The Accurate Prediction of Disordered Regions in Protein Sequences Using Machine Learning Approaches

A thesis submitted for the degree of
Doctor of Philosophy

Pengfei Han
B.App.Sci. (Hons.)
School of Computer Science and Information Technology
College of Science, Engineering and Health
RMIT University
Melbourne, Victoria, Australia.

November, 2011

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Pengfei Han
School of Computer Science and Information Technology
RMIT University
November, 2011

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr Xiuzhen (Jenny) Zhang, Dr Zhi-Ping Feng and Prof. Ray Norton, for their continuous help and guidance during my study. Jenny is always energetic and ready to help. Zhi-Ping is enthusiastic for every discussion. Ray always finds time in his busy schedules to correct my papers and thesis. From them I learnt valuable research skills, and they are an important source of encouragement for me.

I would like to thank School of Computer Science and Information Technology for helping me in various aspects. I have been fortunate to share office with a group of innovative postgraduate students from Search Engine Group at RMIT University who made my days at RMIT University most enjoyable.

I have to thank my family, my wife Zhen Li, my parents Mr. Yiqun Han, Mrs. Fengqin Wang for their endless love and continuous support throughout my study.

Throughout my candidature, I have been supported by an Australian Postgraduate Award.

Credits

Preliminary versions of some results and discussions in this thesis have been previously published. Chapters four, five, six and seven contain material that appeared in the following publications:

- Proceedings of the fourth Australiasian Data Mining Conference [Han et al., 2005]
- Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining [Han et al., 2007]
- Journal of Computational Biology [Han et al., 2006]
- Journal of Molecular and Biochemical Parasitology [Feng et al., 2006]
- BMC Bioinformatics [Han et al., 2009a;b]

These papers are all co-authored with my supervisors Dr Xiuzhen (Jenny) Zhang, Dr Zhi-Ping Feng and Prof. Ray Norton.

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Meaning
AAC	Amino Acid Composition
AAindex	Amino Acid index
AUC	Area Under ROC Curve
BLAST	Basic Local Alignment Search Tool
CASP	The Critical Assessment of protein Structure Prediction
DisProt	The Database of Protein Disorder
DisProt	Predictor of Protein Disorder
DR	Disordered Region
DSSP	Define Secondary Structure of Proteins
EBI	European Bioinformatics Institute
IUP	Intrinsically Unstructured Protein
MSD	Macromolecular Structure Database group
NCBI	National Centre for Biotechnology Information
NMR	Nuclear Magnetic Resonance
NN	Neural Network
PDB	Protein Data Bank
PDBe	Protein Data Bank in Europe
PIR	Protein Information Resource
PONDR	Predictors Of Nature Disordered Region
PSSM	Position Specific Scoring Matrix
ROC	Receiver Operating Characteristic
SCOP	Structural Classification of Proteins
SVM	Support Vector Machine
UniProt	Universal Protein Resource
WT	Wavelet Transform

Abstract

A major challenge in the post-genome era is to determine the function of proteins. The traditional structure-function paradigm assumes that the function of a protein is contingent on it folding into a stable three-dimensional structure. However many proteins contain intrinsic unstructured or Disordered Regions (DRs) under physiological conditions, and yet they still carry important functions. Determination of the disordered regions in proteins is therefore an important step towards the determination of their functions. Traditional experimental approaches are generally time consuming and expensive. The efficient and cost-effective computer aided automatic prediction of DRs is thus an attractive alternative. To this end, we propose the novel application of machine learning models and physicochemical features extracted from protein sequences for predicting long, short and global disorder in proteins.

To improve the understandability of disorder prediction, rule based predictors are proposed, which are not only able to predict DRs, but can also quantify previously unknown associations between order disorder status and sequences. The prediction process is transparent and simple to explain.

As DRs of different lengths possess different properties, to achieve a high accuracy of prediction, we propose predictors specific to long, short and global disorder prediction. These predictors are distinct from each other in terms of their features, the machine learning models used, and the methods of prediction. We thoroughly investigate the database of physicochemical properties of amino acid indices and select the indices most correlated with disorder. Based on these properties, novel feature transforms including autocorrelation and wavelet transforms (WTs) are applied to DR prediction. According to the results of cross-validation tests, our long DR predictor based on autocorrelation achieves the highest accuracy of prediction among long DR predictors at an AUC (Area Under ROC Curve) value of 89.5%. A short DR predictor based on WTs achieves an AUC value of 88.7%, which is comparable to

the most accurate short DR predictors. The global DR predictor achieves an AUC value of 96.1%, close to the optimal value.

A major bottleneck of large scale DR prediction is the time efficiency constraint that is attributed to slow feature generation stages and complicated prediction methods. Both our long and short DR predictors are built from simple methods of prediction and feature space. Our web service for long DR prediction can process an uploaded file of multiple sequences.

Chapter 1

Introduction

Proteins are important molecules in all organisms. Proteins have different levels of structures including primary, secondary and tertiary structures. Loosely speaking, the primary structure is the primary sequence. The secondary structure includes α helices, β sheets (strands) and coils. The tertiary structure gives a three dimensional arrangement of all residues. Existing structural biology experiments suggest that all the information needed for a protein to adopt its native conformation is encoded in its primary structure. If this is really so, then theoretically it is possible to derive rules by analysing sequences with known structures, and then apply these rules to predict the tertiary structure of any given sequence[Attwood and Parry-Smith, 2001]. Unfortunately, the relationship between a sequence and structure is not that simple.

The traditional structure-function paradigm interprets function in terms of a specific three-dimensional structure. According to Dunker and Obradovic [2001], for more than 100 years, the functions of proteins have been believed to depend on these structures.

In contrast to regular secondary structures which have a regular conformation, many proteins contain flexible regions that lack specific tertiary structures. In some cases proteins are entirely flexible ensembles. These flexible regions are called disordered regions (DRs). DRs can hinder the crystallisation process and have a lower complexity [Romero et al., 1999; 2001], a different amino acid composition, hydropathy, charge and coordination number [Romero et al., 1997a;b; 1998; Dunker et al., 1998; 2001] from the structured (ordered) regions. Associations between DRs and some functions related to diseases such as cancer [Peng et al., 2006; Russell and Gibson, 2008] have been determined. These discoveries contradict the generally accepted structure-function paradigm of modern protein science [Daughdrill et al., 2005].

DRs have attracted increasing interest from the research community. The number of PubMed[1] hits on disordered regions/proteins was zero before 1990 and this figure increased to around fifteen between 1995 and 1999. The number of publications has dramatically increased since 2000. In 2004, around 50 papers related to these proteins were published, while more than 160 papers were published in 2007 alone [Russell and Gibson, 2008].

It takes a long time for traditional experimental approaches to annotate DRs in a single protein sequence, which is a bottleneck for large scale research on disorder. This has led to research into computational approaches for automatic DR annotation. Machine learning approaches have been used for the prediction of DRs. Since the first DR predictor, PONDR, was proposed in 1997 [Li et al., 1999], more than twenty DR predictors have been developed and published. Most predictors [Li et al., 1999; Linding et al., 2003a; Obradovic et al., 2003; Peng et al., 2006; Hirose et al., 2007] are based on machine learning approaches such as Neural Networks (NNs), Support Vector Machine (SVM) or ensemble learning. We next introduce the history of DR prediction, then existing DR predictors are summarised based on accuracy, efficiency and understandability respectively.

1.1 Disorder Prediction-A Historical View

Generally, early DR predictors do not stand up to modern approaches in terms of accuracy of prediction. However, they have made significant contributions by introducing a clearer definition of disorder or by proposing pioneering algorithms.

The first DR predictor, PONDR, had no clear definitions of disorder and its disordered database was built by extracting sequences from a protein data bank and from literature reports. Limited effort was made by PONDR to define disorder and standardise the searching process for proteins with DRs. However, PONDR applied a novel machine learning approach, neural networks, to design a predictor. Three region-specific predictors were built that predicted a disordered status on a residue-by-residue basis.

DisEMBL [Linding et al., 2003a] was another early computational tool for prediction of DRs in protein sequences. In contrast to PONDR, DisEMBL focused on different definitions of disorder and introduced a new definition called “hot loops”. In addition to proposing a theoretical definition of disorder, an operational specification for building disordered databases according to these definitions was provided. As three separate predictors were built from

[1] `http://www.ncbi.nlm.nih.gov/pubmed/`, the online database of scientific literature provided by the US National Library of Medicine

different training databases, DisEMBL decreased the risk of missing DRs in disorder prediction.

Disorder predictor GlobPlot [Linding et al., 2003b] is also one of the most cited early approaches in this field. GlobPlot predicts disorder by adopting a novel "rule based" strategy and calculates rules from physicochemical properties of amino acids without using machine learning models. Therefore, the time efficiency of GlobPlot is better than most predictors based on machine learning models.

All of these early disorder predictors, PONDR, DisEMBL and GlobPlot, have web servers which are very useful for target selection and the design of constructs as needed for many biochemical studies, particularly structural biology and structural genomics projects.

1.2 Disorder Prediction-A Learning Perspective

Accuracy

It has been pointed out that DRs of different length show statistical differences [Radivojac et al., 2004] and long DRs possess different physicochemical properties from short DRs. The learning approaches applied to long DR prediction may not be applicable to short DR prediction.

CASP (Critical Assessment of Techniques for Protein Structure Prediction) competitions provide a platform for the comparison of DR predictors. It is known that target sequences of the CASP competitions are biased towards short DRs (<30 residues) and are deplete of long (≥30 residues) DRs [Grana et al., 2005; Bordoli et al., 2007]. In the CASP7 competition the most accurate three predictors are registered as human experts [Bordoli et al., 2006]. The highest accuracy is achieved by the ISTZORAN group [Peng et al., 2006] with an AUC[2] value of 86%. The next three most accurate predictors are registered as automatic servers with the highest AUC value of 83.7% from the group DISOPRED [Ward et al., 2004]. Most top ranked predictors [Ward et al., 2004; Cheng et al., 2005b; Peng et al., 2006] in the CASP7 competition involved training datasets that were dominated by short DRs.

In the most recent CASP8 competition completed at the end of 2008, the most accurate prediction was from a server, the mahmood-torda-server, with an AUC value of 91.6%. The runner up was a human expert, CBRC-DP_DR, with an AUC value of 91.0% [Prilusky et al., 2008]. The higher accuracy of prediction in the CASP8 competition may be attributable to

[2] AUC (Area Under Curve) is used to measure the accuracy of a predictor. The optimal result of an AUC is 100% and a detailed explanation can be found in Chapter 3.

the wide application of ensemble learning approaches and relatively easier targets [Prilusky et al., 2008].

In comparison to short DR prediction, the highest accuracy of prediction for long (>30 residues) DRs is around an AUC value of 87% [Schlessinger et al., 2007b; Hirose et al., 2007]. Both Schlessinger et al. [2007b] and Hirose et al. [2007] have applied disordered training datasets consisting only of long DRs.

Many recent DR predictors make use of ensemble learning, in which bagging or boosting approaches have been adopted [Hirose et al., 2007; Ishida and Kinoshita, 2008; McGuffin, 2008]. These predictors usually have two levels, with the first level being the output of other predictors or base predictors, and the second level machine learning model making a final prediction of order/disorder status. According to Ishida and Kinoshita [2008] and Hirose et al. [2007], these predictors, which reprocess the results of prediction of other individual predictors, can achieve a higher accuracy than any individual ones. However, due to their complexity, the time needed for prediction can be extremely long (up to one hour to predict a single sequence) [Ishida and Kinoshita, 2008]. More importantly, due to a high level of similarity among base predictors, the accuracy improvement is limited (1.7% more accurate than individual predictors, as shown by Ishida and Kinoshita [2008]).

In summary, to make the best use of the length specific property of DRs and in order to achieve a high accuracy of prediction, different training datasets of DRs are applied to build predictors tackling short and long DR prediction problems separately. More work related to short DR prediction has been carried out through the CASP competitions. To build a predictor that is more accurate than all the others in predicting both long and short DRs is difficult. Therefore, ensemble learning based DR predictors are proposed to combine the merits of the other predictors. However, the speed of prediction and the level of improvement are two limits to consider with ensemble predictors.

Efficiency

In the CASP7 competition, all six top ranked predictors applied alignment features in order to build their prediction models. Alignment is a slow process and predictors based on this are not suitable for high throughput analysis. Complex machine learning models such as that of ensemble learning can also prolong the prediction process significantly.

Alternatively, some successful DR predictors are built without the involvement of alignment or machine learning, such as the application of a position-specific energy prediction matrix [Dosztányi et al., 2005a;b] and the hydrophobicity and net charge of amino acids [Uversky et al., 2000; Prilusky et al., 2005]. These predictors do not need to load machine learning models from hard disks and are usually very fast at prediction. However, their accuracy is generally comparable or lower than predictors based on machine learning approaches.

As a result, if a large scale DR prediction is the priority, time consuming alignment and complex learning models should be avoided when building DR predictors. Non-alignment based features are especially needed for short DR predictors.

Understandability

The understandability of most existing DR predictors based on machine learning models needs improvement. Using the feature space of Amino Acid Composition (AAC, to be discussed in Chapter 3) as an example, many DR predictors [Li et al., 1999; Linding et al., 2003a; Obradovic et al., 2003; Peng et al., 2006] do not address the relationship between the features and the results of prediction. Although Peng et al. [2006] discovered and listed some "promoting" residues of order/disorder, more sophisticated and important rules (patterns) between AAC and order/disorder status have not been investigated. Those rules can provide better understandability of the problem of DR prediction; for example, in terms of the quantified correlation between the composition of several different kinds of amino acids and the order/disorder status of residues in protein sequences. A predictor that could uncover this relationship would be very useful.

1.3 Contributions

In this thesis, we study the accurate prediction of various DRs in protein sequences using machine learning approaches. To this end, we investigated novel applications of machine learning models for various DR predictions. In addition to accuracy, we address problems such as understandability and efficiency.

- We investigate several machine learning models for disorder prediction including the decision tree, random forest and SVM. Our model can achieve a higher accuracy of prediction than most other existing DR predictors based on SVM and ensemble learning [Hirose et al., 2007]. To the best of our knowledge, this is the first application of the decision tree and random forest in disorder prediction.

- The AAindex (Amino Acid index) database [Kawashima et al., 1999] is a database of amino acid physicochemical properties. Indices most related to order/disorder status are selected and used to predict long DRs after autocorrelation transformation. We demonstrate that this novel transformation, along with several other commonly applied features, can achieve a higher accuracy of prediction than other published long DR predictors. More importantly, the speed of prediction is fast, since all features can be calculated easily and the method of prediction is simple.

  Most online DR prediction servers [Linding et al., 2003a;b; Ward et al., 2004; Cheng et al., 2005b; Dosztányi et al., 2005b; Prilusky et al., 2005; Galzitskaya et al., 2006a; MacCallum] predict only one sequence at a time (submission), due to the time consuming prediction process. For convenient high throughput analysis, we set up a web service[3] that provides an online service for the large scale prediction of sequences. This makes it much easier for a researcher to get prediction results for a proteome which may have hundreds or more proteins.

- Most successful short DR predictors have applied evolutionary information such as alignment results during the training and prediction stages [Ward et al., 2004; Cheng et al., 2005b; Peng et al., 2006]. This information can improve the accuracy of prediction for many sequences but is not applicable to some proteins (orphan proteins).

  A novel application of wavelet transforms of indices selected from the AAindex database is proposed for short DR prediction. Wavelet transforms use basis localised in time and frequency in order to represent nonstationary signals. They have the advantage of providing analysis opportunities from different angles and distances. More importantly, wavelet transforms are calculated from the sequence itself, and do not rely on evolutionary information. Our experiments show that the application of wavelet transforms achieves a higher accuracy of prediction than many commonly used feature groups.

