for the predictions of protein backbone dynamics that would provide comparable accuracy.

Chapter 5.
DISCUSSION AND PERSPECTIVES

5.1. Protein backbone dynamics predictions

The predictions of protein backbone dynamics are in their infancy. Intrinsic protein disorder is mostly treated as a binary property (a residue or a region in a protein is either ordered, or disordered) and the vast majority of computational methods used to study intrinsically disordered proteins are disorder/order classification methods (see section 1.5.1 and Chapter 3). Some simulation techniques were also used to study IDPs. These techniques are limited by the size of the protein and availability of a starting structure with which they can attempt to simulate disorder (section 1.5.2).

This work introduced two approaches to produce accurate backbone dynamics predictions *de novo* from sequence – FRAGFOLD-IDP (Chapter 2) and the consensus predictor combining the results of DynaMine, a machine learning-based predictor (sections 1.5.2.4, 2.1.2.1 and 2.5.8.4) with FRAGFOLD-IDP (Chapter 4). According to the analyses included in this work, these methods are current state-of-the-art and outperform all *de novo* computational methods. The directions presented here may be practically useful, as the consensus predictor is able to provide useful predictions for 39% of analysed cases. Some cases included in the dataset are difficult or impossible to predict, because of tertiary effects that were pinpointed (compare section 2.5.2.1), but were not excluded for the sake of fair comparison (section 2.2.1). This means that the consensus predictor should be able to provide useful predictions for about 50% of the general population of disordered proteins without further methodological improvements. The next generation of such methods could perform even better, but even with the current performance the methods presented in this work could help to solve some of the outstanding problems (discussed below in sections 5.3, 5.4 and 5.5).

5.2. FRAGFOLD-IDP as a disorder/order classifier

FRAGFOLD-IDP was compared to other methods in its capacity to predict protein backbone dynamics (section 2.5.8) and perform disorder/order classification (Chapter 3). Disorder/order classification is not the main purpose of this method, nor a major focus of this work, as it is argued that protein disorder is not a binary property and interpreting it in this fashion reduces the amount of information about this phenomenon (see section 2.1.1).

Nevertheless, disorder/order classification is the most popular computational technique used to study intrinsically disordered proteins and over the years a variety of methods have been developed to study this issue (compare sections 1.5.1 and 3.2.4). It was therefore desirable to place FRAGFOLD-IDP as a novel method in the context of well-established and thoroughly tested disorder/order classifiers.

Overall, FRAGFOLD-IDP is not better than the top disorder prediction methods, but its performance lies closely to some widely used methods and performs very well in the low FPR region (compare section 3.3). FRAGFOLD-IDP was developed to solve a different, although related problem, yet it performs well in comparison to the current state-of-the-art in disorder classification. Comparing FRAGFOLD-IDP performance on the task of disorder/order classification with its performance on the task of predicting protein backbone dynamics (sections 3.3 and 2.5.8) makes it clear that the performance in disorder/order classification is not correlated with the ability to predict accurate protein backbone dynamics (discussed in section 3.4).

5.3. Efficiency of FRAGFOLD-IDP method in predicting disordered ensembles

Predicting the ensembles of intrinsically disordered proteins explicitly is difficult (compare section 2.5.7). Not only does it require that one finds the correct fold of the protein, but one is also required to correctly predict its per-residue fluctuations. However, the method developed in this work, FRAGFOLD-IDP, shows that it is not necessary to find the correct fold of the protein to be able to predict its backbone dynamics accurately. This in turn suggests that intrinsic protein disorder is a local property of the polypeptide chain.

Unlike in the field of protein structure prediction (concentrated on ordered proteins), there is no consensus as to what are the criteria for a good or excellent backbone dynamics predictions. For the purpose of this work some intuitions were derived based on other works predicting per-residue fluctuations from known 3D structures (section 2.5.1). FRAGFOLD-IDP itself is the most effective single method for the prediction of protein backbone dynamics (section 2.5.8). It was able to provide good predictions for 33% of cases and excellent ones for 17.5% (section 2.5.2).

As mentioned above, what FRAGFOLD-IDP in fact predicts are not structural ensembles, but protein backbone dynamics represented by per-residue RMSD profiles. A step towards predicting actual disordered ensembles (as highlighted in section 5.1) could be made using contact predictions. It was shown that predicted residue-residue contacts improve the quality of *de novo* models up to 4-fold (Kosciolek and Jones, 2015, 2014). The structure prediction success rate, measured by TM-score (metric of structure prediction quality) in this work (discussed in section 2.5.7) is somewhat lower than expected comparing to a set of globular proteins (Kosciolek & Jones, 2014; compare section 2.5.7). So by taking advantage of residue-residue contact prediction, the quality of structure prediction in FRAGFOLD-IDP could be improved and ensembles of disordered proteins could be generated.

However, the role of predicted contacts in disordered regions is still unknown. One possibility is that they correspond to the folded (functional) state of the protein. In which case, it is possible that using contact predictions could actually produce folded states of disordered proteins (low per-residue RMSD values in the disorder profile; as in the case of matching FRAGFOLD fragments; section 2.5.7). Another possibility is that predicted contacts could correspond to the average conformation of the disordered region, as it was shown for conformational changes in proteins (Morcos et al., 2013). In this case, contact predictions should not have a substantial impact on predicting protein backbone dynamics in disordered regions.

Finally, there is a possibility that the coevolutionary signal in disordered regions is depleted. In general, intrinsically disordered regions in proteins show lower sequence conservation than ordered regions (Brown et al., 2011, 2010). But when subjected to random sequence mutations biased towards order promoting residues, disorder is not conserved, as opposed to secondary structure (Schaefer et al., 2010). Nevertheless, protein disorder itself is evolutionarily conserved (Schlessinger et al., 2011).

Those observations in tandem suggest that the hypothesis that there is no coevolutionary signal in disordered regions (or that it is depleted), is unlikely. However, should it prove to be true, contact predictions would help to fold ordered regions and the predictions of backbone dynamics should remain independent of structure predictions.

5.4. Biological significance of the current study

The first conclusion of general biological significance is the confirmation that the predictions of protein backbone dynamics are possible. Hence, this property is encoded in the protein sequence, similarly to disorder as a state (compare the discussion in section 2.1.1 and section 1.5).

Protein intrinsic disorder is a state related to protein function (as highlighted in sections 1.1.2 and 1.2). FRAGFOLD-IDP and the consensus predictor produce only information on protein backbone dynamics. However, as described in Chapter 3 (section 3.2.2) per-residue protein backbone dynamics can be related to annotated protein disorder with high accuracy.

Intrinsic protein disorder is not a binary property and not all conformational states are permitted in disordered ensembles (compare section 1.3.4). Disease-associated mutations need not cause disorder-to-order transitions (as highlighted below, in section 5.5; Vacic *et al.*, 2012; Uversky *et al.*, 2014). The majority of disease-associated mutations can be classified as disorder-to-disorder transitions, likely impacting the ability of the protein to interact with its binding partners, or changing the properties of the disordered ensemble. Therefore, going beyond the binary disorder-order classification is indispensable to be able to grasp the impact of those changes. Accurate predictions of protein backbone dynamics may open up the possibility to study the changes of the disordered state in response to external factors i.e. to perform disorder design (see section 5.5) and, in future, other biomedical applications such as the design of small molecules to alter the disordered state (Heller et al., 2015; Jin et al., 2013).

In more general terms, protein backbone dynamics predictions could be related to functional information in proteins. Such predictions could either serve as a source of information for protein function prediction methods, or be used to guide experiments aimed at investigating the structure and function of those proteins deemed likely to be disordered.

5.5. Disorder design

Disorder design is a term I would like to use to describe conformational transitions in IDPs (either disorder-to-order, or order-to-disorder) upon amino acid substitution, or via some other external factor (e.g. the binding of a small molecule). Similar to protein design, where methods are developed to try and predict sequences of proteins showing specific folds (Khoury et al., 2014; Kuhlman et al., 2003), in disorder design the aim is to take control over the phenomenon of protein intrinsic disorder. Being able to successfully perform disorder design would greatly improve the understanding of this phenomenon.

This understanding could benefit medical applications focused on diseases associated with intrinsically disordered proteins. It was argued, that missense mutations often impact disordered regions causing loss or gain of function effects and perturbations in protein interaction networks (Dembinski et al., 2014; Vacic and Iakoucheva, 2012; Vacic et al., 2012). The authors of these works also showed, based on disorder predictor results, that disorder-to-order transition mutations are enriched in disease, when compared to neutral evolutionary substitutions (Vacic et al., 2012).The utility of disorder design was also suggested to be an important source of information in facilitating the understanding of how natural variation in disordered regions affects the emergence of new phenotypes (Babu et al., 2012).

Disorder/order classification methods generally perform poorly on the disorder design task, i.e. using single-point mutations to trigger disorder-to-order, or order-to-disorder transitions (Ali et al., 2014). There is only some anecdotal evidence coming from single predictions on individual proteins using single sequence-based predictors that such design is possible (Dembinski et al., 2014; Vacic and Iakoucheva, 2012; Vacic et al., 2012). In general, poor performance of disorder predictors in disorder design is not surprising. Most of the methods use sequence profiles to perform the predictions (compare section 1.5.1) and point mutations do not impact the sequence profile significantly. Recently, some sequence-based machine learning methods were developed to tackle the problem of predicting disorder-to-order (and order-to-disorder) transitions upon mutation specifically (Ali et al., 2014; Anoosha et al., 2015). The first of these methods, PON-Diso (Ali et al., 2014), reported only aggregate results (combined disorder-to-disorder, order-to-order, etc.), omitting the comparison authors performed for disorder predictors. The work also relied on cross-validation based on a dataset of 101 cases from 31 proteins, making it likely that the method was in fact overtrained on the data, as no measures were taken to separate data coming from the same proteins. The second work reported an impressive average accuracy of 90% using an SVM-based model (Anoosha et al., 2015). The authors used the same dataset as in PON-Diso, but removed disorder-to-disorder and disorder-to-order cases, effectively erasing the most biologically relevant case of disorder-to-order transitions. Again, no measures were taken to ensure the method was not overtrained on the input data.

FRAGFOLD-IDP, somewhat similarly to other disorder predictors, also uses profile information. The method utilizes this information to perform secondary structure predictions and pre-select protein fragments from the FRAGFOLD library (compare sections 2.1.3 and 2.3.2). It therefore does not explicitly rely on profile information while performing the dynamics predictions. This shows promise in providing more reliable predictions for disorder design. It is certainly one of the attractive future directions that could be explored using FRAGFOLD-IDP. One of possible obstacles in doing so is the relative paucity of data. The largest known study to date used only 31 proteins (101 mutations) with only 3 cases of disorder-to-order transitions (Ali et al., 2014). An exploratory in-house dataset based on mobiDB (using only structural data, either from X-ray or NMR) is slightly larger and contains 7 cases coming from different proteins. This amount of data, regardless of possible FRAGFOLD-IDP performance, is unlikely to provide robust conclusions about the ability of computational methods to perform disorder design.

A problem related to protein design, which could be more computationally accessible is the design for protein dynamics. It is hypothesized that proteins are not only subject to selective pressure based on their structural properties, but also their local dynamic properties. An example of this could be DHFR protein family (Bhabha et al., 2013). *E. Coli* DFHR and human DHFR share significant structural similarity, but because of different dynamic properties it was shown that human DHFR cannot substitute its homolog in bacterial cells. So, subject to data availability, this “dynamics design” could be an interesting intermediate step towards disorder design. Here, more substantial sequence changes are observed which trigger some changes in protein disorder (dynamics) profiles. Therefore, this problem seems to be suitable for FRAGFOLD-IDP.

5.6. Future developments of FRAGFOLD-IDP

Some additional experiments to expand the analyses performed by FRAGFOLD-IDP should include the studies of sequences predicted by the method. It would be useful to determine, whether FRAGFOLD-IDP predictions are biased by low complexity regions, or bear some other sequence features that could help to discriminate easy and difficult targets for the method.

FRAGFOLD-IDP is a method that is able to predict protein backbone dynamics *de novo* from sequence. Performing broader comparisons including non-PDB data could be beneficial. Available NMR PDB data on disordered proteins are quite limited – the dataset used in this study consisted of all proteins that fulfilled the criteria established in section 2.2.1 which selected only 200 proteins. There are more proteins now, as the dataset used in Chapter 2 was assembled using mobiDB v. 1.2, the current version of mobiDB is 2.2. There are also longer (than 150 residues) proteins which could be explored. The authors of DynaMine paper used over 2,000 proteins (chemical shifts data) in their study (Cilia et al., 2013). An extensive comparison of the relationship between structure prediction and backbone dynamics predictions quality was also performed and it was shown that the correct fold is not necessary to obtain high quality backbone dynamics predictions (section 2.5.7). Hence, protein tertiary structure information is not necessary to evaluate FRAGFOLD-IDP further. Data such as chemical shifts (CS), or experimental NMR order parameters (S^2) could be used to broaden the spectrum of investigations. But this has some drawbacks – the RCI method (Berjanskii and Wishart, 2008), translating chemical shifts to order parameters shows a relatively low correlation between CS and S^2 ($r = 0.685$). On the other hand, the success of DynaMine shows that using this source of data displays great promise (Cilia *et al.*, 2013; compare sections 1.5.2.4, 2.1.2.1 and 2.5.8.4).

Another possible development was already highlighted previously – to study whether using sequence fragments, instead of the entire sequences would impact FRAGFOLD-IDP predictions (compare sections 2.5.7 and 2.6). The independence of structure and protein backbone dynamics predictions suggests that the FRAGFOLD-IDP approach samples only local protein backbone conformations. Therefore, using only sequence fragments should be sufficient to obtain all of the necessary information. Being able to simulate fragments would speed-up the calculations, as simulation time grows exponentially with chain length. This would make FRAGFOLD-IDP simulations more practical in comparison to methods such as DynaMine. The run time of the calculations is not the only possible benefit of simulating sequence fragments – it should also be possible to simulate longer proteins using FRAGFOLD-IDP. At present, FRAGFOLD is able to fold proteins up to around 200 residues, because of both computational and complexity reasons (Jones, 2001; Kosciolek and Jones, 2014), but using sequence fragments would allow larger sequences to be handled.