  We have compared different feature groups that are applied in short DR prediction. Our experiments show that after combining wavelet transforms with other feature groups, the accuracy of prediction can be improved up to 17.5%. Wavelet transforms are thus very suitable for predicting short DRs.

- We propose a decision tree based global DR predictor, which specifies whether the sequence is completely (or nearly completely) ordered/disordered. AACs are training

[3] `http://dmg.cs.rmit.edu.au/IUPforest/IUPforest-L.php`

features used to build the decision tree. Each rule from the tree quantifies the relationship between AAC and order/disorder status. A prominent feature of our model is its strong understandability.

1.4 Organisation of the Thesis

The remainder of this thesis is organised as follows.

In Chapter 2, the necessary background knowledge of the biology involved is introduced. It is followed by a survey of related work in DR prediction. Specifically, work regarding long disorder predictors, short disorder predictors, comprehensive disorder predictors and global disorder predictors is discussed.

Preliminaries are given in Chapter 3. Different terms concerning disorder are described first, followed by an introduction of the databases that are used in our research. Given that DRs do not occur uniformly in protein sequences, the distribution of DRs in different databases is analysed. Then, features used in our models for DR prediction as well as the theory of alignment are described. Finally, evaluation metrics for DR prediction are introduced.

In Chapter 4, a decision tree based long DR predictor is proposed. This predictor applies a window strategy during the learning/prediction procedure through which the precise position of long DRs can be identified.

IUPforest-L, a long DR prediction model based on the random forest and autocorrelation of amino acid physicochemical property indices, is described in Chapter 5. We study the performance of the random forest under different numbers of trees. The performance of novel feature autocorrelation derived from selected indices from a physicochemical property database is compared with existing feature groups. A combination of autocorrelation alongside existing features achieves the highest accuracy of prediction.

In Chapter 6, we present the short DR predictor based on wavelet transforms. Forty-five wavelets are tested, and only those most related to the order/disorder status of residues are selected. Experiments show that selected wavelet transforms lead to more accurate prediction than other nonalignment based feature groups in short DR prediction. Compared with alignment based predictors, our predictor achieves comparable accuracy of prediction in much less time. Continuing with the fast short DR predictor introduced, we present a more complicated short DR predictor combining wavelet transform and other features. Experiments show that this approach improves the accuracy of short DR prediction and achieves a higher accuracy than the winner server of the CASP7 competition.

In Chapter 7, various global disorder predictors are proposed. These predictors exploit the feature space AAC and reduced AAC. For query sequences, the models make a binary prediction of the ordered/disordered status. Biological meaningful rules are derived from the decision tree and these rules are discussed. The accuracy of prediction from different models and features is compared.

We conclude our study in Chapter 8, where the work of the thesis is summarised and future research problems are discussed.

Chapter 2

Background

In this chapter, we first introduce the biological background for DR prediction in Section 2.1, specifically the protein sequence structure and function. The main purpose of Section 2.1 is to make the thesis self-contained. Readers familiar with protein structure can skip this section. We review the definition of DRs in Section 2.2. According to the length of DRs predicted, we categorise existing DR predictors into four groups. They are: long DR predictors, short DR predictors, comprehensive DR predictors and global disorder predictors, described from Section 2.3 to Section 2.6. In Section 2.7, a summary of DR predictors is provided from the perspective of methods of prediction, features and training datasets.

2.1 The Structure-Function Paradigm of Proteins

Protein molecules are composed of amino acids (also called residues). The spatial organisation of a protein, its shape in three dimensions, is crucial to the understanding of its function [Lodish et al., 2004]. If chains contain less than 20-30 residues they are generally called peptides, while polypeptides can contain as many as 4,000 residues. The term polypeptide and protein are often used interchangeably when a protein refers to a polypeptide or a complex of polypeptides [Lodish et al., 2004]. Proteins can also work together to achieve a particular function, and they often associate to form stable complexes [Maton et al., 1993]. Amino acids as building blocks of proteins get names from their structure, which consists of an amino group (-NH2) and a carboxyl group (-COOH) connected through a central carbon atom. Most amino acids (except proline) have the structure 2HN-CHR-COOH, where R is a variable side chain attached to the central carbon atom. The building process that combines amino acids into a protein is shown in A), B) and C) of Figure 2.1 [Clark, 2005].

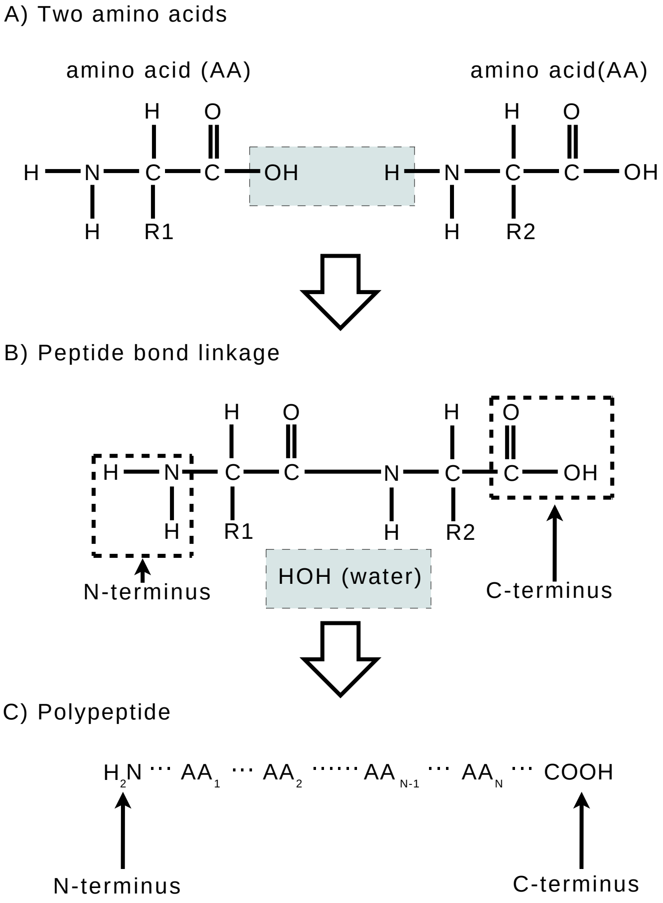


Figure 2.1: From amino acids to a protein

Two amino acids are linked through a peptide bond between COOH and NH2, with water eliminated in the process. Then successive amino acids are joined in this manner to form a polypeptide/protein.

As noted in B) and C) of Figure 2.1, one end of the linear polypeptides has a free amino group (-NH2) called the N-terminus and the other end has a free carboxyl group (-COOH) called the C-terminus [Clark, 2005].

Corresponding to 20 variations on the side chain R, there are 20 naturally occurring amino acids and these are listed in Table 2.1 [Kyte and Dolittle, 1982; Cooper and Hausman, 2004]. These 20 amino acids show different properties including polarity, charge and hydropathy. N and P in the column “Side chain polarity” represent negative and positive polarity. The

Table 2.1: 20 amino acids

Amino Acid	3-Letter	1-Letter	Side chain polarity	Side chain charge (PH7)	Hydropathy index
Alanine	Ala	A	N	O	1.8
Arginine	Arg	R	P	+	-4.5
Asparagine	Asn	N	P	O	-3.5
Aspartic acid	Asp	D	P	-	-3.5
Cysteine	Cys	C	N	O	2.5
Glutamic acid	Glu	E	P	-	-3.5
Glutamine	Gln	Q	P	O	-3.5
Glycine	Gly	G	N	O	-0.4
Histidine	His	H	P	+	-3.2
Isoleucine	Ile	I	N	O	4.5
Leucine	Leu	L	N	O	3.8
Lysine	Lys	K	P	+	-3.9
Methionine	Met	M	N	O	1.9
Phenylalanine	Phe	F	N	O	2.8
Proline	Pro	P	N	O	-1.6
Serine	Ser	S	P	O	-0.8
Threonine	Thr	T	P	O	-0.7
Tryptophan	Trp	W	N	O	-0.9
Tyrosine	Tyr	Y	P	O	-1.3
Valine	Val	V	N	O	4.2

annotation of "+", "-" and "O" in the column "Side chain charge" represents positive, negative and neutral charge, respectively. The percentages of 20 amino acids in a sequence are called the AAC. Amino acids can be grouped into reduced-AACs according to the properties of the relevant amino acids. Different groups of amino acids will be discussed further in later chapters.

Proteins can be described at different levels including primary, secondary and tertiary structures.

1. Primary structure. The primary structure (sequence) of a protein is established by the number, kind and sequence of amino acid units composing the polypeptide chain or chains making up the molecule. The primary structure determines the three-dimensional shape into which the protein folds. Therefore amino acid sequences are of primary importance in establishing protein shapes.

2. Secondary structure. There are several definitions of secondary structure. Danish pro-

tein chemist K.U.Linderstrom-Lang gives a widely accepted definition that secondary structure is: "the assignment of helices, sheets and the hydrogen-bonding pattern of the main chain" [Lesk, 2002]. Most proteins consist of several segments of α helices and/or β sheets separated from each other by various loop (coil) regions. α helix and β sheet secondary structures are strongly held in particular conformations by virtue of the large number of hydrogen bonds. Loop regions, conversely, are usually quite flexible and can vary in length and shape, allowing the overall polypeptide to fold into a compact tertiary structure. There are other definitions of secondary structure. The DSSP (Define Secondary Structure of Proteins) [Kabsch and Sander, 1983] program divides helices, sheets and loops further into eight types of secondary structures. However the three-element definition is used more often.

3. Tertiary structure. After a molecule has adopted its secondary structure, such as a helix or a beta sheet, it folds further. The tertiary structure refers to the overall conformation of a polypeptide chain, which is the three-dimensional arrangement of all its amino acid residues. The tertiary structure is primarily stabilised by hydrophobic interactions between the nonpolar side chains, hydrogen bonds between polar side chains and peptide bonds. These interactions and bonds hold the secondary structures, α helices, β sheets and loop regions compactly together [Lodish et al., 2004].

The primary structure and secondary structure annotation of a sample protein taken from PDB [1] [Berman et al., 2000] (Protein Data Bank) with code 1JC9 (chain A) are shown in Figure 2.2. The PDB database contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. Given a protein in PDB, its primary, secondary and tertiary structures can be retrieved. In Figure 2.2, the secondary structure annotation consisting of α helices and β sheets is annotated in red and green, respectively. Most black segments correspond to loop regions, while grey segments are regions without any structure. The tertiary structure of chain A of protein 1JC9 is shown in Figure 2.3 from two different angles. α helices and β sheets are labelled in red and green. All other parts are loops. The folding of the secondary structure is easy to see. Unstructured regions shown in grey segments in Figure 2.2 are too flexible to be presented in the tertiary structure of Figure 2.3.

The existing structure-function paradigm states that the amino acid sequence of a protein determines its three-dimensional structure, which in turn determines its function. This view

[1] `http://www.rcsb.org/pdb/home/home.do`

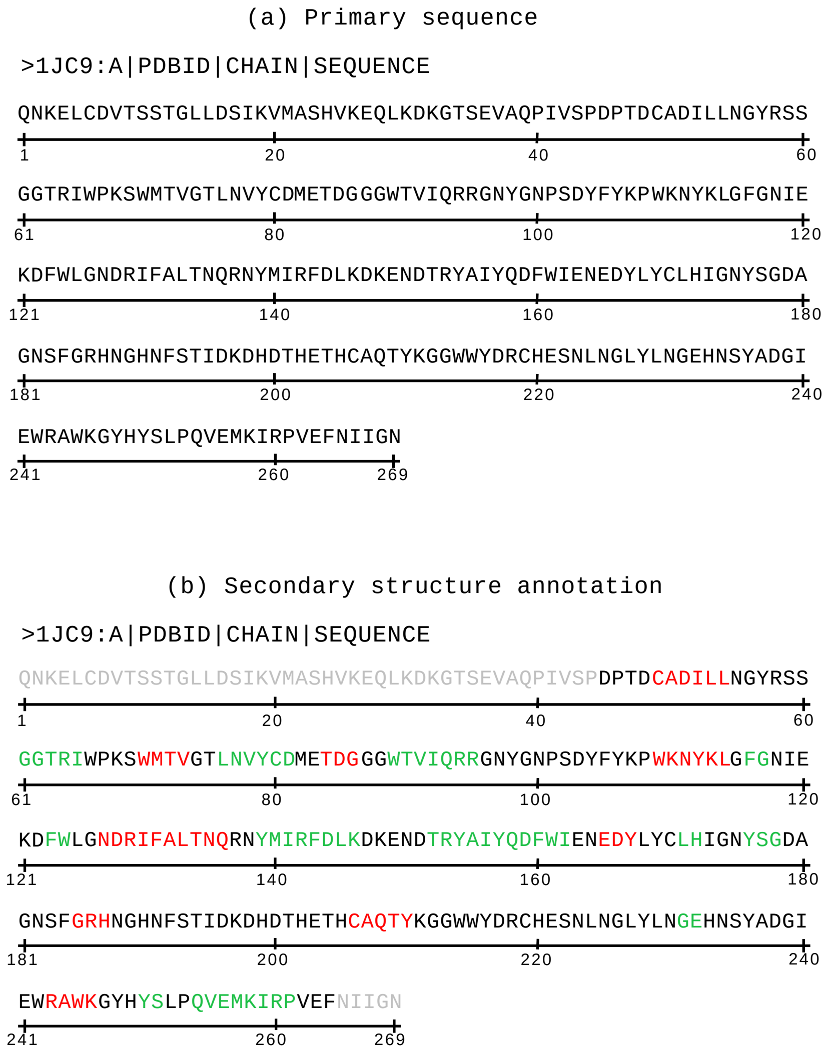

(a) Primary sequence
>1JC9:A|PDBID|CHAIN|SEQUENCE
QNKELCDVTSSTGLLDSIKVMASHVKEQLKDKGTSEVAQPIVSPDPTDCADILLNGYRSS
1
20
40
60
GGTRIWPKSWMTVGTLNVYCDMETDGGGWTVIQRRGNYGNPSDYFYKPWKNYKLGFGNIE
61
80
100
120
KDFWLGNDRIFALTNQRNYMIRFDLKDKENDTRYAIYQDFWIENEDYLYCLHIGNYSGDA
121
140
160
180
GNSFGRHNGHNFSTIDKDHDTHETHCAQTYKGGWWYDRCHESNLNGLYLNGEHNSYADGI
181
200
220
240
EWRAWKGYHYSLPQVEMKIRPVEFNIIGN
241
260
269
(b) Secondary structure annotation
>1JC9:A|PDBID|CHAIN|SEQUENCE
QNKELCDVTSSTGLLDSIKVMASHVKEQLKDKGTSEVAQPIVSPDPTDCADILLNGYRSS
1
20
40
60
GGTRIWPKSWMTVGTLNVYCDMETDGGGWTVIQRRGNYGNPSDYFYKPWKNYKLGFGNIE
61
80
100
120
KDFWLGNDRIFALTNQRNYMIRFDLKDKENDTRYAIYQDFWIENEDYLYCLHIGNYSGDA
121
140
160
180
GNSFGRHNGHNFSTIDKDHDTHETHCAQTYKGGWWYDRCHESNLNGLYLNGEHNSYADGI
181
200
220
240
EWRAWKGYHYSLPQVEMKIRPVEFNIIGN
241
260
269

Figure 2.2: Primary structure and secondary structure annotation of chain A of the protein 1JC9 (PDB access number)

has been engrained in protein science for a long time [Anson and Mirsky, 1925; Gutte and Merrifield, 1969; Daughdrill et al., 2005].

2.2 Disordered Proteins

The dominant view of structure-function paradigm has been challenged by recent experiments [Dunker and Obradovic, 2001]. Sometimes loops are missing from high-resolution structures, and these loops are known to be essential for functionality [Bloomer et al., 1978;

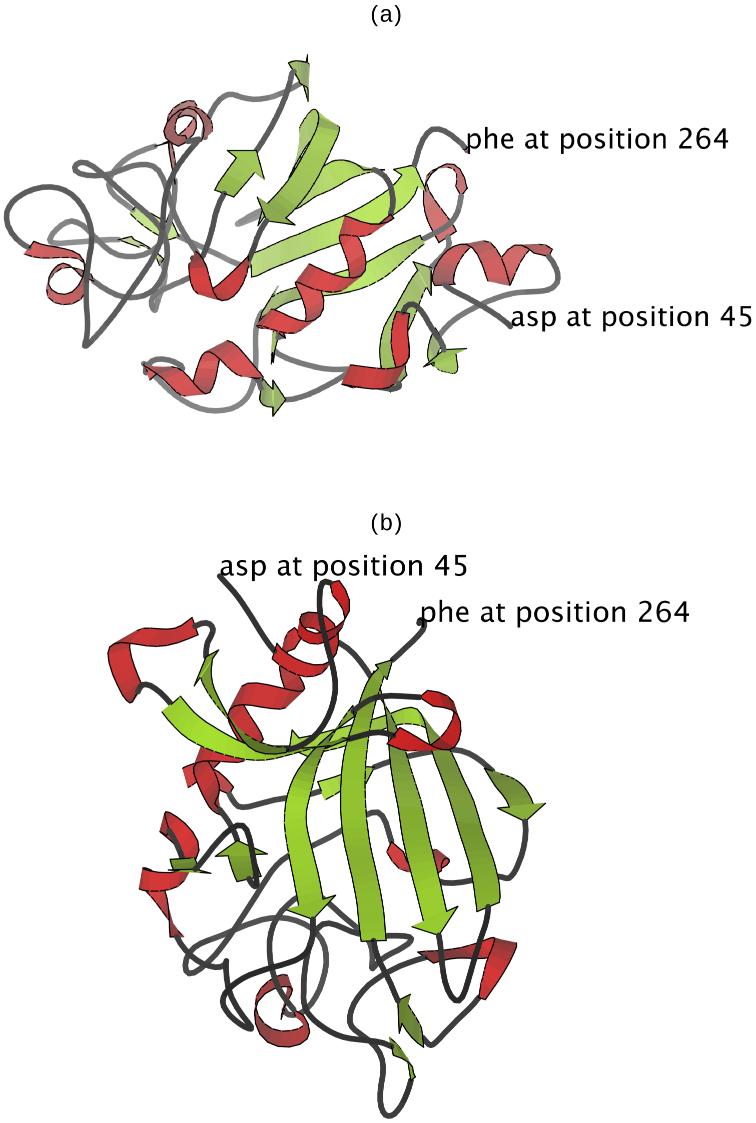


Figure 2.3: Tertiary structure of chain A of the protein 1JC9 (PDB access number)

Bode et al., 1978; Daughdrill et al., 2005]. Nuclear Magnetic Resonance (NMR) spectroscopy has shown that some proteins with known biological functions do not possess stable, defined structures in solution [Williams, 1978]. In contrast to the dominant view, these proteins display functions requiring the disordered state [Dunker et al., 2002a].

Generally, DRs are defined as entire proteins or regions of proteins that lack a fixed tertiary structure [Tompa, 2002; Ferron et al., 2006]. Nevertheless, lots of work has been done towards a more accurate “definition” of the term disorder and the creation of disorder databases. Due to the lack of a clear single definition of disorder, Linding et al. [2003a] have applied three definitions including loops/coils defined by DSSP, hot loop constitutes which are refined subsets of loops/coils, and missing coordinates in the X-Ray structure as defined in PDB. However, there are still limits to these three definitions. Loops/coils are not necessarily disordered and helices and sheets are not necessarily ordered. The definition of hot loops relates to temperature factors (B factors). However, these B factors can vary due to the effects of local packing and the structure environment. Missing coordinates in the X-Ray structure can be attributed to many reasons and may not be caused by disorder.

The CASP competition defines a residue as disordered if it meets either of the following two requirements [Bordoli et al., 2007]:

- X-Ray: no coordinates present for crystallised residues (SEQRES).
- NMR: residues whose conformation is not sufficiently defined by NMR restraints.

The first requirement of disorder is one of the definitions by Linding et al. [2003a] and has been widely used. However, segments without coordinates from crystallised residues are generally biased towards short regions. Considering the second definition of DRs from the CASP competition, the total number of structures solved by NMR is quite limited compared to that found by X-Ray in PDB. According to the PDB statistics in 2009, only 12.7% of protein structures are solved by NMR, while 86.8% of structures are solved by X-Ray.

The DisProt database (`http://www.disprot.org/`) [Sickmeier et al., 2007] is a published curated dataset in which each sequence contains at least one DR through searching the relevant literature and biological databases.

Dunker et al. [2002a] observe that backbones of ordered protein regions have the same Ramachandran angle, while disordered protein regions have different, often dynamic Ramachandran angles. DRs are divided into two major classes [Dunker et al., 2002a]:

1. Random coil like DRs. The backbones of proteins appear to be highly extended and similar to β sheets. However, due to the lack of long range contacts, the backbone angles fluctuate rapidly, which is different from β sheets. Side chains of this kind of DR exhibit motional characteristics more like those of random coils but with a backbone secondary structure more like that of the ordered state [Dunker et al., 2001].

2. Molten globule like DRs. Molten globules [Ohgushi and Wada, 1983] include various types of partially folded protein states found in mildly denaturing conditions or at high

> temperature. In general, molten globules have some proportion of secondary structures, but do not have stable tertiary structures. Molten globule like DRs include both molten globular regions and similar kinds of regions in which secondary structures do not exist.

The definition of disorder from Dunker et al. [2002a] tends to be more comprehensive and clearer than other definitions of disorder. Especially, molten globule like DRs suggest that some secondary structures can be disordered if they are unable to form stable three dimensional structures.

Different terms have been used to describe proteins or regions that fail to form specific three dimensional structures. These have been labelled as flexible [Pullen et al., 1975], mobile [Cary et al., 1978], partially folded [Linderstrom-Lang and Schellman, 1959], natively denatured [Schweers et al., 1994], natively unfolded [Weinreb et al., 1996], intrinsically unstructured [Wright and Dyson, 1999], and intrinsically disordered [Dunker et al., 2001].

According to Daughdrill et al. [2005], none of these terms or combinations are completely appropriate, however, the terms (intrinsically/natively) unstructured/disordered [Wright and Dyson, 1999; Dunker et al., 2001; Weathers et al., 2004; Thomson and Esnouf, 2004; Dyson and Wright, 2005; Tompa, 2005; Dosztányi et al., 2005a; Vullo et al., 2006; Feng et al., 2006; Schlessinger et al., 2007b] and unfolded [Uversky, 2002a;b; Garbuzynskiy et al., 2004; Coeytaux and Poupon, 2005; Prilusky et al., 2005; Galzitskaya et al., 2006a] have been more commonly used than others. There is substantial overlap among the terms unfolded, unstructured and disordered in the context of disorder. Both unfolded and unstructured imply a lack of backbone organisation [Daughdrill et al., 2005]. "Intrinsically" and "natively" are also used interchangeably.

Despite the complexity of definitions and terms, DRs present some well known properties. A series of papers [Romero et al., 1997b; Li et al., 1999; Romero et al., 2001; Dunker and Obradovic, 2001; Ferron et al., 2006] show that DRs have low sequence complexity, biased amino acid composition and high flexibility, being devoid of a stable secondary structure.

A variety of functions are related to DRs [Dunker et al., 2002a] including DNA/RNA/protein recognition, modulation of specificity/affinity of protein binding, molecular threading and activation by cleavages.

Proteins with DRs occur widely in nature. The DR predictor PONDR [Romero et al., 1997a;b; Li et al., 1999; Romero et al., 2001] has been used to predict the percentage of sequences with long DRs ($\geq$ 40 consecutive disordered residues) in eukaryote, bacteria and archaea genomes [Uversky et al., 2005]. The eukaryote proteins exhibit the most disorder content and 52-67% of these proteins were predicted to contain long regions of disorder, while 16-45% and 26-51% of bacteria and archaea proteins were predicted to contain long DRs [Uversky et al., 2005]. Ward et al. [2004] predict that long (>30 residue) DRs occur in 2.0% of achaean, 4.2% of eubacterial and 33.0% of eukaryotic proteins. All these results highlight that disorder appears to be common in nature and correlates with "biological complexity" [Dunker et al., 2000; Russell and Gibson, 2008]. This agrees with biological findings that unstructured proteins are more related to the signalling, regulation or control events in higher organisms and rarely have a role in routine processes such as metabolism [Russell and Gibson, 2008].

Different experimental methods can be used to characterise natively disordered proteins. NMR spectroscopy and crystallography are two commonly used approaches. One difficulty in using NMR is that different protons can have the same or very similar chemical shifts, which become greater as a protein becomes larger. For this reason NMR is usually restricted to small proteins or peptides. In addition, structure determination by NMR is a time consuming process, requiring interactive data analysis by trained scientists. X-Ray crystallography determines the arrangement of atoms within a protein crystal by analysing angles and intensities of scattered X-Ray beams. From angle and intensity information, a crystallographer can produce a three-dimensional picture of the density of electrons within a crystal. But as the crystal becomes larger and more complex, the picture provided by X-Ray crystallography becomes less well-resolved. In some cases it is hard to crystallise a protein because of disorder regions, but not every failure of crystallisation attributes to disorder [Shimizu et al., 2007b].

Thus, experimental approaches to DR determination are time consuming, complicated and expensive compared with automatic predictors. They are not applicable to large scale analysis of disorder, nor are they suitable techniques for some macromolecules.

It has been reported that long DRs present different properties from short DRs. Obradovic et al. [2005] reported that, given the training feature AAC, predictors built on different length groups of DRs outperformed the all-length predictor by 9-14%. This indicates a length dependency of the AAC of DRs. In another study [Radivojac et al., 2004], a set of short DRs ($\leq$10 residues) were extracted from PDB and then compared to long DRs, as well as regions of high B-factor order and low B-factor order. This study showed that short DRs exhibited significantly different AACs compared to long DRs and appeared to be more similar to the high B-factor ordered regions.

In the next several sections, we will review four types of disorder prediction:

1. Long DRs: pinpoint the long DRs in proteins.
2. Short DRs: pinpoint the short DRs in proteins.
3. Comprehensive DRs: predict DRs of various lengths.
4. Global disorder: predict protein sequences or chains that are largely disordered.

For each type, we will make further divisions according to the methods used for different predictors. We are especially interested in the machine-learning models that are used. Training databases, feature spaces and accuracies of different predictors are also introduced. However, accuracies of predictors are not always comparable, because different definitions of disorder [Cheng et al., 2005b; Peng et al., 2005] are adopted and various collections of disorder/order proteins are applied as training data.

Some databases and terms have been introduced previously and others that are used in the following sections will be explained briefly here. Further detail will be presented in Chapter 3. First, a list of databases is explained. PDB-select-25 [Hobohm and Sander, 1994] is a subset of structures obtained from PDB [Berman et al., 2000] that shows less than 25% amino acid sequence homology. The CASP7 (Critical Assessment of protein Structure Prediction) database refers to the disordered database used in the seventh CASP competition. The SCOP (Structural Classification of Proteins) [Murzin et al., 1995] database provides a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structures are known. The MSD (Macromolecular Structure database group) [Boutselakis et al., 2003] changed its name to PDBe (Protein Databank in Europe) in 2009, which is a project for the collection, management and distribution of data about macromolecular structures, derived from PDB. SWISS-PROT [Bairoch and Apweiler, 2000], a manually curated biological database of protein sequences, was created in 1986 and has since been developed by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). TrEMBL (Translated EMBL) [Bairoch and Apweiler, 2000] is a computer-annotated protein sequence database supplementing the SWISS-PROT protein sequence data bank. As translation is not entirely perfect, proteins predicted by the TrEMBL database can be hypothetical. The PIR (Protein Information Resource) [Barker et al., 1998] was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource mainly oriented to assist the propagation and standardisation of protein annotation. UniProt (Universal Protein Resource) [Consortium, 2007] is the central resource for storing and interconnecting information from large and disparate sources and databases. It is the most comprehensive catalogue of protein sequences and functional annotation and it is collaboration between the EBI, SIB and PIR. Finally, NRL-3D [Pattabiraman et al., 1990] is a sequence-structure database derived from PDB and searchable within PIR.

A list of terms is explained below. Solvent accessibility refers to surface area exposed to solvent surrounding the protein in its native form. The PSSM (position specific scoring matrix) is a type of scoring matrix in which amino acid substitution scores are given separately for each position of a sequence based on its multiple sequence alignments. The ROC curve is the abbreviation for the receiver operating characteristic curve. It is a graphical curve of the sensitivity versus (1-specificity) for a binary classification system given different thresholds.

2.3 Prediction of Long DRs

We divide long DR predictors into three types according to their prediction models:

1. The Dunker lab has adopted Neural Network (NN) and ensemble learning based on NNs to build a number of DR predictors, including the PONDR[2] (Predictors of Nature DR) [Romero et al., 1997a;b; Li et al., 1999; Romero et al., 2001] and DisProt[3] [Vucetic et al., 2003; Radivojac et al., 2003; Obradovic et al., 2003; Peng et al., 2005] series of predictors.

2. Predictors such as Spritz [Vullo et al., 2006] and POODLE-L [Hirose et al., 2007] adopt SVM or ensemble learning based on SVM to build the model.

3. Predictors that do not apply machine learning models. Predictors IUPred [Dosztányi et al., 2005b] and Ucon [Schlessinger et al., 2007b] are based on pairwise interaction energy and amino acid contact potential. The predictor NORSp [Liu et al., 2002; Liu and Rost, 2003] is a “rule based” predictor, annotating DRs by finding no regular secondary structure regions.

2.3.1 NN and Ensemble Learning Based Predictors

PONDR is the first tool designed specifically for the prediction of protein disorder. Some DR predictors in PONDR are integrated with DisProt, a series of predictors mainly for long DR prediction developed by the Dunker lab. Given that NNs have been applied in almost all PONDR and DisProt predictors, we will describe them together.

[2] `http://www.pondr.com`

[3] `http://www.ist.temple.edu/disprot/Predictors.html`

The default predictor of PONDR is VL-XT [Romero et al., 1997a; Li et al., 1999; Romero et al., 2001], an integrated long DR predictor combining a predictor for the internal region (PONDR VL1 [Romero et al., 1997a]) and two predictors for the N- and C- terminal regions. To enable prediction from the first to the last residue in a protein, VL1 and predictors for N- and C-terminal regions predict their respective domains. Structured regions and DRs in the training dataset of VL1 are constructed from PDB. Ten features selected from AAC, hydropathy scales, flexibility index, α-moment and β-moment are applied to train a NN. VL-XT achieves a 5-fold cross-validation accuracy of around 83%. Three independent predictors for the prediction of the N, C and internal regions are the major contributions of VL-XT.