Finally, simulating protein fragments could allow FRAGFOLD-IDP to treat protein termini and mid-sequence regions separately. Although FRAGFOLD-IDP, unlike DynaMine, does not seem to overestimate the amount of disorder at the protein termini (compare Figure 50 and Cilia *et al.*, 2013, including Supplementary Information), it is a quite common practice to treat disordered regions at the termini separately to mid-sequence disordered regions (e.g. VSL2 predictor; Peng *et al.*, 2006).

FRAGFOLD-IDP predictions could also be used to obtain ensembles of intrinsically disordered proteins, instead of predicting protein backbone dynamics alone. For this purpose, intra-protein contact predictions could be used within FRAGFOLD, as discussed in section 5.3.

5.7. Limitations of studying disorder

The predictions of protein backbone dynamics add another dimension to our knowledge about proteins. It is an exciting area, but like with any computational approach, it is limited by the availability and reliability of the experimental data at hand. Disorder is a prevalent phenomenon that is notoriously difficult to grasp experimentally (Dyson and Wright, 2005). Several experimental techniques which are used to study ordered proteins largely fail when it comes to intrinsically disordered proteins (e.g. X-ray and EM). Further complicating the study is the observation the disorder a metastable state susceptible to the changes in the environment.

Since many intrinsically disordered proteins are responsible for regulation (Dunker and Uversky, 2008; Ward et al., 2004b), their behaviour is often controlled by post-translational modifications (PTM), such as phosphorylation (Bah et al., 2015). PTMs preferentially occur in intrinsically disordered regions (Theillet et al., 2014). They can cause disorder-to-order or order-to-disorder transitions and alter binding affinities.

Molecular crowding of the cellular environment can also impact the conformational ensembles of intrinsically disordered proteins, as it was proven by both NMR experiments (Cino et al., 2012) and MD simulations (Qin and Zhou, 2013).

Even in cases where intrinsically disordered proteins were treated in a binary disorder/order fashion, it was shown that the classification of residues can change upon environmental variations in the experimental conditions (Mohan et al., 2009). The experimental conditions included temperature, pH, salt concentrations, as well as, significant changes (disorder-to-order transitions) upon point mutations.

Hence, going beyond the binary classification makes the predictions even more susceptible to some subtle changes. This poses a major challenge on the interpretation of the data produced. At the same time, it shows that the selecting Spearman's rank correlation (R_S) as an evaluation metric was a good choice (compare sections 2.3.4.3 and 2.4.4), since it makes the evaluation less prone to minor changes in the disorder profiles.

5.8. Concluding remarks

With FRAGFOLD-IDP I show that the *de novo* predictions of protein backbone dynamics are possible and can be accurate. Using only sequence information, FRAGFOLD-IDP achieves state-of-the-art results that are significantly better than a naïve approach based on secondary structure predictions. On a related task of disorder/order classification, FRAGFOLD-IDP performs well, on par with sophisticated machine learning-based approaches designed to specifically to tackle the disorder/order classification problem. In this work I also introduce a neural network-based consensus predictor which combines FRAGFOLD-IDP and DynaMine predictions, along with a set of physicochemical features and secondary structure information, to produce significantly better results than any of the input methods. The consensus predictor generated good results, comparable to the methods which simulate protein backbone dynamics from structure, for 39% of the analysed cases.

The history shows that the study of proteins greatly benefits from the interplay of computational and experimental techniques. That was the case for the studies of ordered proteins, in protein design and disorder/order classification. To take the studies of IDPs further, we need more experimental data under a range of experimental conditions to help and evaluate the computational method better. At the same time, the computational techniques need to show their worth by guiding the experimental studies to spearhead new discoveries in the field.

APPENDICES

A. List of abbreviations

AUC (PR)	Area Under Precision-Recall Curve
AUROC	Area Under Receiver Operating Characteristic Curve
FPR	False Positive Rate
IDP	Intrinsically Disordered Protein
MCC	Matthews Correlation Coefficient
MD	Molecular Dynamics
MMC	Metropolis Monte Carlo
MoCA	Mobility Continuous Assessment
MoRF	Molecular Recognition Features
MSA	Multiple Sequence Alignment
PTM	Post-translational Modifications
REMC	Replica Exchange Monte Carlo
R_S	Spearman’s Rank Correlation
SA	Simulated Annealing
TPR	True Positive Rate

B. Software information

i. General

3D protein models were generated using PyMOL version 1.7 (www.pymol.org). Plots were generated using Python matplotlib library version 1.4.3 and MS Excel. All wrappers and scripts were generated using BASH and Pymol 2.7.X scripts

ii. Chapter 2

Structural ensembles were generated using FRAGFOLD version 4.62 (available on request from Professor David T. Jones (d.t.jones@ucl.ac.uk))

Structural similarities were calculated using TM-score (available for download from: zhanglab.umich.edu/TMscore)

Structural superposition methods used:

(1) ProFit (http://www.bioinf.org.uk/programs/profit/index.html)

(2) Theseus (http://theseus3d.org/)

Clustering algorithms used:

(1) TMclust, RMSDclust (available on request from Professor David T. Jones (d.t.jones@ucl.ac.uk))

(2) MaxCluster (http://www.sbg.bio.ic.ac.uk/~maxcluster/)

(3) PFClust (http://chemistry.st-andrews.ac.uk/staff/jbom/group/PFClust.zip)

(4) SPICKER (http://zhanglab.ccmb.med.umich.edu/SPICKER/)

Methods used for comparisons with FRAGFOLD-IDP:

(1) PROFbval (https://rostlab.org/owiki/index.php/PROFbval)

(2) DynaMine (http://dynamine.ibsquare.be/download/)

iii. Chapter 3

Disorder predictors:

(1) IUpred (`http://iupred.enzim.hu/`)

(2) DISOPRED3 (`http://bioinfadmin.cs.ucl.ac.uk/downloads/DISOPRED/`)

(3) ESpritz (obtained from the Authors; `http://protein.bio.unipd.it/download/`)

(4) VSL2 (`http://www.dabi.temple.edu/disprot/download/VSL2.tar.gz`)

iv. Chapter 4

Consensus predictor neural network was implemented in Python using PyBrain machine learning library version 0.3 (`http://pybrain.org/`)

REFERENCES

Abelev, G.I., 1971. Alpha-fetoprotein in ontogenesis and its association with malignant tumors. Adv. Cancer Res. 14, 295–358.

Ali, H., Urolagin, S., Gurarslan, Ö., Vihinen, M., 2014. Performance of protein disorder prediction programs on amino acid substitutions. Hum. Mutat. 35, 794–804.

Allison, J.R., Varnai, P., Dobson, C.M., Vendruscolo, M., 2009. Determination of the Free Energy Landscape of α-Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements. J. Am. Chem. Soc. 131, 18314–18326.

Anderson, C.W., Appella, E., 2003. Handbook of Cell Signaling, Handbook of Cell Signaling. Elsevier.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G., 2007. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36, D419–D425.

Anfinsen, C.B., Haber, E., Sela, M., White, F.H., 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 47, 1309–14.

Anoosha, P., Sakthivel, R., Gromiha, M.M., 2015. Prediction of protein disorder on amino acid substitutions. Anal. Biochem. 491, 18–22.

Arnone, A., Bier, C.J., Cotton, F.A., Day, V.W., Hazen, E.E., Richardson, D.C., Yonath, A., Richardson, J.S., 1971. A high resolution structure of an inhibitor complex of the extracellular nuclease of Staphylococcus aureus. I. Experimental procedures and chain tracing. J. Biol. Chem. 246, 2302–2316.

Babu, M.M., Kriwacki, R.W., Pappu, R. V., 2012. Versatility from Protein Disorder. Science (80-.). 337, 1460–1461.

Babu, M.M., van der Lee, R., de Groot, N.S., Gsponer, J., 2011. Intrinsically disordered proteins: regulation and disease. Curr. Opin. Struct. Biol. 21, 432–440.

Bah, A., Vernon, R.M., Siddiqui, Z., Krzeminski, M., Muhandiram, R., Zhao, C., Sonenberg, N., Kay, L.E., Forman-Kay, J.D., 2015. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. Nature 519, 106–109.

Baker, C.M., Best, R.B., 2013. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. Wiley Interdiscip. Rev. Comput. Mol. Sci. 4, 182–198.

Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B.J., Boone, C., Bader, G.D., Myers, C.L., Kim, P.M., 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 12, R14.

Berjanskii, M., Wishart, D.S., 2006. NMR: prediction of protein flexibility. Nat. Protoc. 1, 683–688.

Berjanskii, M. V, Wishart, D.S., 2008. Application of the random coil index to studying protein flexibility. J. Biomol. NMR 40, 31–48.

Berjanskii, M. V, Wishart, D.S., 2007. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic Acids Res. 35, W531–W537.

Berjanskii, M. V, Wishart, D.S., 2005. A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts. J. Am. Chem. Soc. 127, 14970–14971.

Bertini, I., Felli, I.C., Gonnelli, L., Kumar M., V., Pierattelli, R., 2011. 13 C Direct-Detection Biomolecular NMR Spectroscopy in Living Cells. Angew. Chemie Int. Ed. 50, 2339–2341.

Bhabha, G., Ekiert, D.C., Jennewein, M., Zmasek, C.M., Tuttle, L.M., Kroon, G., Dyson, H.J., Godzik, A., Wilson, I.A., Wright, P.E., 2013. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat. Struct. Mol. Biol. 20, 1243–9.

Bodart, J.-F., Wieruszeski, J.-M., Amniai, L., Leroy, A., Landrieu, I., Rousseau-Lescuyer, A., Vilain, J.-P., Lippens, G., 2008. NMR observation of Tau in Xenopus oocytes. J. Magn. Reson. 192, 252–257.

Brangwynne, C.P., Tompa, P., Pappu, R. V, 2015. Polymer Physics of Intracellular Phase Transitions. Nat. Phys. 11, 899–904.

Brown, C.J., Johnson, A.K., Daughdrill, G.W., 2010. Comparing models of evolution for ordered and disordered proteins. Mol. Biol. Evol. 27, 609–21.

Brown, C.J., Johnson, A.K., Dunker, A.K., Daughdrill, G.W., 2011. Evolution and disorder. Curr. Opin. Struct. Biol. 21, 441–6.

Brueschweiler, R., Wright, P.E., 1994. NMR Order Parameters of Biomolecules: A New Analytical Representation and Application to the Gaussian Axial Fluctuation Model. J. Am. Chem. Soc. 116, 8426–8427.

Bueren-Calabuig, J.A., Michel, J., 2015. Elucidation of Ligand-Dependent Modulation of Disorder-Order Transitions in the Oncoprotein MDM2. PLOS Comput. Biol. 11, e1004282.

Bujnicki, J.M., Elofsson, A., Fischer, D., Rychlewski, L., 2001. Structure prediction meta server. Bioinformatics 17, 750–751.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., Babu, M.M., 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell 46, 871–883.

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., Mornon, J.P., 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell. Mol. Life Sci. 53, 621–645.

Cavanagh, J., Fairbrother, W.J., Palmer, A.G., Skelton, N.J., Rance, M., 2007. Protein NMR Spectroscopy, Protein NMR Spectroscopy. Elsevier.

Chen, J., 2012. Towards the physical basis of how intrinsic disorder mediates protein function. Arch. Biochem. Biophys. 524, 123–31.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., 2010. MolProbity : all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D Biol. Crystallogr. 66, 12–21.

Chen, Y., Campbell, S.L., Dokholyan, N. V, 2007. Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection. Biophys. J. 93, 2300–2306.

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., Vranken, W.F., 2014. The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res. 42, W264–W270.

Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., Vranken, W.F., 2013. From protein sequence to dynamics and disorder with DynaMine. Nat. Commun. 4, 2741.

Cino, E.A., Karttunen, M., Choy, W.-Y., 2012. Effects of molecular crowding on the dynamics of intrinsically disordered proteins. PLoS One 7, e49876.

Cozzetto, D., Jones, D.T., 2013. The contribution of intrinsic disorder prediction to the elucidation of protein function. Curr. Opin. Struct. Biol. 23, 467–472.

Cuff, J. a, Barton, G.J., 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40, 502–511.

Cumberworth, A., Lamour, G., Babu, M.M., Gsponer, J., 2013. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem. J. 454, 361–369.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kähäri, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M.J., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., Flicek, P., 2014. Ensembl 2015. Nucleic Acids Res. 43, D662–9.

Dames, S.A., Aregger, R., Vajpai, N., Bernado, P., Blackledge, M., Grzesiek, S., 2006. Residual Dipolar Couplings in Short Peptides Reveal Systematic Conformational Preferences of Individual Amino Acids. J. Am. Chem. Soc. 128, 13508–13514.

Das, R.K., Crick, S.L., Pappu, R. V, 2012. N-Terminal Segments Modulate the α-Helical Propensities of the Intrinsically Disordered Basic Regions of bZIP Proteins. J. Mol. Biol. 416, 287–299.

Das, R.K., Pappu, R. V, 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc. Natl. Acad. Sci. U.S.A. 110, 13392–13397.

Daughdrill, G.W., Borcherds, W.M., Wu, H., 2011. Disorder Predictors Also Predict Backbone Dynamics for a Family of Disordered Proteins. PLoS One 6, e29207.

De Simone, A., Cavalli, A., Hsu, S.-T.D., Vranken, W., Vendruscolo, M., 2009. Accurate Random Coil Chemical Shifts from an Analysis of Loop Regions in Native States of Proteins. J. Am. Chem. Soc. 131, 16332–16333.

Dembinski, H., Wismer, K., Balasubramaniam, D., Gonzalez, H. a, Alverdi, V., Iakoucheva, L.M., Komives, E. a, 2014. Predicted disorder-to-order transition mutations in IκBα disrupt function. Phys. Chem. Chem. Phys. 16, 6480–5.

Deutsch, H.F., 1991. Chemistry and Biology of α-Fetoprotein, in: Advances in Cancer Research. pp. 253–312.

Di Domenico, T., Walsh, I., Martin, A.J.M., Tosatto, S.C.E., 2012. MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 28, 2080–2081.

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., Gibson, T.J., 2008. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front. Biosci. 13, 6580–6603.

Disfani, F.M., Hsu, W.-L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., Kurgan, L., 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28, i75–i83.

Dosztányi, Z., Csizmok, V., Tompa, P., Simon, I., 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433–3434.

Dosztányi, Z., Csizmók, V., Tompa, P., Simon, I., 2005b. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347, 827–39.

Dosztányi, Z., Meszaros, B., Simon, I., 2010. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief. Bioinform. 11, 225–243.