The XL1 [Romero et al., 1997b] of PONDR has three predictors to investigate the accuracy of DR prediction for different lengths. Both ordered and disordered training datasets are built by searching PDB and selecting a set of proteins each having at least one DR with more than seven residues [Romero et al., 1997b]. These DRs are partitioned into three datasets according to their length and are called the SDR (Short Disorder Region), MDR (Medium Disorder Region) and LDR (Long Disorder Region) predictors, respectively. Ten features are selected from the AAC, flexibility [Vihinen et al., 1994], hydropathy [Kyte and Dolittle, 1982] and hydrophobic moments [Eisenberg et al., 1982] to train predictors based on SDR, MDR and LDR. For each predictor, a NN architecture is determined with ten inputs, one hidden layer of six units and a single output unit. The accuracy of 5-fold cross-validation can reach 71%, 76% and 77% for SDR, MDR and LDR respectively while the model trained by total DRs has a 5-fold cross-validation result of only 62%. These results reveal that DRs of different lengths have different AAC, hydropathy and flexibility properties and it is worthwhile building separate predictors for DRs of different length.

DisProt includes four major predictors VL2, VL3, VL3H and VL3E (predictor VL3P is part of VL3E) mainly for the prediction of long DRs.

In VL2 [Vucetic et al., 2003], disordered proteins are divided into different groups (flavours) through a competition algorithm. The partition procedure iterates until a stable partition is obtained, where the further partition of proteins will not lead to a higher accuracy of prediction. The training and prediction features include AAC and Shannon entropy [Shannon, 1948] calculated from the AAC in windows. The disordered and ordered training sequences are derived from PDB-Select-25. As a result, three different flavours of disorder are found. The general disorder predictor trained by total DRs achieves accuracies of prediction of 84.3% and 71.5% for ordered and disordered sequences. In contrast, three flavour specific predictors achieve accuracies of prediction of 81.5%-85.8% and 83.1%-86.9% for ordered and disordered sequences, which increases the prediction accuracy of disordered residues significantly, while maintaining the prediction accuracy of ordered residues.

VL3 [Radivojac et al., 2003], VL3H [Obradovic et al., 2003; Peng et al., 2005] and VL3E [Peng et al., 2005] of DisProt are all based on the ensemble of NNs and share a similar training dataset. The set of disordered training proteins is based on the training dataset of VL2 [Vucetic et al., 2003] introduced above. The set of ordered proteins is selected from the PDB. For each of the three predictors, a set of NNs is first trained on a balanced order and disorder dataset randomly sampled from examples available, before a majority voting scheme is employed to combine them into a single predictor.

The VL3H predictor uses the same features and prediction model as VL3 and hypothesizes that including homologues in the disorder training dataset would improve prediction accuracy. Each sequence in the disordered dataset is aligned to the original datasets of Vucetic et al. [2003] in order to expand the training dataset.

VL3E is a simple combination of results of prediction from predictors VL3H and VL3P. All predictors from VL3H and VL3P are used by VL3E for the majority voting procedure. VL3P utilises evolutionary information and PSSM as its main features. For a sequence of length N, an $N \times 20$ family profile is constructed based on the multiple alignment of homologues found during a PSI-BLAST [Altschul et al., 1990] search.

The overall accuracy of four predictors (VL2, VL3, VL3H and VL3E) is 80.9%, 83.6%, 84.7% and 86.7%, respectively. Thus, the more powerful ensemble prediction model used for VL3 results in a significantly higher accuracy over the VL2 predictor. In addition, the improvement from VL3 to VL3E is 3.1%, which can be attributed to the evolutionary information used.

2.3.2 SVM and Ensemble Learning Based Predictors

Spritz [Vullo et al., 2006] is based on two SVM models in order to pinpoint DRs from sequences. Two models are trained and benchmarked on two datasets, representing long and short DRs. The long (> 45 residues) disordered training dataset is built from the DisProt dataset. The corresponding ordered dataset is built from PDB-Select-25. The short disordered training dataset is from PDB. This short training dataset is fairly imbalanced and only 3.2% of residues are disordered. Both classifiers use features including 20 amino acid frequencies computed from multiple sequence alignment and secondary structure prediction results.

As the short disordered training dataset is imbalanced, the SVM model applies asymmetric costs with a larger penalty for misclassification of disorder. Spritz achieves very balanced AUC accuracies of 82% and 85% in 5-fold cross-validation on short and long disordered datasets. Meanwhile, the performance of the short DR predictor on the long DR dataset and the long DR predictor on the short DR dataset has been investigated. Corresponding AUC values are only 60% and 80% respectively. These results emphasize the importance of separate predictors for long and short DRs. The performance of Spritz confirms the findings of predictor XL1, in which length specific predictors achieve a higher accuracy of prediction than predictors trained by complete DRs.

POODLE-L [Hirose et al., 2007] is a two-level ensemble learning SVM prediction system for predicting long DRs (≥ 40 residues). The ordered training dataset is collected from PDB. The disordered training data is collected from DisProt [Vucetic et al., 2005].

The first level SVM models predict the disorder of 40-residue segments in sequences. Given six feature groups including hydrophobicity, charge, sequence complexity, AAC, secondary structure and average number of contacts, there are 63 possible combinations. The ten best performing DR predictors of various feature combinations are selected as base predictors for the next level.

The second level SVM model divides the output of the first-level, the disorder probability, into ten groups using an increment of 0.1. Given a window, the number of residues in each group can be calculated, which then form the input features of the second level SVM model. With ten basic predictors at the first level and three different window sizes, the final disorder probability of an amino acid is calculated as the average value over 18 probabilities (the two largest and smallest probabilities of each window size are omitted). Cross validation results of POODLE-L illustrate that the two-level SVM model can find 1.7% more disordered residues over the one level approach, while the accuracy on ordered residues is the same. Blind test accuracy of POODLE-L reaches 87.3%, higher than other predictors including VL3H, VSL2, DISOPRED2, FoldUnfold, IUPred, RONN, FoldIndex and DisEMBL.

2.3.3 Non Machine Learning Predictors

Several predictors are based on the idea that DRs cannot fold because their amino acids cannot make inter-residue interactions sufficient enough to overcome the large unfavourable entropic penalty accompanying folding [Tompa, 2008]. Predictors based on this principle either apply simple statistical comparisons [Galzitskaya et al., 2006a], use pairwise potentials to predict the structural state of a sequence [Schlessinger et al., 2007b], or estimate the total potential inter-residue interaction energy of a chain [Dosztányi et al., 2005a;b].

IUPred [Dosztányi et al., 2005a;b] is based on the general view that the primary structure of a globular protein determines its native conformation, and therefore its minimum global energy in conformational space. DRs are predicted by estimating their total pairwise inter-residue energy [Dosztányi et al., 2005a]. The position-specific estimated total energy is calculated from the AAC of a segment (window) and a 20 (residues) by 20 (residues) interaction matrix. In the blind test, IUPred can achieve a higher or comparable accuracy of prediction to PONDR (VL3H), DISOPRED2 and GlobPlot. IUPred is ranked seventh with an AUC value of 77.7% among all server predictors in the CASP7 competition. Different parameters have been adopted by IUPred for prediction of long and short DRs.

Ucon [Schlessinger et al., 2007b] computes DRs from protein-specific contact and generic pairwise potentials. Ucon can identify unstructured regions involved in protein protein binding, that may be missed by other methods. Two datasets have been applied to measure the performance and optimise the accuracy of prediction. The disordered dataset contains proteins from DisProt (version 3.0) and the ordered dataset is built from PDB. The prediction method contains three major steps. First, a two-dimensional contact map of each protein is predicted by PROFcon [Punta and Rost, 2005] and is a a predictor of long-range contacts in a protein. Each dot in the two-dimensional map represents a predicted probability of two residues interacting. Then, the two-dimensional map is multiplied with statistical pairwise potentials to derive a position-specific score for a sequence. Finally, a threshold is defined, above which a residue is labelled as unstructured. In the 5-fold cross-validation, Ucon can achieve an AUC value of 91.2%. Both Ucon and IUPred calculate an energy-related score, while Ucon is more accurate than IUPred with the successful application of prediction results of long range contacts. The energy-related score can be important to predict the lack of a regular structure.

NORSp [Liu et al., 2002; Liu and Rost, 2003] is a publicly available predictor for DRs in proteins. NORSp predicts long DRs by identifying regions without a regular secondary structure (NORS). NORS regions are identified by combining prediction results of PROFphd [Rost, 2001] for secondary structure prediction, PHDhtm [Rost et al., 1996] for transmembrane helix prediction and COILS [Lupas, 1996] for coiled-coil region prediction. As NORSp is essentially a secondary structure predictor, a potential problem of NORSp is that regions without a regular secondary structure are not necessarily disordered. Structures such as the Kringle domain (PDB code: 1krn) are almost entirely without a regular secondary structure in their native state, but still have a tertiary structure where the basic building block is coils [Linding et al., 2003a].

2.4 Prediction of Short DRs

The prediction of short DRs (<30 residues) is a different task to the prediction of long DRs [Romero et al., 1997b], mainly because short DRs have different features to long DRs. In addition, short DR annotations are less reliable and short DRs are insufficient. As will be shown in more detail in Section 3.1.4, most DRs of the CASP7 targets are short (<30 residues). We therefore consider many top ranked CASP7 predictors as short DR predictors because of their supreme performance in that competition. However, this does not imply these predictors cannot be applied to long DR prediction.

The training features of the five short DR predictors in this section are quite similar. Most of them have applied PSSM [Ward et al., 2004; Shimizu et al., 2005; Cheng et al., 2005b; Ishida and Kinoshita, 2007] and secondary structure prediction results [Ward et al., 2004; Cheng et al., 2005b]. However, the methods of prediction are quite different. Predictors POODLE-S [Shimizu et al., 2005], DISOPRED2 [Ward et al., 2004] and DISpro [Cheng et al., 2005b] use a single model for prediction, while PrDOS [Ishida and Kinoshita, 2007] and DISOclust [McGuffin, 2008] make a prediction from two different models. Loosely speaking, PrDOS and DISOclust can be considered as applications of ensemble learning.

IUPred [Dosztányi et al., 2005b], XL1 [Romero et al., 1997b] and Spritz [Vullo et al., 2006] have built both long and short DR predictors for DRs of different length. But in each of them, the difference between long and short predictors is not significant, which can be due to different parameters being used (IUPred) or different training sequences (XL1 and Spritz). Therefore, their corresponding short DR predictors will not be discussed in this section. The winning human expert of the CASP7 competition ISTZORAN [Peng et al., 2006] will be introduced in Section 2.5 due to its specific design to obtain high accuracy of prediction for both long and short DRs.

2.4.1 Single Model Predictors

CBRC-DR [Shimizu et al., 2005; 2007b; Hirose et al., 2007] is the runner up human expert in the CASP7 competition. CBRC-DR includes three predictors that are specialised for the prediction of global disorder (POODLE-W of Section 2.6), long DRs (POODLE-L of Section 2.3), and short DRs (POODLE-S [Shimizu et al., 2005]). The training dataset of

POODLE-S is derived from DisProt2.2 and PDB. Given that the AAC has different properties in different regions of sequences, a sequence is divided into seven regions. In each region, data with the same label has a similar AAC tendency (separated by its Chi-square score). PSSM is grouped based on selected physicochemical properties for each of seven regions. A grouped PSSM is calculated in a sliding-window and the features are fed into a SVM to build learning models. The accuracy of cross-validation reveals that predictors based on seven-region specific reduced sets of PSSM perform better than directly reduced sets of PSSM and slightly outperformed by the commonly used three-region (N- terminal, internal, C- terminal) specific reduced sets of PSSM. In the blind test on the CASP6 targets, the seven-region specific approach is clearly better than the other two approaches.

DISOPRED2 [Ward et al., 2004] is the best server predictor in the CASP7 competition with an AUC value of 83.7%. The training dataset of DISOPRED2 is from PDB. Residues with missing atomic coordinates are defined as disordered. It is imperfect, as missing residues can also arise as an artefact of the crystallisation process, such as rigid body wobble or crystal contacts. However, this appears to be the most effective means of identifying DRs in the absence of further experimental characterisation of the protein sequence [Ward et al., 2004]. The training features of DISOPRED2 include sequence profile (PSSM) and prediction results from a secondary structure predictor PSIPRED [Bryson et al., 2005]. These features, under a sliding-window, are input to the first layer SVM learning model. The outputs of this SVM are input to the second level smoothing NN for the final results of the prediction. Results of cross-validation tests show that DISOPRED2 can achieve an AUC value of 86.8%.

DISpro [Cheng et al., 2005b] is the runner up server in the CASP7 competition. Its training dataset is built from PDB. The training features of DISpro include sequence profile (PSSM), predicted secondary structure and relative solvent accessibility. DISpro has applied a novel 1D-Recursive NN (1D-RNN) as the learning model. The 1D-RNN can handle inputs with variable length and allow classification decisions to be made based on information outside of the traditional local input window widely used by most DR predictors. DISpro achieves an AUC value of 82.2% in the CASP7 competition. DISpro has been updated by Hecker et al. [2008], where a more recent training dataset is built from PDB. The updated DISpro achieves a higher accuracy of prediction than that of DISOPRED2. As indicated by Cheng et al. [2005b], DISpro is significantly more accurate for the prediction of short DRs than for long DRs.

2.4.2 Two Model Predictors

The third most accurate human expert in CASP7 is Fais [Ishida et al., 2006; Ishida and Kinoshita, 2007] with an AUC value of 84.4%. The corresponding server of Fais is PrDOS [Ishida and Kinoshita, 2007], a prediction system composed of two component predictors, based on local amino acid sequence information and template proteins.

The training dataset of PrDOS is built from PDB. Both component predictors in PrDOS use the PSI-BLAST search to generate a PSSM. The input features of the first predictor are PSSM information and spaces annotating whether a window is beyond termini. These inputs are sent to the SVM to generate the model of the first predictor. The second predictor is template-based. The prediction is made by the alignments of homologues with structures determined proteins. Finally, the results of the two predictors are combined by weighted average. The major contribution of PrDOS is the second predictor, a novel alignment based DR predictor without a training dataset or machine learning model.

DISOclust [McGuffin, 2008] is an unsupervised method of DR prediction that investigates disorder prediction from a novel three-dimensional structure-based approach. The premise of DISOclust is that ordered residues within a protein should be conserved in a three-dimensional space among multiple models, where residues that vary or are consistently missing may correlate with disorder. Given a query sequence, the corresponding fold recognition models are obtained from the LOMETS server [Wu and Zhang, 2007], which aligns sequences to structures. Then, prediction of disorder involves prediction of the per-residue error in the multiple fold recognition models followed by a simple analysis of the conservation of per-residue error across all models. This conserved error is used to represent the probability of a residue being in order/disorder status. The performance of DISOclust has been measured on two datasets: CASP7 and the DisProt supplemental dataset[4]. On the DisProt supplemental dataset, only 23 (11.6%) sequences contain long regions of disorder (>30 residues). The AUC value of DISOclust itself on the CASP7 targets is 81.66%. However, when DISOclust is combined with DISOPRED2, a consensus taking the average of scores from DISOclust and DISOPRED2, the AUC value can reach 88.02%, higher than all predictors in the CASP7 competition. However, the AUC value of DISOclust on 23 selected DisProt targets containing long DRs is only 69.6%, which is significantly lower than the DisProt subset containing only short regions (<30 residues) of disorder with an AUC value of 78.6%.

[4] `http://www.disprot.org/data/missingxray/missingXray.080503.zip`

2.5 Comprehensive Prediction of DRs

There are generally two categories of predictors for both long and short DR prediction in sequences. Some recent disorder predictors build several member predictors specific for short and long DRs. Given a query sequence, these ensemble member predictors initially make predictions and then their outputs are combined by a meta predictor to give the final results of prediction. Alternatively, some disorder predictors use existing disorder predictors to predict training databases. The output of these predictors is input to the second level ensemble learning model. According to the accuracy of cross-validation, these ensemble learning based predictors are generally more accurate than disorder predictors that do not apply ensemble learning.

Many early DR predictors [Linding et al., 2003a;b; Yang et al., 2005; Coeytaux and Poupon, 2005] were not designed specifically for long or short DR prediction. They use training datasets with a relatively balanced number of residues from long and short DRs. We believe this may due to the limited knowledge of DRs during the early stage of DR prediction. Generally, these predictors have contributed lots of innovative ideas, however they tend to be less accurate than recent comprehensive DR predictors tailored to improve the accuracy of prediction for both long and short DRs.

In this section we will first discuss the recent comprehensive DR predictors and then describe some representatives of early comprehensive DR predictors.

2.5.1 Ensemble Learning Based Predictors

The two most highly ranked predictors in CASP7 are ISTZORAN [Peng et al., 2006] and CBRC-DR [Shimizu et al., 2005; Hirose et al., 2007; Shimizu et al., 2007b]. Both build component predictors for long and short DRs first, and outputs from component predictors are combined to build a final predictor. The combination procedure of CBRC-DR is not published and we will initially discuss ISTZORAN in this section. Then, the most recent methods of GeneSilico MetaServer [Rowski and Bujnicki, 2003], metaPrDOS [Ishida and Kinoshita, 2008] and MD [Schlessinger et al., 2009] are introduced.

ISTZORAN, registered as a human expert, won the CASP7 competition. The DR prediction server of ISTZORAN is VSL2 [Peng et al., 2006]. It designs two predictors for prediction of long (>30 residues) and short DRs (≤30 residues), specifically. Then, a meta predictor is trained to integrate the specialised predictions into the final prediction. The disordered and ordered training datasets of ISTZORAN are derived from DisProt (version 1.2) and PDB.

To build two length specialised predictors, DRs of more than 30 residues are used to train the long DR predictor VSL2-L, and DRs of less than 30 residues are used to train the short DR predictor VSL2-S.

For the two component predictors VSL2-L and VSL2-S, features are calculated for each residue with a sliding window centred at that residue. These features include AAC, PSSM and results of secondary structure prediction from predictors PHDsec [Rost and Sander, 1993b;a] and PSIPRED [Jones, 1999; Bryson et al., 2005]. Compared with the long DR predictor, a smaller window is used for the short DR predictor. A meta predictor, M1, is trained to assign the appropriate weight to prediction results from VSL2-L and VSL2-S. The ten-fold cross-validation results show that VSL2 can achieve an AUC value of 90.5%. With the support of human expert knowledge, the blind test accuracy of ISTZORAN on the CASP7 targets is 86% for AUC.

The GeneSilico MetaServer [Rowski and Bujnicki, 2003] is the server ranked third in the CASP7 competition, with an AUC value of 80.4%. As a web server for protein disorder prediction, it facilitates access to several structure prediction methods through a single web interface. It claims that from the results of assessments, better structure prediction can be obtained after combining results produced from several different methods given that they have different strengths and weaknesses [Rowski and Bujnicki, 2003]. However, the details of which DR predictors have been applied are not published.

metaPrDOS [Ishida and Kinoshita, 2008] processes the results of prediction of seven independent DR predictors with a SVM model. These predictors include PrDOS, DISOPRED2, DisEMBL, VSL2P [Peng et al., 2006], DISpro, IUPred and POODLE-S. The training dataset of metaPrDOS is built from PDB. At the start of training, each predictor predicts all chains and the numeric score of disorder of each residue is obtained. Then a SVM model integrates the scores and builds the final model. According to the results of a ten-fold cross-validation test, metaPrDOS can achieve an AUC value of 90.4%, higher than the most accurate component predictor (88.7%). In the blind test on the CASP7 targets, metaPrDOS achieves an AUC value of 87.7% and outperforms other prediction methods including ISTZORAN and CBRC-DR.

MD [Schlessinger et al., 2009] obtains results of prediction from four independent DR predictors including DISOPRED2, IUPred, NORSnet [Schlessinger et al., 2007a] and Ucon. Results of prediction of these independent predictors are combined with features including prediction results of flexibility, predicted secondary structure, local sequence profiles, solvent accessibility, the presence of low complexity regions, AAC and sequence length as inputs to a standard feed-forward NN with backpropagation [Rost and Sander, 1993a]. The training dataset of MD is derived from DisProt (version 3.4). The improvement of MD over constituent DR predictors is more than 3% in AUC.

In summary, both metaPrDOS and MD are successful ensemble predictors built from existing DR predictors. From the perspective of learning, this approach has made use of bagging [Breiman, 1996] and boosting [Freund and Schapire, 1996].

2.5.2 Non Ensemble Learning Based Predictors

Some early DR predictors can be considered as comprehensive DR predictors including DisEMBL [Linding et al., 2003a], RONN [Yang et al., 2005], PreLink [Coeytaux and Poupon, 2005] and GlobPlot [Linding et al., 2003b]. They typically do not apply ensemble learning. DisEMBL [Linding et al., 2003a] and RONN [Yang et al., 2005] are based on a NN, and both PreLink and GlobPlot adopt a "rule based" approach calculated from physicochemical properties of amino acids without using machine learning models.

Three definitions of disorder are proposed in DisEMBL: loops/coils defined by DSSP [Kabsch and Sander, 1983], hot loops, a refined subset of loops/coils that has a high degree of mobility, and residues with missing coordinates in X-Ray structures. As a result, three different predictors are built. The predictor based on the loops/coils definition is essentially a secondary structure predictor. For the prediction of ordered and disordered hot loops, separate networks are trained. For the prediction of missing coordinates, an ensemble of two sets of networks is formed. Accuracy of cross-validation shows that at a sensitivity of 64%, hot loops is the most accurate predictor with a corresponding false positive rate of only 1.3%. Loop/coil and missing coordinates have much higher false positive rates at 10% and 16%, respectively. In comparison with PONDR, the missing coordinates predictor achieves marginally higher accuracy than the PONDR predictors.

The Regional Order NN (RONN) [Yang et al., 2005] is an application of the "bio-basis function neural network" [Thomson et al., 2003] pattern recognition algorithm to the detection of DRs in proteins. More specifically, RONN aligns query sequences to sequences of known order/disorder states. Scores of alignment are used to classify query sequences by a trained NN.

PreLink [Coeytaux and Poupon, 2005] predicts DRs based on two properties of DRs. First, DRs have a biased amino acid composition. Second, DRs usually contain no or small hydrophobic clusters (segments containing mainly hydrophobic residues). The DRs are called "linker regions" in PreLink, meaning that no three-dimensional coordinates are available for these regions. The linker set (L) is created by aligning the PDB protein sequences with corresponding SWISS-PROT protein sequences and extracting the non-aligned fragments. The reference structured set (S) is the ensemble of PDB sequences. Three rules for prediction of DRs are derived based on the probability ratio (calculated from the AAC) and the hydrophobic cluster distances. In the blind test on the CASP5 targets, PreLink achieves a sensitivity value of 87% when the specificity value is 99.8%.

GlobPlot [Linding et al., 2003b] applies a propensity scale to explore potential globular (ordered) and flexible (disordered) regions in protein sequences. The globular or disordered status depends on the running sum of the propensity from amino acids. The disordered training set of proteins is taken from the SCOP database (version 1.59) (`http://scop.mrc-lmb.cam.ac.uk/scop/`) [Conte et al., 2002; Chandonia et al., 2002]. The ordered set contains residues extracted from DSSP [Kabsch and Sander, 1983]. For each residue, the tendency for disorder can be expressed as $P = RC - SS$ where RC and SS are the propensity for a given amino acid to be in "random coil" and regular "secondary structure", respectively. GlobPlot is essentially an optimised secondary structure predictor. To decide the final disorder propensity of a target residue, the running sum method adds up the disorder tendencies of residues before the target. The globular and disordered segments are selected using a simple peak finder algorithm.

2.6 Prediction of Global Disorder

The binary global classification of proteins as largely disordered or not is a gross approximation of the actual biological situation. The prediction of global disorder is relatively easy as the precise locations of DRs do not need to be specified.

Global disorder predictors in the literature are generally trained from completely ordered segments from PDB and disordered segments from DisProt and PDB. Prediction models of global disorder predictors are quite different. Some predictors [Weathers et al., 2004; Shimizu et al., 2007b] adopt machine learning approaches to build models from training data, that are then used to predict query sequences. Other predictors [Prilusky et al., 2005; Galzitskaya et al., 2006a] make predictions by calculating physicochemical properties of query sequences, and no machine learning models are involved. The later predictors can achieve comparable accuracies of prediction with machine learning based predictors, but with better human understandability and time efficiency.

2.6.1 Machine Learning Model Based Global Disorder Predictors

Machine learning approaches have been used to predict global disorder. AACs or reduced-AACs are the commonly used features. Weathers et al. [2004] applied a SVM model trained on disordered and ordered proteins to examine the contribution of various parameters (vectors) in recognising disordered protein sequences. AACs and different reduced-AAC groups are used to construct the predictor. The AAC based model can achieve an accuracy of 87±2%, while the reduced-AAC based models can achieve an accuracy between 62±3% and 82±1%.

In general, Weathers et al. [2004] have illustrated the importance of AAC and reduced-AAC in DR prediction and have demonstrated that SVM can be a useful machine learning model for the prediction of disorder. They have also shown that hydrophobicity groups give a better indication of disorder than other groups such as charge.

POODLE-W [Shimizu et al., 2007b] is a more recent predictor, predicting mostly disordered proteins by using structure-unknown protein data. It applies a Spectral Graph Transducer (SGT) [Joachims, 2003], a binary classification algorithm based on the K-Nearest-Neighbour (KNN) graph and semi-supervised learning. Various types of measurements of the nearest neighbour are applied including ACC, physicochemical properties and sequence similarity.

POODLE-W is trained by structure solved sequences as well as structure unknown sequences. The disordered (positive) and ordered (negative) proteins are from DisProt and PDB. The unlabelled protein dataset is from UniProt 50, SWISS-PROT and TrEMBL. The result of the five-fold cross-validation test reveal that this predictor successfully located 72.3% of disordered sequences, while only 2.3% of ordered sequences were misclassified.

2.6.2 Global Disorder Predictors Based on Sequence Physicochemical Properties

Approaches that do not use machine learning models are very time efficient for disorder prediction. Web servers FoldIndex [Prilusky et al., 2005] and FoldUnfold [Garbuzynskiy et al., 2004; Galzitskaya et al., 2006a;b;c] predict disorder based on physicochemical properties without using machine learning models. FoldIndex implements the algorithm of Uversky et al. [2000] and makes a binary prediction of order/disorder based on average residue hydrophobicity and the net charge of sequences. The ordered dataset of FoldIndex is from SWISS-PROT [Bairoch and Apweiler, 2000] and its supplement, TrEMBL. The database of natively unfolded proteins is built from proteins that are reported as being close to typical unfolded polypeptide chains. Given a protein sequence, a high mean hydrophobicity and low absolute value of the mean net charge indicate that the protein is likely to be ordered.

FoldIndex can also be used to predict the binary status of each residue in sequences with the help of a sliding window including a small segment of residues. FoldIndex is the first attempt at quantifying the relationship between hydrophobicity, net charge and DR with a simple equation. Independent blind evaluation successfully predicts 77% of unfolded and 88% of folded proteins. This accuracy is better than or comparable to some machine learning based approaches by Linding et al. [2003b] and Jones and Ward [2003]. Given the extremely fast speed of prediction, FoldIndex is a suitable baseline predictor for the evaluation of disorder predictors.

As natively unfolded proteins do not have sufficient energetic interactions to form a stabilised conformation, FoldUnfold calculates the expected average number of contacts per residue from an amino acid sequence to predict whether a protein is folded or unfolded. More specifically, the expected average number of contacts is the sum of the average contacts of all residues, divided by the number of residues in the amino acid sequence.

The ordered training dataset of FoldUnfold includes ideally folded proteins from PDB and SCOP [Murzin et al., 1995]. The database of natively unfolded proteins includes proteins reported as being close to typical unfolded chains by Uversky et al. [2000] and proteins from SWISS-PROT [Bairoch and Apweiler, 2000]. The prediction accuracy of FoldUnfold is 89%, which exceeds that of the hydrophobicity index [Kyte and Dolittle, 1982] (83%).

2.6.3 Cumulative Distribution Function Based Global Disorder Predictors

The Cumulative Distribution Function (CDF) [Sprent, 1993] describes the probability that a variant X takes on a value less than or equal to a number x. The CDF has been applied to the prediction of completely disordered proteins by Dunker et al. [2000] and Oldfield et al. [2005a]. These predictors rely on the results of prediction from DR predictors that specify the probability of disorder for each residue in a query sequence.

A global DR predictor based on CDF was proposed by Dunker et al. [2000]. The disordered dataset contains completely disordered chains selected from PDB, SWISS-PROT and PIR [Barker et al., 1998]. The ordered database is constructed from randomly selected segments in NRL-3D [Pattabiraman et al., 1990].

The DR predictor VL-XT described in Section 2.3.1 makes a prediction for a sequence first. Then the predicted scores of all the residues in the sequence can be used to draw a cumulative histogram. A CDF curve can be calculated from the histogram. At any point on the CDF curve, the ordinate value gives the proportion of residues with a score less than or equal to the abscissa value. Given that scores of prediction are between 0 and 1, the corresponding curve always begins at the point (0,0) and ends at the point (1,1). Proteins with higher scores of prediction have curves with low cumulative values and proteins with lower scores of prediction have curves with high cumulative values. Thus, a boundary line can be determined to separate completely ordered and disordered proteins. The boundary line (composed of points) is calculated by minimising the total error in the training dataset. The overall accuracy of global prediction can reach 83.3% [Dunker et al., 2000].

This CDF based predictor [Dunker et al., 2000] has also been combined with a net charge-hydropathy distribution predictor [Uversky et al., 2000] by Oldfield et al. [2005a] to predict mostly disordered proteins. This new method makes a prediction according to the consensus of the two predictors.

Experiments report that the classification accuracy with net charge-hydropathy distribution is 83% overall; 76% for disordered proteins and 91% for ordered proteins. The classification accuracy for CDF (VL-XT score) is 88% overall, while 87% and 90% wholly disordered and ordered proteins, respectively, are correctly classified. CDF is generally more accurate at predicting completely disordered and ordered proteins. Oldfield et al. [2005a] use a weighted combination of prediction results from both CDF and net charge-hydropathy distribution predictors. This method gives very balanced accuracy of prediction and 95% and 90% of completely ordered and disordered sequences, respectively, are correctly predicted. Obviously this consensus method achieves a higher accuracy than when using two base predictors.

2.7 Summary and Discussion

In this section, we will summarise the predictors discussed in previous sections. The best known DR predictors for long DRs, short DRs, comprehensive DRs and global disorder predictions are listed in Tables 2.2, 2.3, 2.4 and 2.5. Each predictor is described by its URL, method of prediction, features and learning databases.

2.7.1 Methods

Given that machine learning and non-machine learning approaches have been applied to all four types of disorder predictors, we will discuss the association between these approaches and various types of disorder prediction.

Machine Learning Based DR Predictors

Machine learning based DR predictors rely on a training dataset with labelled disordered and ordered regions. A machine learning model is built from the training dataset which can then be used to predict query sequences. From Tables 2.2, 2.3, 2.4 and 2.5, NN and SVM learning approaches have been widely used in all four types of disorder prediction, while ensemble learning approaches are commonly used for long, short and comprehensive DR prediction.