Dosztányi, Z., Meszaros, B., Simon, I., 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25, 2745–2746.

Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43, W389–W394.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D.T., Longhi, S., Metallo, S.J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R. V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J.L., Tompa, P., Uversky, V.N., 2014. What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disord. Proteins 1, e24157.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., Obradović, Z., 2002. Intrinsic disorder and protein function. Biochemistry 41, 6573–6582.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, W., Garner, E.C., Obradovic, Z., 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59.

Dunker, A.K., Obradovic, Z., 2001. The protein trinity—linking function and disorder. Nat. Biotechnol. 19, 805–806.

Dunker, A.K., Oldfield, C.J., Meng, J., Romero, P., Yang, J.Y., Chen, J.W., Vacic, V., Obradovic, Z., Uversky, V.N., 2008. The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9 Suppl 2, S1.

Dunker, A.K., Uversky, V.N., 2008. Signal transduction via unstructured protein conduits. Nat. Chem. Biol. 4, 229–30.

Dyson, H.J., Wright, P.E., 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6, 197–208.

Dyson, H.J., Wright, P.E., 2004. Unfolded Proteins and Protein Folding Studied by NMR. Chem. Rev. 104, 3607–3622.

Edwards, Y.J., Lobley, A.E., Pentony, M.M., Jones, D.T., 2009. Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. Genome Biol. 10, R50.

Efremov, R.G., Sazanov, L. a, 2011. Structure of the membrane domain of respiratory complex I. Nature 476, 414–420.

Eliezer, D., 2009. Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 19, 23–30.

Esteban-Martín, S., Fenwick, R.B., Salvatella, X., 2010. Refinement of Ensembles Describing Unstructured Proteins Using NMR Residual Dipolar Couplings. J. Am. Chem. Soc. 132, 4626–4632.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42, D222–30.

Fischer, E., 1894. Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der Dtsch. Chem. Gesellschaft 27, 2985–2993.

Fisher, C.K., Huang, A., Stultz, C.M., 2010. Modeling Intrinsically Disordered Proteins with Bayesian Statistics. J. Am. Chem. Soc. 132, 14919–14927.

Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S.D., Koike, R., Hiroaki, H., Ota, M., 2014. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res. 42, D320–5.

Fukuchi, S., Sakamoto, S., Nobe, Y., Murakami, S.D., Amemiya, T., Hosoda, K., Koike, R., Hiroaki, H., Ota, M., 2012. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res. 40, D507–11.

Fuxreiter, M., 2012. Fuzziness: linking regulation to protein dynamics. Mol. BioSyst. 8, 168–177.

Ganguly, D., Chen, J., 2009. Atomistic Details of the Disordered States of KID and pKID. Implications in Coupled Binding and Folding. J. Am. Chem. Soc. 131, 5214–5223.

Gough, J., Karplus, K., Hughey, R., Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. 313, 903–19.

Grimmler, M., Wang, Y., Mund, T., Cilensek, Z., Keidel, E.-M., Waddell, M.B., Jäkel, H., Kullmann, M., Kriwacki, R.W., Hengst, L., Cilenšek, Z., Keidel, E.-M., Waddell, M.B., Jäkel, H., Kullmann, M., Kriwacki, R.W., Hengst, L., 2007. Cdk-inhibitory activity and stability of p27Kip1 are directly regulated by oncogenic tyrosine kinases. Cell 128, 269–280.

Gront, D., Kmiecik, S., Kolinski, A., 2007. Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. J. Comput. Chem. 28, 1593–1597.

Gsponer, J., Futschik, M.E., Teichmann, S.A., Babu, M.M., 2008. Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. Science (80-.). 322, 1365–1368.

Habchi, J., Tompa, P., Longhi, S., Uversky, V.N., 2014. Introducing protein intrinsic disorder. Chem. Rev. 114, 6561–88.

Heller, G.T., Sormanni, P., Vendruscolo, M., 2015. Targeting disordered proteins with small molecules using entropy. Trends Biochem. Sci. 40, 491–496.

Henriques, J., Cragnell, C., Skepö, M., 2015. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. J. Chem. Theory Comput. 11, 150616124431001.

Higo, J., Nishimura, Y., Nakamura, H., 2011. A Free-Energy Landscape for Coupled Folding and Binding of an Intrinsically Disordered Protein in Explicit Solvent from Detailed All-Atom Computations. J. Am. Chem. Soc. 133, 10448–10458.

Hollstein, M., Sidransky, D., Vogelstein, B., Harris, C., 1991. p53 mutations in human cancers. Science (80-.). 253, 49–53.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., Dunker, a. K., 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323, 573–584.

Ishida, T., Kinoshita, K., 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. 35, W460–W464.

Ito, Y., Mikawa, T., Smith, B.O., 2012. In-Cell NMR of Intrinsically Disordered Proteins in Prokaryotic Cells, in: Methods in Molecular Biology. pp. 19–31.

Jamroz, M., Kolinski, A., Kmiecik, S., 2014. CABS-flex predictions of protein flexibility compared with NMR ensembles. Bioinformatics 30, 2150–2154.

Jamroz, M., Kolinski, A., Kmiecik, S., 2013a. CABS-flex: server for fast simulation of protein structure fluctuations. Nucleic Acids Res. 41, W427–W431.

Jamroz, M., Orozco, M., Kolinski, A., Kmiecik, S., 2013b. Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field. J. Chem. Theory Comput. 9, 119–125.

Janin, J., Sternberg, M.J.E., 2013. Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biol. Rep. 5, 2.

Jensen, M.R., Ruigrok, R.W., Blackledge, M., 2013. Describing intrinsically disordered proteins at atomic resolution by NMR. Curr. Opin. Struct. Biol. 23, 426–435.

Jin, F., Yu, C., Lai, L., Liu, Z., 2013. Ligand clouds around protein clouds: a scenario of ligand binding with intrinsically disordered proteins. PLoS Comput. Biol. 9, e1003249.

Jones, D.T., 2001. Predicting novel protein folds by using FRAGFOLD. Proteins 45, 127–132.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.

Jones, D.T., 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins 29, 185–191.

Jones, D.T., Bryson, K., Coleman, A., McGuffin, L.J., Sadowski, M.I., Sodhi, J.S., Ward, J.J., 2005. Prediction of novel and analogous folds using fragment assembly and fold recognition. Proteins 61, 143–151.

Jones, D.T., Cozzetto, D., 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31, 857–863.

Jones, D.T., McGuffin, L.J., 2003. Assembling novel protein folds from super-secondary structural fragments. Proteins 53, 480–485.

Jones, D.T., Singh, T., Kosciolek, T., Tetchner, S., 2015. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31, 999–1006.

Jones, D.T., Ward, J.J., 2003. Prediction of Disordered Regions in Proteins From Position Specific Score Matrices. Proteins 53, 573–578.

Karush, F., 1950. Heterogeneity of the Binding Sites of Bovine Serum Albumin 1. J. Am. Chem. Soc. 72, 2705–2713.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., Phillips, D.C., 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181, 662–666.

Khoury, G.A., Smadbeck, J., Kieslich, C.A., Floudas, C.A., 2014. Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol. 32, 99–109.

Kolinski, A., 2004. Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. 51, 349–371.

Kosciolek, T., Jones, D.T., 2015. Accurate contact predictions using coevolution techniques and machine learning. Proteins n/a–n/a.