- *Neural Network.* Most PONDR and DisProt long DR predictors developed by the Dunker lab are based on NN or ensemble learning models [Romero et al., 2001; 1997b]. Short DR predictors such as DISOPRED [Jones and Ward, 2003], DISpro and early comprehensive DR predictors such as DisEMBL and RONN are also based on NN.

  Roughly speaking, a NN involves a set of connected input and output units where there is a weight for each connection. During the learning phase, a network learns by adjusting the weights so as to predict the correct class of the input samples [Han and Kamber, 2000]. Advantages of NNs include high tolerance to noisy data and the ability to classify patterns which have not appeared in the training data. These merits make NNs suitable for the task of disorder prediction, which always has noisy training datasets due to experimental limitations and the inherent complexity of biological phenomenon.

- *SVM.* Many DR predictors have applied SVM to develop their prediction models. These predictors include the global disorder predictor proposed by Weathers et al. [2004], long DR predictors such as Spritz and POODLE-L, short DR predictors such as POODLE-S, DISOPRED2 and PrDOS and comprehensive DR predictors such as VSL2 and MD. Interestingly, some early predictors based on NNs have changed to use a SVM model in their latest version; for example, DISOPRED [Jones and Ward, 2003] and the predictors developed by the Dunker lab.

  Theoretically, a SVM learning model seeks an optimal separating hyperplane from data points of two different classes in space where the margin is maximal. Only those data points that are support vectors at the margin determine the hyperplane. SVM has the advantage of modelling complex problems accurately. It also performs well on datasets that have many attributes where NNs may not perform well under these circumstances.

- *Meta/Ensemble learning.* Meta-servers or meta-predictors apply the ensemble learning approach to tackle DR prediction problems. These predictors include long DR predictors, the VL3 series and POODLE-L, the short DR predictors DISOclust and PrDOS and comprehensive DR predictors GeneSilico, metaPrDOS, MD and VSL2.

  For example, metaPrDOS first finds component predictors, and then a SVM model is trained by outputs from these predictors before a decision plane is created. This process essentially corresponds to the boosting strategy except for the way in which the margin is measured or the way that the weight vectors are optimised [Rätsch et al., 2000]. Naturally, metaPrDOS also suffers from the shortcomings of bagging and boosting. The accuracy of an ensemble predictor depends on both its prediction accuracy and the variation in the component predictors [Breiman, 1996; 2001a]. Most of the seven predictors in metaPrDOS depend on similar features and use a similar algorithm, which may give closely related results of prediction and limit the improvement of the prediction accuracy of the meta predictor.

Non-Machine Learning Based DR Predictors

According to Tables 2.2, 2.3, 2.4 and 2.5 some DR predictors do not apply any machine learning approaches. Non-machine learning based DR predictors have the advantage of good understandability by humans and fast speed of prediction. Since short and comprehensive DR prediction tasks are relatively more complex and require more sensitive prediction, non-machine learning based DR predictors generally converge at global disorder and long DR prediction. The accuracy of some non-machine learning based global and long DR predictors [Galzitskaya et al., 2006a; Schlessinger et al., 2007b] is comparable with the best performing machine learning based DR predictors.

- *Hydrophobicity and net charge relation.* Uversky et al. [2000] observe that DRs usually have low mean hydrophobicity and high net charge. More importantly, this relationship has been quantified by a simple equation. With the introduction of a sliding window [Prilusky et al., 2005], this strategy enables the prediction of the disorder/order status of each residue in a given sequence.

- *Lack of secondary structure.* Native disorder can exist in structures such as extended random coil proteins with negligible secondary structure or in molten globules, which

Table 2.2: Long DR predictors (IUPred, Spritz and XL1 also have short DR predictor versions)

Predictor	Method	Features	Databases	
			Order	Disorder
IUPred [Dosztányi et al., 2005b] `http://iupred.enzim.hu/`	minimum global energy	20 by 20 interaction matrix	PDB-select	literature
NORSp [Liu and Rost, 2003] `http://cubic.bioc.columbia.edu/services/NORSp/`	detection of no regular secondary structure region	prediction results of secondary structure (alignment), transmembrane helices and coiled-coil regions	NA	NA
POODLE-L [Hirose et al., 2007] `http://mbs.cbrc.jp/poodle/poodle-l.html`	ensemble learning of two-level SVM model	physicochemical properties, sequence complexity, encoded AAC; outputs of first level SVM	PDB	DisProt
Spritz [Vullo et al., 2006] `http://distill.ucd.ie/spritz/`	nonlinear SVM, separate predictors based on different databases for short and long DR prediction	amino acid frequency from multiple alignment, predicted results of secondary structure	PDB-select, PDB	DisProt, PDB
Ucon [Schlessinger et al., 2007b] `http://www.rostlab.org/newwebsite/services/ucon/`	calculated amino acid contact potential	predicted residue-residue interaction (alignment), statistical pairwise potentials	PDB	DisProt
VL-XT [Romero et al., 2001] `http://www.pondr.com/`	three NN predictors for N, C terminal and internal areas	AAC and physicochemical properties	PDB, NRL-3D	PDB-select
VL2 [Vucetic et al., 2003] `http://www.ist.temple.edu/disprot/Predictors.html`	NN, ordinary least squares algorithm applied to training dataset partition	AAC and Shannon entropy	PDB-select	PDB-select
VL3 series [Obradovic et al., 2003; Peng et al., 2005] `http://www.ist.temple.edu/disprot/predictor.php`	ensemble learning, multiple NNs, majority voting scheme	AAC, Shannon entropy, flexibility, PSSM (VL3P, VL3E)	PDB	PDB-select
XL1 Romero et al. 1997b `http://www.pondr.com/`	three independent predictors predicting short, medium and long DRs	AAC, flexibility, hydropathy and hydrophobic moments	PDB	PDB

have regular secondary structure elements but do not condense into a stable globular fold. However, DRs generally include a high percentage of proline and frequently lack regular secondary structure [Romero et al., 2001; Vucetic et al., 2003; Radivojac et al., 2004; 2007; Schlessinger et al., 2009]. NORSp predicts long (>70 residues) regions devoid of secondary structure elements. Prediction results of secondary structure, solvent

Table 2.3: Short DR predictors

Predictor	Method	Features	Databases	
			Order	Disorder
DISOclust [McGuffin, 2008] http://www.reading.ac.uk/bioinf/DISOclust/	unsupervised method and calculate per-residue error from multiple fold recognition models	predicted multiple fold recognition models (alignment)	NA	NA
DISOPRED2 [Ward et al., 2004] http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	cascaded SVM classifier followed by NN for smoothing	PSSM and predicted secondary structure	PDB	PDB
POODLE-S [Shimizu et al., 2005] http://mbs.cbrc.jp/poodle/poodle-s.html	seven-region specific SVM predictors	reduced PSSM based on physicochemical properties	PDB	DisProt
PrDOS [Ishida and Kinoshita, 2007] http://prdos.hgc.jp	the weighted average of two DR predictors based on SVM and alignment template	PSSM and space for SVM predictor; results of multiple sequence alignment for template predictor	PDB	PDB
DISpro [Cheng et al., 2005b] http://scratch.proteomics.ics.uci.edu/	1D-recursive NNs	PSSM and predicted secondary structure and solvent accessibility	PDB	PDB

accessibility, membrane helices and coiled-coil are used to assign regular or non-regular secondary structures. Although the underlying principle of NORSp for DR prediction is yet to be theoretically justified, its accuracy is comparable to many other disorder predictors.

- *Contact potentials.* Several predictors are based on the idea that DRs cannot fold because their amino acids cannot make inter-residue interactions sufficient enough to overcome the unfavourable entropic penalty accompanying folding. The predictors based on this principle either apply statistical comparisons [Galzitskaya et al., 2006a], use pairwise potential to predict the structural state of a sequence [Schlessinger et al., 2007b] or estimate the total potential inter-residue interaction energy of a chain [Dosztányi et al., 2005a;b].

- *Propensity.* Propensity-based predictors assess if a given amino-acid feature is enriched or depleted within a segment defined by a sliding window over a sequence.

  GlobPlot first calculates the propensities for disorder/globularity databases, then a running sum function is applied, which determines the potential globular (ordered)

Table 2.4: Comprehensive DR predictors

Predictor	Method	Features	Databases	
			Order	Disorder
DisEMBL [Linding et al., 2003a] http://dis.embl.de/	NN learning models; three predictors based on three different definitions of DR	sequence in windows	PDB	DSSP, PDB
GeneSilico [Rowski and Bujnicki, 2003] https://genesilico.pl/meta2/	Meta server, combined results of several different methods	predicted results of several different methods	NA	NA
GlobPlot [Linding et al., 2003b] http://globplot.embl.de/	rule based, running sum of the amino acid propensity	amino acid propensity in random coil or secondary structure	DSSP	SCOP
MD [Schlessinger et al., 2009] http://cubic.bioc.columbia.edu/services/md/	NN learning from results of other DR predictors, secondary structure predictors, and sequence features	prediction results of other DR predictors, predicted secondary structure, flexility, solvent accessibility, PSSM, etc	DisProt	DisProt
metaPrDOS [Ishida and Kinoshita, 2008] http://prdos.hgc.jp/cgi-bin/meta/top.cgi	SVM model built from prediction results of other DR predictors	prediction results of seven independent DR predictors	PDB	PDB
PreLink [Coeytaux and Poupon, 2005] http://genomics.eu.org/spip/PreLink	rule based, ratio of multinomial probability	probability ratio (calculated from amino acid composition) and the hydrophobic clusters distance	PDB	PDB
RONN [Yang et al., 2005] http://www.strubi.ox.ac.uk/RONN	bio-basis function Neural Network	score of multiple sequence alignment	MSD	MSD
VSL2 [Peng et al., 2006] http://www.ist.temple.edu/disprot/predictorVSL2.php	two predictors specific for short and long DRs (>30 residues) prediction; a meta predictor then integrate specialised predictors	PSSM, AAC, predicted secondary structure	PDB	PDB, DisProt

and flexible (disordered) regions in protein sequences.

PreLink also calculates the amino acid distribution in structured and unstructured regions first. The multinomial probability of the unknown fragment can then be calculated, which determines whether it belongs to a structured or unstructured region.

2.7.2 Features

Some features extracted from primary sequences are widely used for various types of disorder prediction. These include AAC related features, evolutionary information, prediction results

Table 2.5: Global disorder predictors

Predictor	Method	Features	Databases	
			Order	Disorder
FoldIndex [Prilusky et al., 2005] `http://bip.weizmann.ac.il/fldbin/findex`	equation based on hydrophobicity and charge	net charge and average hydrophobicity score	SWISS-PROT	literature
FoldUnfold [Galzitskaya et al., 2006a] `http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi`	expected average number of contacts per residue	optimal set of artificial parameters of contacts for 20 residues	PDB, SCOP	SWISS-PROT
POODLE-W [Shimizu et al., 2007b] `http://mbs.cbrc.jp/poodle/poodle-w.html`	KNN graph and semi-supervised learning	AAC, physicochemical properties	PDB	DisProt

of secondary structure and solvent accessibility.

AAC Related Features

The first set of features introduced in DR prediction is AAC related features. It is known that many DRs have low complexity and biased amino acid composition, and AAC is a common feature in the prediction of disorder [Dunker et al., 2001; Vucetic et al., 2003]. Subsequent studies indicate that sequence regions with low complexity nearly always correspond to non-folding segments, whereas DRs do not always possess low sequence complexity [Romero et al., 2001; Radivojac et al., 2007]. Some amino acids are substantially more abundant in DRs than in the common folded proteins, whereas others are rare or absent. In general, the bias favours hydrophilic residues and discriminates against hydrophobic ones. Many physicochemical properties such as the hydrophobicity index can be used to group the AAC into reduced-AAC. Various predictors [Romero et al., 1997b; Li et al., 1999; Vucetic et al., 2003; Radivojac et al., 2003; Weathers et al., 2004; Peng et al., 2006; Hirose et al., 2007] of all four types of disorder prediction have adopted AAC or reduced-AAC based on physicochemical properties as training features. Predictors such as PreLink [Coeytaux and Poupon, 2005] predict unstructured regions with three rules calculated from AAC. The commonly used Shannon entropy is also derived from AAC. AAC related features can be calculated easily from sequences and are particularly suitable for large scale prediction of disorder.

Evolutionary Information

PSSM is another feature commonly used by DR predictors. It has been reported that DRs appear to be conserved through evolution [Iakoucheva et al., 2001; Lise and Jones, 2004]. However Peng et al. [2005] reported that DRs often evolve faster. More recently, Chen et al. [2006b] reported that sequence conservation in Conserved Disorder Prediction (CDP) regions varies. On average, sequence conservation in CDP regions is slightly lower than in regions of conserved order. However, one observation holds that many DR predictors achieve higher prediction accuracy after considering PSSM (or related evolutionary) information. PSSM has been adopted by most short and comprehensive DR predictors [Ward et al., 2004; Peng et al., 2005; Shimizu et al., 2005; Cheng et al., 2005b; Vullo et al., 2006; Ishida et al., 2006; Peng et al., 2006; Ishida and Kinoshita, 2007; Schlessinger et al., 2007b] as part of the training features.

The application method of PSSM in DR prediction is one thing that needs to be considered. Instead of applying 20 numbers for each target residue as training features, the local PSSM information of the target residue should also be taken into consideration. VSL2 uses the simplest approach by averaging the PSSM in the input window before feeding it into the machine learning model. Conversely, DISOPRED2 takes all the PSSM information in the input windows to the learning model. DISpro has applied the 1D-Recursive NN as the training model, so outputs of order/disorder status depend on the PSSM information of entire sequences.

Prediction Results of Secondary Structure and Solvent Accessibility

It is known that secondary structure and solvent accessibility properties are related to the disorder/order status of residues. Prediction results of secondary structure and solvent accessibility have been used as training features by many successful long, short and comprehensive DR predictors including NORSp [Liu et al., 2002; Liu and Rost, 2003], POODLE-L [Hirose et al., 2007], Ucon [Schlessinger et al., 2007b], DISOPRED2 [Ward et al., 2004], DISpro [Cheng et al., 2005b], Spritz [Vullo et al., 2006] and VSL2 [Peng et al., 2006].

2.7.3 Training Datasets

Training datasets are one of the most significant differences among the four types of DR predictors. DisProt contains many long DRs and these long DRs are commonly used by global and long disorder predictors. In contrast, DRs derived from PDB are usually short and mainly used by short and comprehensive DR predictors. Structured training datasets are usually built from PDB with different selection criteria.

DisProt

DisProt [Obradovic et al., 2003; Vucetic et al., 2005] is a published database of protein disorder. As a curated database, DisProt provides structure and function information about proteins that lack a fixed three-dimensional structure under putatively native conditions, either in their entirety or in part. All available structural/functional information in DisProt is obtained through an exhaustive search of the relevant literature and biological databases [Dunker et al., 2002a]. For each disordered protein, the database includes: the name of the protein, various aliases, access codes, amino acid sequence, location of the DRs, and methods used for structural characterisation. Since most DRs in DisProt are long, DisProt is a good option for training predictors specifically for long DR or global disorder prediction. Global disorder and long DR predictors, Spritz [Vullo et al., 2006], POODLE-W [Shimizu et al., 2007b], POODLE-L [Hirose et al., 2007] and Ucon [Schlessinger et al., 2007b] have built their training dataset from DisProt. Only a limited number of short DR predictors such as POODLE-S [Shimizu et al., 2005] have included DRs from DisProt into their training datasets.