Kosciolek, T., Jones, D.T., 2014. De Novo Structure Prediction of Globular Proteins Aided by Sequence Variation-Derived Contacts. PLoS One 9, e92197.

Koshland, D.E., 1959. Enzyme flexibility and enzyme action. J. Cell. Comp. Physiol. 54, 245–258.

Kosol, S., Contreras-Martos, S., Cedeño, C., Tompa, P., 2013. Structural characterization of intrinsically disordered proteins by NMR spectroscopy. Molecules 18, 10802–28.

Kozlowski, L.P., Bujnicki, J.M., 2012. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics 13, 111.

Kragelj, J., Ozenne, V., Blackledge, M., Jensen, M.R., 2013. Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts. ChemPhysChem 14, 3034–3045.

Krzeminski, M., Marsh, J.A., Neale, C., Choy, W.-Y., Forman-Kay, J.D., 2013. Characterization of disordered proteins with ENSEMBLE. Bioinformatics 29, 398–399.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–8.

Kurowski, M.A., 2003. GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31, 3305–3307.

Laskowski, R., Rullmann, J.A., MacArthur, M., Kaptein, R., Thornton, J., 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR 8, 477–486.

Latysheva, N.S., Flock, T., Weatheritt, R.J., Chavali, S., Babu, M.M., 2015. How do disordered regions achieve comparable functions to structured domains? Protein Sci. 24, 909–22.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C. a., Sheffler, W., Davis, I.W., Cooper, S., Treuille, A., Mandell, D.J., Richter, F., Ban, Y.-E.A., Fleishman, S.J., Corn, J.E., Kim, D.E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J.J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J.J., Kuhlman, B., Baker, D., Bradley, P., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 487, 545–574.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lei, H., Duan, Y., 2007. Improved sampling methods for molecular simulation. Curr. Opin. Struct. Biol. 17, 187–191.

Li, X., Romero, P., Rani, M., Dunker, A.K., Obradovic, Z., 1999. Predicting Protein Disorder for N-, C-, and Internal Regions. Genome informatics 10, 30–40.

Lieutaud, P., Canard, B., Longhi, S., 2008. MeDor: a metaserver for predicting protein disorder. BMC Genomics 9 Suppl 2, S25.

Linding, R., 2003. GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res. 31, 3701–3708.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B., 2003. Protein Disorder Prediction. Structure 11, 1453–1459.

Lindorff-Larsen, K., Best, R.B., Depristo, M.A., Dobson, C.M., Vendruscolo, M., 2005. Simultaneous determination of protein structure and dynamics. Nature 433, 128–32.

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., Shaw, D.E., 2012. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. J. Am. Chem. Soc. 134, 3787–3791.

Lipari, G., Szabo, A., 1982. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. J. Am. Chem. Soc. 104, 4546–4559.

Liu, J., Tan, H., Rost, B., 2002. Loopy Proteins Appear Conserved in Evolution. J. Mol. Biol. 322, 53–64.

Lundström, J., Rychlewski, L., Bujnicki, J., Elofsson, A., 2001. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci. 10, 2354–2362.

Mao, A.H., Crick, S.L., Vitalis, A., Chicoine, C.L., Pappu, R. V, 2010. Net Charge per Residue Modulates Conformational Ensembles of Intrinsically Disordered Proteins. Proc. Natl. Acad. Sci. U.S.A. 107, 8183–8188.

Mark, W.Y., Liao, J.C.C., Lu, Y., Ayed, A., Laister, R., Szymczyna, B., Chakrabartty, A., Arrowsmith, C.H., 2005. Characterization of segments from the central region of BRCA1: An intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? J. Mol. Biol. 345, 275–287.

Marsh, J. a, Singh, V.K., Jia, Z., Forman-Kay, J.D., 2006. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. Protein Sci. 15, 2795–2804.

Marsh, J. a., Forman-Kay, J.D., 2012. Ensemble modeling of protein disordered states: Experimental restraint contributions and validation. Proteins 80, 556–572.

Martin, A.J.M., Walsh, I., Tosatto, S.C.E., 2010. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. Bioinformatics 26, 2916–2917.

Mavridis, L., Nath, N., Mitchell, J.B., 2013. PFClust: a novel parameter free clustering algorithm. BMC Bioinformatics 14, 213.

McLachlan, A.D., 1982. Rapid comparison of protein structures. Acta Crystallogr. Sect. A 38, 871–873.

Mitra, P., Shultis, D., Zhang, Y., 2013. EvoDesign: de novo protein design based on structural and evolutionary profiles. Nucleic Acids Res. 41, W273–W280.

Mittag, T., Forman-Kay, J.D., 2007. Atomic-level characterization of disordered protein ensembles. Curr. Opin. Struct. Biol. 17, 3–14.

Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., Kurgan, L., 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26, i489–i496.

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., Kryshtafovych, A., 2011. Evaluation of disorder predictions in CASP9. Proteins 79, 107–118.

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., Fidelis, K., 2014. Assessment of protein disorder region predictions in CASP10. Proteins 82, 127–137.

Morcos, F., Jana, B., Hwa, T., Onuchic, J.N., 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc. Natl. Acad. Sci. U.S.A. 110, 20533–8.

Moritsugu, K., Terada, T., Kidera, A., 2012. Disorder-to-Order Transition of an Intrinsically Disordered Region of Sortase Revealed by Multiscale Enhanced Sampling. J. Am.

Chem. Soc. 134, 7094–7101.

Moult, J., Fidelis, K., Zemla, A., Hubbard, T., 2003. Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 53, 334–339.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Musayeva, K., Henderson, T., Mitchell, J.B., Mavridis, L., 2014. PFClust: an optimised implementation of a parameter-free clustering algorithm. Source Code Biol. Med. 9, 5.

Naganathan, A.N., Orozco, M., 2011. The Native Ensemble and Folding of a Protein Molten-Globule: Functional Consequence of Downhill Folding. J. Am. Chem. Soc. 133, 12154–12161.

Nehrt, N.L., Peterson, T.A., Park, D., Kann, M.G., 2012. Domain landscapes of somatic mutations in cancer. BMC Genomics 13, S9.

Nodet, G., Salmon, L., Ozenne, V., Meier, S., Jensen, M.R., Blackledge, M., 2009. Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings. J. Am. Chem. Soc. 131, 17908–17918.

Noivirt-Brik, O., Prilusky, J., Sussman, J.L., 2009. Assessment of disorder predictions in CASP8. Proteins 77, 210–216.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztányi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., Dunker, A.K., Gough, J., 2013. D2P2: database of disordered protein predictions. Nucleic Acids Res. 41, D508–D516.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K., 2003. Predicting intrinsic disorder from amino acid sequence. Proteins 53 Suppl 6, 566–72.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Dunker, a. K., 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 61, 176–182.

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

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., Thornton, J., 1997. CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1109.

Orosz, F., Ovadi, J., 2011. Proteins without 3D structure: definition, detection and beyond. Bioinformatics 27, 1449–1454.

Ota, M., Koike, R., Amemiya, T., Tenno, T., Romero, P.R., Hiroaki, H., Dunker, A.K., Fukuchi, S., 2013. An assignment of intrinsically disordered regions of proteins based on NMR structures. J. Struct. Biol. 181, 29–36.