DR annotation is still not complete in DisProt and many unstructured regions are not covered by DisProt [Oldfield et al., 2005a]. In the CASP7 competition, none of the top six predictors is solely trained by DRs from DisProt.

PDB

In contrast to DisProt, which obtains disordered information from the literature, CASP competitions have a definition for targets of DRs (refer to Section 2.2). According to this definition, many predictors build their training datasets by searching PDB. In PDB most of the structures come from proteins that have been successfully crystallised [Ward et al., 2004]. DRs are annotated by comparing SEQRES, ATOM tags or by referencing REMARK 465 tags. The SEQRES records contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule under study. It is the primary sequence of backbone residues. The ATOM records keep the atomic coordinates for standard residues and atomic coordinates that describe the position of an atom in the asymmetric unit of the crystal structure [Glusker et al., 1994]. The assumption is that missing atomic coordinates (ATOM record) are related to the flexibility of disorder. The REMARK 465 tags directly list the residues that are present in the SEQRES records but not in the coordinate sections.

Many short or comprehensive DR predictors including DISOPRED2 [Ward et al., 2004], DISpro [Cheng et al., 2005b], VSL2 [Peng et al., 2006], DisEMBL [Linding et al., 2003a], metaPrDOS [Ishida and Kinoshita, 2008], PrDOS [Ishida and Kinoshita, 2007] and Fais [Ishida et al., 2006; Ishida and Kinoshita, 2007] build their training datasets from PDB. One advantage of this approach is that each sequence contains both ordered and disordered residues and it is not necessary to build another structured dataset. The training dataset is similar to real query sequences that are partially disordered. Predictors built from this kind of dataset can be more sensitive and suitable for short DR prediction. In addition to short and comprehensive DR predictors, some long DR predictors such as XL1 [Romero et al., 1997b], VL1 [Romero et al., 1997a] and VL2 [Vucetic et al., 2003] have included DRs from PDB into their training datasets.

PDB is a structured database and it is not surprising that most DR predictors including FoldUnfold, VL1, VL-XT, VL3, VL3H, VL3E, POODLE-L, DISOPRED2, DISpro and VSL2 extract ordered segments from it to build ordered datasets.

Redundancy Reduction of Training Datasets

Redundancy reduction is a crucial issue in building training datasets of disorder [Linding et al., 2003a; Miller et al., 2008]. In the last 15 years, the number of proteins in PDB has increased over 20 fold and it is known that PDB contains considerable redundancy in sequence and structure[5]. Proteins sharing the same evolutionary origin often have structural similarity. The DisProt dataset also has a redundancy problem, as most DRs in this dataset come from literature reports and a redundancy reduction strategy is not applied. A high level of redundancy in training datasets can lead to bias in the training procedure. Similarly, performance assessments and accuracy comparison can be inaccurate due to redundancy. Many DR predictors [Linding et al., 2003a;b; Cheng et al., 2005b] have carefully removed redundant sequences before building prediction models.

Various approaches have been applied to reduce the redundancy in training datasets. CD-HIT [Li and Godzik, 2006] is a program for clustering large protein databases at a high

[5] `http://www.pdb.org/pdb/static.do?p=general_information/news_publications/newsletters/2001q3/red_red.html`

sequence identity threshold. The program removes redundant sequences and generates a database only of the representatives. PDB users can apply CD-HIT to remove redundant sequences at different cut-off values and keep representatives only. PDB-select [Boberg et al., 1992; Hobohm and Sander, 1994] is another algorithm using a representative list of protein chains to reduce redundancy. The SCOP [Murzin et al., 1995; Conte et al., 2002] database is a comprehensive ordering of all proteins of known structure according to their evolutionary and structural relationships. Protein domains in SCOP are grouped into species and are hierarchically classified into families, superfamilies, folds and classes. Therefore, datasets can be homology reduced based on SCOP [Linding et al., 2003a]. A dataset containing only one chain from each SCOP superfamily has a very low level of redundancy.

These redundancy reduction approaches introduced above are generic methods and can be applied to various biological databases.

Chapter 3

Data and Evaluation

In this chapter, in Section 3.1 we will describe the databases based on which training and test datasets of our predictors will be developed. In Section 3.2, features used in our DR prediction will be described. Finally in Section 3.3 we will illustrate metrics for measuring predictors.

3.1 Databases

From Sections 3.1.1 to 3.1.4, we introduce various databases used to train and test our disordered prediction models. Many DR predictors apply alignment based features [Bordoli et al., 2007]. This approach brings evolutionary information into the prediction process and has been proven to be very effective. Two commonly used alignment databases nr and Uniref100 are introduced in Section 3.1.5. Finally, the redundancy reduction process is discussed in Section 3.1.6.

3.1.1 DisProt

DisProt [Sickmeier et al., 2007] is a curated published database of protein disorder and was established by searching the relevant literature and biological databases. Many long DRs are included in this database and this database is distinguished in that molten globule-like proteins [Sickmeier et al., 2007] are included within the definition of disorder.

There are annotation overlaps in DisProt. For example, two DRs may be annotated as 10-20 and 5-25. In constructing our training databases, overlapping DRs are merged to produce longer DR annotations. DisProt updates regularly and later versions usually contain more Intrinsically Unstructured Proteins (IUPs), which are proteins containing long DRs. The

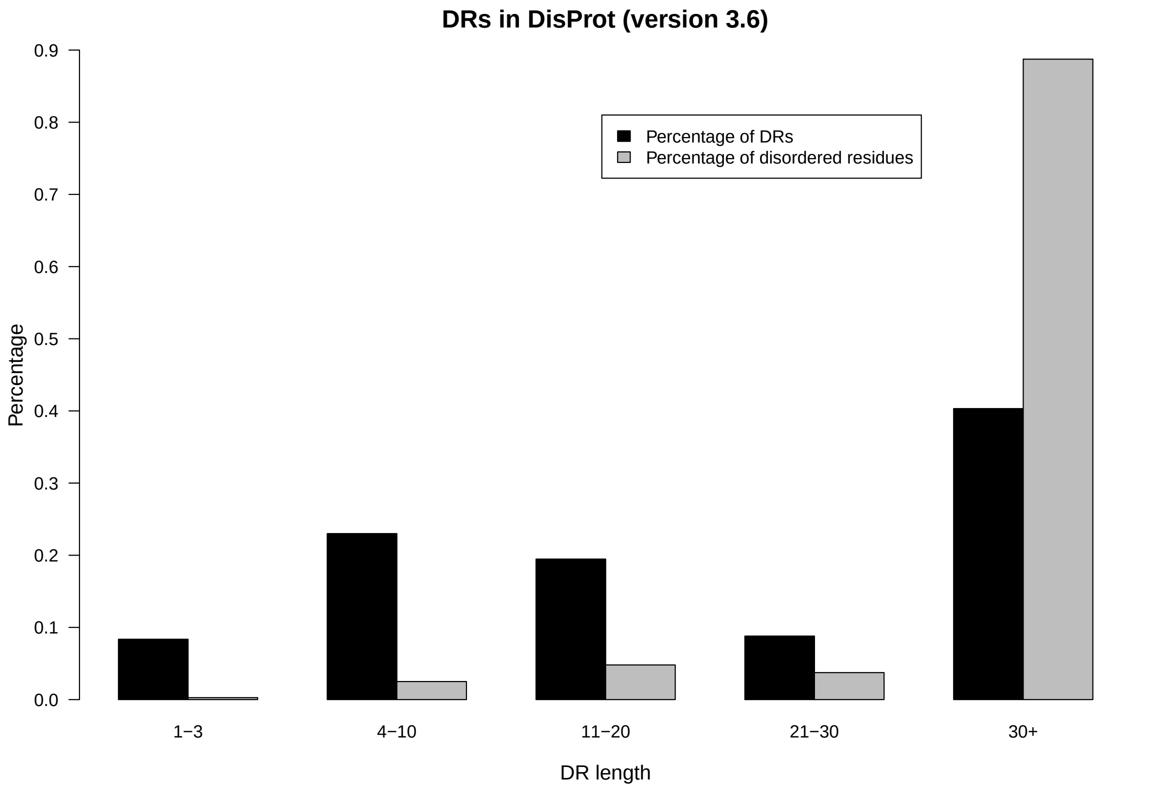


Figure 3.1: Distribution of length of DRs in DisProt (version 3.6)

number of long (> 30 residues) DRs from different versions of DisProt is shown in Table 3.1. Long DRs extracted from DisProt 2.1 where 2.1 is the version number are applied to train our global disorder predictor and this database is called *GDDB*. Long DRs extracted from DisProt 2.2 and 3.6 are applied to train our long DR predictors and these databases are called *LDDB1* and *LDDB2*. Unannotated areas in DisProt may not be considered as ordered given that the order/disorder status of these areas are unknown. These regions can contain DRs that have not yet been discovered.

Most DRs in DisProt are long with more than 30 residues. The percentage of length of DRs from DisProt 3.6 is illustrated in Figure 3.1. Residues from long DRs (>30 residues) dominate, while only a small percentage of residues (<15%) are from short DRs.

Table 3.1: Number of long DRs and residues from long DRs in DisProt

Version	Long DRs	Residues from long DRs
2.1	176	25,172
2.2	204	28,386
3.6	352	47,251

3.1.2 PDB

The PDB [Berman et al., 2000] database is mainly a structure database. Both structured regions and DRs can be discovered in PDB. As of 21st July 2009, there are over 59,000 structures stored in PDB and most structures were solved by X-Ray crystallography. For each protein in the database, PDB provides coordinate files listing atoms in the protein, and their three-dimensional location in space. If a structure was solved by X-Ray and some residues have missing coordinates, these residues are likely to be in disordered status. A DR dataset can be built by extracting these residues in PDB. Although this procedure does not strictly follow the definition of disorder, as missing coordinates may be attributed to various other reasons, it appears to be the most effective means of identifying DRs in the absence of further experiments. Usually DRs retrieved from PDB are short and contain less than 30 residues. Those residues in between DRs are usually considered ordered if the three-dimensional locations of them are known.

To formulate a curated dataset of DRs, the following process is applied. Firstly, 7.6% of the protein chains in PDB solved by X-Ray crystallography are obtained. Each chain contains at least one region of disorder of at least three residues in length. The residue status of order or disorder is based on the existence of an ATOM field (coordinate) for the C_α atom of a residue in the PDB file. A missing ATOM field indicates that a residue is disordered. Those proteins that are less than 30 amino acids in length, or have a resolution coarser than 2.5 Å are filtered out. Homologous protein chains are removed by UniqueProt [Mika and Rost, 2003].

The final dataset has 214,465 residues, and 13,831 (6.45%) residues from 1,775 DRs are annotated as disordered. 76.27% of disordered residues are from short DRs. There are 1,098 DRs containing at least four residues. In Figure 3.2 the distribution of DRs in this dataset is shown and this dataset is called *SDB*. It is clear that the majority of DRs are short and most disordered residues are from short DRs.

We have discovered that residues at the N and C termini in *SDB* are more likely to be disordered. The distribution of disordered residues is shown in Figure 3.3. In Figure 3.3

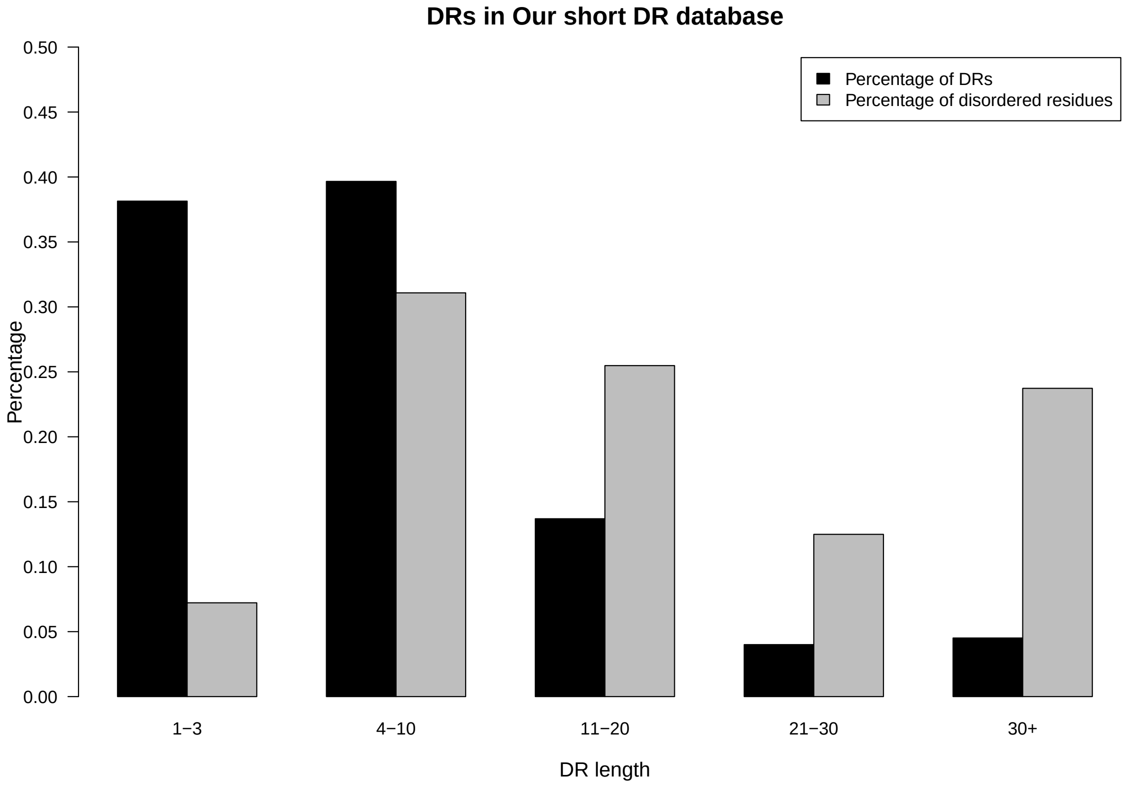


Figure 3.2: Distribution of length of DRs in our short DR database SDB

(a), the Y axis represents the percentage of disordered residues in N, C or internal areas. Generally, residues at smaller N/C terminal areas are more likely to be disordered. With ten residues at N and C termini, 38.3% of N terminal residues are disordered and 39.9% of C terminal residues are disordered. Only 4.1% of residues in internal areas are disordered. Although the internal area has a much lower content of disorder, it still includes many disordered residues. In Figure 3.3 (b), the Y axis represents the percentage of disordered residues among all disordered residues. With ten residues at the N and C termini, it shows that the internal area contains 59.4% of total disordered residues while the N and C termini contain 19.9% and 20.7% of total disordered residues.