Ozenne, V., Bauer, F., Salmon, L., Huang, J. -r., Jensen, M.R., Segard, S., Bernado, P., Charavay, C., Blackledge, M., 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28, 1463–1470.

Palazzesi, F., Prakash, M.K., Bonomi, M., Barducci, A., 2015. Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. J. Chem. Theory Comput. 11, 2–7.

Penel, S., Morrison, R.G., Dobson, P.D., Mortishire-Smith, R.J., Doig, A.J., 2003. Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. Protein Eng. 16, 957–961.

Penel, S., Morrison, R.G., Mortishire-Smith, R.J., Doig, A.J., 1999. Periodicity in α-helix lengths and C-capping preferences. J. Mol. Biol. 293, 1211–1219.

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

Peng, Z., Mizianty, M.J., Kurgan, L., 2014a. Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins 82, 145–158.

Peng, Z., Mizianty, M.J., Xue, B., Kurgan, L., Uversky, V.N., 2012. More than just tails: intrinsic disorder in histone proteins. Mol. Biosyst. 8, 1886–901.

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., Kurgan, L., 2014b. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. 72, 137–151.

Pentony, M.M., Ward, J.J., Jones, D.T., 2010. Computational resources for the prediction and analysis of native disorder in proteins. Methods Mol. Biol. 604, 369–93.

Potenza, E., Di Domenico, T., Walsh, I., Tosatto, S.C.E., 2015. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res. 43, D315–20.

Potoyan, D. a., Papoian, G.A., 2011. Energy Landscape Analyses of Disordered Histone Tails Reveal Special Organization of Their Conformational Dynamics. J. Am. Chem. Soc. 133, 7405–7415.

Powers, R., Clore, G.M., Garrett, D.S., Gronenborn, A.M., 1993. Relationships between the precision of high-resolution protein NMR structures, solution-order parameters, and crystallographic B factors. J. Magn. Reson. Ser. B 101, 325–327.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Sussman, J.L., 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21, 3435–3438.

Qin, S., Zhou, H.-X., 2013. Effects of Macromolecular Crowding on the Conformational Ensembles of Disordered Proteins. J. Phys. Chem. Lett. 4.

Radivojac, P., Obradović, Z., Brown, C.J., Dunker, A.K., 2003. Prediction of boundaries between intrinsically ordered and disordered protein regions. Pac. Symp. Biocomput. 216–27.

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., Dunker, A.K., 2004. Protein flexibility and intrinsic disorder. Protein Sci. 13, 71–80.

Rand, W.M., 1971. Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. 66, 846–850.

Rao, J.N., Jao, C.C., Hegde, B.G., Langen, R., Ulmer, T.S., 2010. A Combinatorial NMR and EPR Approach for Evaluating the Structural Ensemble of Partially Folded Proteins. J. Am. Chem. Soc. 132, 8657–8668.

Remmert, M., Biegert, A., Hauser, A., Söding, J., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175.

Roberts, G.C.K., 1993. NMR of macromolecules: a practical approach. IRL Press at Oxford University Press.

Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D., 2004. Protein Structure Prediction Using Rosetta, in: Methods in Enzymology. pp. 66–93.

Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Dunker, A.K., 1997. Identifying disordered regions in proteins from amino acid sequence, in: Proceedings of International Conference on Neural Networks (ICNN’97). IEEE, pp. 90–95.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, a K., 1998. Thousands of proteins likely to have long disordered regions. Pacific Symp. Biocomput. 437–448.

Rosato, A., Aramini, J.M., Arrowsmith, C., Bagaria, A., Baker, D., Cavalli, A., Doreleijers, J.F., Eletsky, A., Giachetti, A., Guerry, P., Gutmanas, A., Güntert, P., He, Y., Herrmann, T., Huang, Y.J., Jaravine, V., Jonker, H.R.A., Kennedy, M.A., Lange, O.F., Liu, G., Malliavin, T.E., Mani, R., Mao, B., Montelione, G.T., Nilges, M., Rossi, P., van der Schot, G., Schwalbe, H., Szyperski, T.A., Vendruscolo, M., Vernon, R., Vranken, W.F., de Vries, S., Vuister, G.W., Wu, B., Yang, Y., Bonvin, A.M.J.J., 2012. Blind Testing of Routine, Fully Automated Determination of Protein Structures from NMR Data. Structure 20, 227–236.

Rosato, A., Tejero, R., Montelione, G.T., 2013. Quality assessment of protein NMR structures. Curr. Opin. Struct. Biol. 23, 715–24.

Rost, B., 2005. How to use protein 1D structure predicted by PROFphd., in: Walker, J.E. (Ed.), The Proteomics Protocols Handbook: Methods in Molecular Biology. Humana, pp. 875–901.

Roy, A., Kucukural, A., Zhang, Y., 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. 5, 725–738.

Rupp, B., 2009. Biomolecular crystallography: principles, practice, and application to structural biology. Garland Science.

Schaefer, C., Schlessinger, A., Rost, B., 2010. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics 26, 625–631.

Schlessinger, A., Rost, B., 2005. Protein flexibility and rigidity predicted from sequence. Proteins 61, 115–126.

Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M., Rost, B., 2011. Protein disorder—a breakthrough invention of evolution? Curr. Opin. Struct. Biol. 21, 412–418.

Schlessinger, A., Yachdav, G., Rost, B., 2006. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 22, 891–893.

Shen, Y., Vernon, R., Baker, D., Bax, A., 2009. De novo protein structure generation from incomplete chemical shift assignments. J. Biomol. NMR 43, 63–78.

Shimizu, K., Hirose, S., Noguchi, T., 2007. POODLE-S: Web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics 23, 2337–2338.

Shortle, D., Simons, K.T., Baker, D., 1998. Clustering of low-energy conformations near the native structures of small proteins. Proc. Natl. Acad. Sci. U.S.A. 95, 11158–11162.

Sibille, N., Bernadó, P., 2012. Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. Biochem. Soc. Trans. 40, 955–962.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., Dunker, A.K., 2007. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. 35, D786–D793.

Šikić, M., Tomić, S., Vlahoviček, K., 2009. Prediction of Protein–Protein Interaction Sites in Sequences and 3D Structures by Random Forests. PLoS Comput. Biol. 5, e1000278.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J., Orengo, C.A., 2015. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43, D376–D381.

Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D., 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 37, 171–176.

Simons, K.T., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. 268, 209–225.

Skwark, M.J., Raimondi, D., Michel, M., Elofsson, A., 2014. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. PLoS Comput. Biol. 10, e1003889.

Sormanni, P., Camilloni, C., Fariselli, P., Vendruscolo, M., 2015. The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. J. Mol. Biol. 427, 982–96.

Staneva, I., Huang, Y., Liu, Z., Wallin, S., 2012. Binding of Two Intrinsically Disordered Peptides to a Multi-Specific Protein: A Combined Monte Carlo and Molecular Dynamics Study. PLoS Comput. Biol. 8, e1002682.

Stellwagen, E., Rysavy, R., Babul, G., 1972. The conformation of horse heart apocytochrome c. J. Biol. Chem. 247, 8074–7.

Su, C.-T., Chen, C.-Y., Hsu, C.-M., 2007. iPDA: integrated protein disorder analyzer. Nucleic

Acids Res. 35, W465–W472.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., von Mering, C., 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43, D447–52.

Tamiola, K., Acar, B., Mulder, F. a a, 2010. Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins. J. Am. Chem. Soc. 132, 18000–18003.

Tamiola, K., Mulder, F.A.A., 2012. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. Biochem. Soc. Trans. 40, 1014–1020.

Theillet, F.-X., Binolfi, A., Frembgen-Kesner, T., Hingorani, K., Sarkar, M., Kyne, C., Li, C., Crowley, P.B., Gierasch, L., Pielak, G.J., Elcock, A.H., Gershenson, A., Selenko, P., 2014. Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs). Chem. Rev. 114, 6661–6714.

Theobald, D.L., Steindel, P. a., 2012. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics 28, 1972–1979.

Theobald, D.L., Wuttke, D.S., 2008. Accurate Structural Correlations from Maximum Likelihood Superpositions. PLoS Comput. Biol. 4, e43.

Toretsky, J.A., Wright, P.E., 2014. Assemblages: functional units formed by cellular phase separation. J. Cell Biol. 206, 579–88.

Tronrud, D.E., 1996. Knowledge-Based B-Factor Restraints for the Refinement of Proteins. J. Appl. Crystallogr. 29, 100–104.

Trott, O., Siggers, K., Rost, B., Palmer, A.G., 2008. Protein conformational flexibility prediction using machine learning. J. Magn. Reson. 192, 37–47.

Tsirigos, K.D., Peters, C., Shu, N., Käll, L., Elofsson, A., 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 43, W401–W407.

Uversky, V.N., 2013. A decade and a half of protein intrinsic disorder: Biology still waits for physics. Protein Sci. 22, n/a–n/a.

Uversky, V.N., 2013. Unusual biophysics of intrinsically disordered proteins. Biochim. Biophys. Acta - Proteins Proteomics 1834, 932–951.

Uversky, V.N., Davé, V., Iakoucheva, L.M., Malaney, P., Metallo, S.J., Pathak, R.R., Joerger, A.C., 2014. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. Chem. Rev. 114, 6844–79.

Uversky, V.N., Dunker, A.K., 2013. The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. F1000 Biol. Rep. 5, 1.

Uversky, V.N., Dunker, A.K., 2010. Understanding protein non-folding. Biochim. Biophys. Acta - Proteins Proteomics 1804, 1231–1264.

Uversky, V.N., Gillespie, J.R., Fink, A.L., 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins 41, 415–27.

Uversky, V.N., Oldfield, C.J., Dunker, a K., 2008. Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept. Annu. Rev. Biophys. 37, 215–246.

Vacic, V., Iakoucheva, L.M., 2012. Disease mutations in disordered regions--exception to the rule? Mol. Biosyst. 8, 27–32.

Vacic, V., Markwick, P.R.L., Oldfield, C.J., Zhao, X., Haynes, C., Uversky, V.N., Iakoucheva, L.M., 2012. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. PLoS Comput. Biol. 8, e1002709.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., Dunker, A.K., 2007. Characterization of molecular recognition features, MoRFs, and their binding partners. J. Proteome Res. 6, 2351–2366.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R. V, Tompa, P., Uversky, V.N., Wright, P.E., Babu, M.M., 2014. Classification of intrinsically disordered regions and proteins. Chem. Rev. 114, 6589–631.

van der Schot, G., Zhang, Z., Vernon, R., Shen, Y., Vranken, W.F., Baker, D., Bonvin, A.M.J.J., Lange, O.F., 2013. Improving 3D structure prediction from chemical shift data. J. Biomol. NMR 57, 27–35.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., Sussman, J., Svergun, D.I., Uversky, V.N., Vendruscolo, M., Wishart, D., Wright, P.E., Tompa, P., 2014. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 42, D326–35.

Varadi, M., Vranken, W., Guharoy, M., Tompa, P., 2015. Computational approaches for inferring the functions of intrinsically disordered proteins. Front. Mol. Biosci. 2, 45.

Vavouri, T., Semple, J.I., Garcia-Verdugo, R., Lehner, B., 2009. Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity. Cell 138, 198–208.

Vitalis, A., Pappu, R. V, 2009a. Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules, in: Annual Reports in Computational Chemistry. pp. 49–76.

Vitalis, A., Pappu, R. V., 2009b. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. J. Comput. Chem. 30, 673–699.

Vitalis, A., Wang, X., Pappu, R. V, 2007. Quantitative Characterization of Intrinsic Disorder in Polyglutamine: Insights from Analysis Based on Polymer Theories. Biophys. J. 93, 1923–1937.

Vitalis, A., Wang, X., Pappu, R. V., 2008. Atomistic Simulations of the Effects of Polyglutamine Chain Length and Solvent Quality on Conformational Equilibria and Spontaneous Homodimerization. J. Mol. Biol. 384, 279–297.

Vousden, K.H., Lu, X., 2002. Live or let die: the cell's response to p53. Nat. Rev. Cancer 2, 594–604.

Wallner, B., Larsson, P., Elofsson, A., 2007. Pcons.net: protein structure prediction meta server. Nucleic Acids Res. 35, W369–W374.

Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., Tosatto, S.C.E., 2015. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 31, 201–208.

Walsh, I., Martin, A.J.M., Di Domenico, T., Tosatto, S.C.E., 2012. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28, 503–509.

Walsh, I., Martin, A.J.M., Di Domenico, T., Vullo, A., Pollastri, G., Tosatto, S.C.E., 2011. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic Acids Res. 39, W190–W196.

Wang, R.Y.-R., Han, Y., Krassovsky, K., Sheffler, W., Tyka, M., Baker, D., 2011. Modeling disordered regions in proteins using Rosetta. PLoS One 6, e22060.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., Jones, D.T., 2004a. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20, 2138–2139.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T., 2004b. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. J. Mol. Biol. 337, 635–645.

Wright, P.E., Dyson, H.J., 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

Wu, K.-P., Weinstock, D.S., Narayanan, C., Levy, R.M., Baum, J., 2009. Structural Reorganization of α-Synuclein at Low pH Observed by NMR and REMD Simulations. J. Mol. Biol. 391, 784–796.

Xu, D., Zhang, Y., 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80, 1715–1735.

Xu, D., Zhang, Y., 2011. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. Biophys. J. 101, 2525–2534.

Xu, J., Zhang, Y., 2010. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26, 889–895.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, a. K., Uversky, V.N., 2010. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. Biochim. Biophys. Acta - Proteins Proteomics 1804, 996–1010.

Xue, B., Dunker, a. K., Uversky, V.N., 2012a. The Roles of Intrinsic Disorder in Orchestrating the Wnt-Pathway. J. Biomol. Struct. Dyn. 29, 843–861.

Xue, B., Dunker, A.K., Uversky, V.N., 2012b. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. 30, 137–149.

Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M., 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21, 3369–3376.

Yuan, Z., Bailey, T.L., Teasdale, R.D., 2005. Prediction of protein B-factor profiles. Proteins 58, 905–912.

Zhang, F., Brüschweiler, R., 2002. Contact Model for the Prediction of NMR N−H Order Parameters in Globular Proteins. J. Am. Chem. Soc. 124, 12654–12655.

Zhang, Y., Skolnick, J., 2004a. SPICKER: A clustering approach to identify near-native protein folds. J. Comput. Chem. 25, 865–871.

Zhang, Y., Skolnick, J., 2004b. Scoring function for automated assessment of protein structure template quality. Proteins 57, 702–710.