3.1.3 PDB-select-25

PDB-select-25 [Hobohm and Sander, 1994] is a subset of structures obtained from PDB [Berman et al., 2000] that shows less than 25% amino acid sequence homology. As PDB-select-25 con-

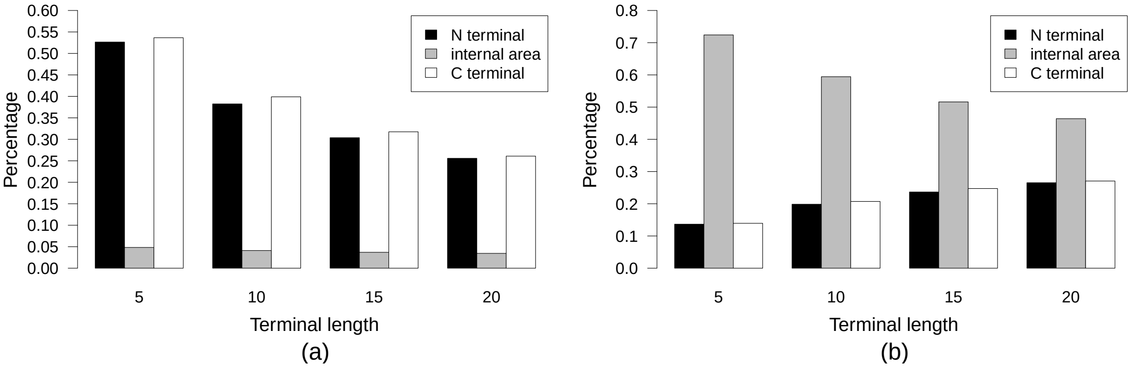


Figure 3.3: Distribution of disordered residues over N terminal, C terminal and internal areas in our short disorder training dataset SDB

tains structured proteins at a low level of homology, reliable ordered regions can be extracted from it for training DR predictors. There are 2,485 protein sequences in PDB-select-25 (Oct.2004 version). They include 366 high resolution crystal structures (<2Å) that are free from missing backbone or side chain coordinates, free from non-standard amino acid residues and with a sequence length of more than 80 residues. These high resolution crystal structures (ordered regions) include 80,324 residues. We apply this ordered database to train long DR predictors discussed in Chapters 4 and 5 as well as a global disorder predictor discussed in Chapter 7. We label this database *ODB*.

3.1.4 CASP7

The CASP competition held every two years provides a platform of comparison among different DR predictors as sequences are released for competition before structures are publicly available.

At the time when this thesis is written, the CASP8 competition has recently completed, but DRs of target sequences are not yet available. We therefore use the CASP7 sequences for evaluating our DR predictors. The CASP7 (`http://predictioncenter.org/casp7/`) dataset includes 170 DRs and 1,189 disordered residues. There are 106 DRs with at least four residues. In Figure 3.4 the percentage of length of DRs in the CASP7 targets is illustrated.

Figures 3.2 and 3.4 have a similar trend and there is a greater number of short DRs (<30 residues) than long DRs. Most residues come from short DRs. This trend is substantially different from the percentage of length of DRs in DisProt shown in Figure 3.1.

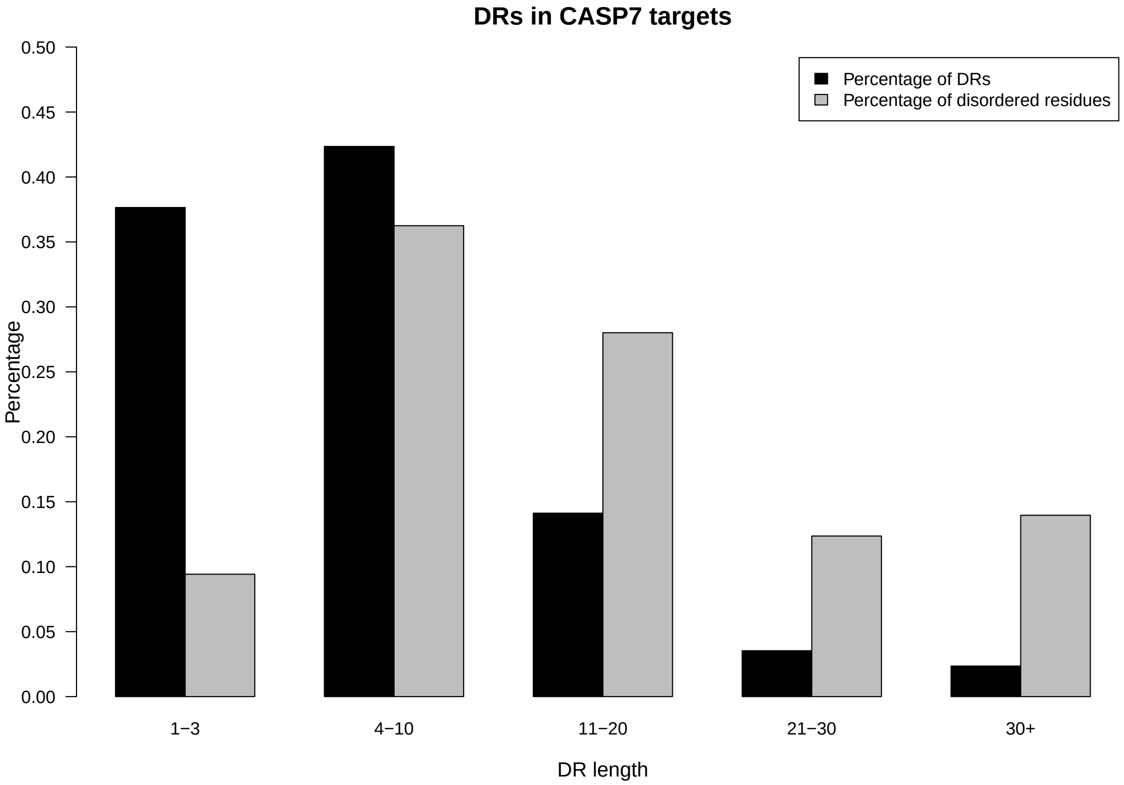


Figure 3.4: Distribution of length of DRs in the CASP7 targets

Targets of CASP7 have shown that N and C termini are more likely to be disordered than internal areas of protein sequences. The internal area is thus more difficult to predict [Bordoli et al., 2006]. In Figure 3.5 the distribution of disordered residues over N, C termini and internal areas in the CASP7 targets is shown. In Figure 3.5 (a), the Y axis represents the percentage of disordered residues in N, C or internal areas. In Figure 3.5 (b), the Y axis represents the percentage of disordered residues among all disordered residues. Both (a) and (b) in Figure 3.5 are similar to Figure 3.3. Residues at short N and C terminal areas are more likely to be disordered; however, the internal area still contains the most disordered residues due to the large number of residues. When N/C terminal areas increase the percentage difference between N/C and internal areas becomes smaller.

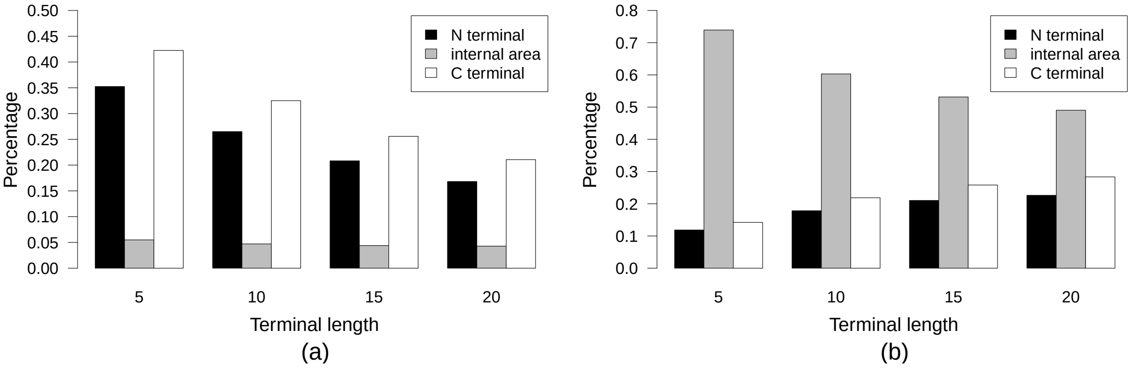


Figure 3.5: Distribution of disordered residues over N terminal, C terminal and internal areas in the CASP7 targets

3.1.5 UniRef100 and nr

The UniRef (UniProt Reference Clusters) databases cluster sequences from the UniProt (Universal Protein Resource) knowledge base and selected UniParc (UniProt Archive) records to form the complete coverage of sequence space at several resolutions while hiding redundant sequences. UniRef100 is a non-redundant protein sequence set merging identical sequences and sub-fragments with 11 or more residues (from any organism) into a single UniRef entry.

The nr (non-redundant) database is a widely used database from the NCBI (the National Centre for Biotechnology Information) for alignment. As the default database used by BLAST, nr contains proteins from Swiss-Prot, PIR, PRF, RefSeq, PDB as well as translations of coding sequences in GenBank.

Most sequences in the database UniRef100 are the same as those in nr since they derive their sequences from a similar set of sequence repositories. Comparing the releases of UniRef100 (Release 9.0, October 31, 2006) and nr (November 1, 2006), UniRef100 has over 180,000 unique sequences while nr only has 9,440 unique sequences. Generally UniRef100 contains more sequences than nr and both databases have unique sequences.

Different from the databases introduced earlier, UniRef100 and nr are not directly used as training/test databases. They are alignment databases that introduce evolution information into DR prediction. DRs have been shown to have a different speed of evolution from ordered regions [Iakoucheva et al., 2001; Lise and Jones, 2004; Peng et al., 2005; Chen et al., 2006b] and evolutionary information has been used successfully in DR prediction [Jones and Ward, 2003; Peng et al., 2006]. As a large number of sequences from different databases are included in UniRef100 and nr, the alignment result of a query sequence is likely to be a group of sequences from the same family. Due to different sequences in nr and UniRef100, results of alignment can be slightly different. We have downloaded the UniRef100 database (release 13.0)[1] and sequence library nr[2] on 18 Apr, 2008 as our alignment databases for training our disorder prediction models.

3.1.6 Redundancy Reduction

Redundancy reduction has been introduced in Chapter 2, and we know a high level of redundancy in training datasets can lead to bias in both training and performance evaluation. To alleviate redundancy, we applied various approaches to build datasets of disorder. All tables and statistics regarding our datasets in Section 3.1 are the findings after redundancy reduction.

For long DRs extracted from DisProt, homologues of DRs in our DisProt datasets are removed by the CD-HIT [Li and Godzik, 2006] algorithm. A threshold of 0.9 for sequence identity is adopted to assure that sequence similarity in each DR dataset is less than 90%.

To decrease the redundancy of the training dataset extracted from PDB, we filtered out homologous protein chains using UniqueProt [Mika and Rost, 2003] with a threshold HSSP value of ten. The HSSP value measures similarity between two sequences taking into account of sequence identity and length. A HSSP value of ten corresponds roughly to a 30% sequence identity for a global alignment of a 250 amino acid length.

One of the major reasons that PDB-select-25 [Hobohm and Sander, 1994] is chosen by us to extract ordered sequences is due to its low level of redundancy. As a subset of PDB [Berman et al., 2000], PDB-select-25 shows less than 25% amino acid sequence homology. The cut-off value here is very stringent and it is unnecessary to have further redundancy reduction processes.

3.2 Features Used in our Models

In this section, the features used to build our disorder predictors are introduced, which include amino acid composition, position specific score matrix, secondary structure and solvent accessibility. These features have proven to be effective in disorder prediction and are adopted in building our disorder predictors.

[1] ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref100/

[2] ftp://ftp.ncbi.nlm.nih.gov/blast/db/

Table 3.2: Reduced amino acids

Reduced AAC groups	Frequency	Residues
Positively charged (P)	F_P	Lys, Arg
Negatively charged (N)	F_N	Asp, Glu
Hydrophobic (H)	F_H	Trp, Phe, Tyr, Leu, Ile, Val, Met
Others (E)	F_E	Ala, Cys, Gly, His, Asn, Pro, Gln, Ser, Thr

3.2.1 Amino Acid Composition Based Features

The AAC is defined as the percentage of amino acids in a sequence/segment. Recall from Table 2.1 that the 20 amino acids can be grouped into different reduced-AAC groups according to their physicochemical properties. It is known that AACs and reduced-AACs are related to the order/disorder status of residues. Generally, disordered regions have a high composition of residues A, R, G, Q, S, P, E and K and low composition of residues W, C, F, I, Y, V, L and N [Dunker et al., 2001; Vucetic et al., 2003]. The high content of hydrophobic residues is a good indicator of the lack of DRs in sequences [Weathers et al., 2004; Shimizu et al., 2007a].

In our research, we group AACs into four categories based on their hydrophobicity and polarity (positively charged, negatively charged, hydrophobic, others) shown in Table 3.2. This reduced-AAC is applied in our long DR predictors in Chapters 4 and 5 and our global disorder predictor of Chapter 7.

3.2.2 Position Specific Score Matrix

The PSSM is a matrix of size 20 by the sequence length. The PSSM is generated by calculating position-specific scores for each position in the alignment and it captures the conservation pattern in the alignment.

PSI-BLAST [Altschul et al., 1990; 1997] is one of the most commonly used sequence profile search methods to generate a PSSM. Given a query sequence to be aligned, in an initial PSI-BLAST search, a PSSM (profile) is constructed from a multiple alignment of the highest scoring pairs. The score of a pairwise sequence alignment is defined by a “substitution matrix”, which assigns a score for aligning any possible pair of residues. Scores in the substitution matrix come from the observed frequency with which that substitution is known to occur among related proteins. Frequently observed substitutions receive positive scores and seldom observed substitutions are given negative scores. The profile is used in place of the original “substitution matrix” to perform a second PSI-BLAST search and the results of

Table 3.3: Part of the PSSM of protein sequence 1JC9

Pos	Res	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	Q	-3	-1	-2	-1	-5	7	4	-4	-2	-5	-4	-1	-3	-5	-3	-2	-3	-4	-4	-4
2	N	-3	0	3	-3	-4	5	-1	-2	2	3	-3	1	-2	-4	-4	-2	-1	-5	-3	-1
3	K	-2	0	-3	-4	-4	3	-1	-2	0	-1	2	3	0	0	-4	-2	0	-4	-1	-1
4	E	-2	4	0	2	-5	2	3	-4	2	-5	-5	0	-4	-5	-4	1	-1	-5	-1	-5
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
266	I	-1	-4	-3	-4	-3	-3	-3	-4	-4	1	2	-3	-2	-3	6	1	1	-5	-4	0
267	I	0	-3	-3	-3	-3	-3	0	-4	-4	4	1	-3	1	-2	2	0	-2	-4	-3	2
268	G	2	-3	-1	-2	-3	-2	-3	6	-3	-4	-4	-2	-3	-4	-3	-1	0	-4	-4	-3
269	N	-2	-1	5	3	-4	0	0	-2	7	-4	-4	-1	-3	-3	-3	-1	-1	-4	-1	-4

each "iteration" are used to refine the profile. This iterative search strategy continues until no new sequences are detected above a defined threshold. This threshold value has to be properly set to make a trade-off between more family members and the risk of picking up unrelated sequences, which weakens family-specific bias in the profile [Mount, 2001; Bhagwat and Aravind, 2008]. The iterative profile generation process makes PSI-BLAST far more capable of detecting distant sequence similarities than a single query sequence.

PSSM has been used in secondary structure prediction [Bryson et al., 2005]. Studies reveal that DRs are evolutionarily conserved and possess biological functions [Ward et al., 2004]. Therefore, evolutionary information in PSSM improves the accuracy of DR prediction [Peng et al., 2006].

In our research, we have applied the PSI-BLAST program blast-2.2.16 to the nr and UniRef100 databases. As a trade-off between time and sensitivity, we adopt three iterative searches. To avoid alignment failure and to include more sequences during the alignment procedure, we set an e value which sets the upper limit of the expected cutoff value to 0.001 when generating a PSSM from nr. As an example, part of a PSSM of sequence 1JC9 (PDB code) whose secondary and tertiary structure were shown in Chapter 2, is illustrated in Table 3.3. The first row of Table 3.3 suggests that according to the results of a multiple sequence alignment, the first residue Q in sequence 1JC9 is more likely to be substituted by residues G or H, while it is unlikely to be substituted by residues F, L or Q. In addition to a 20-column PSSM being generated, information per position and relative weight to pseudo counts are also calculated.

To select sequences closely related to each other and increase the search stringency, the e value is changed to 0.0001 when the UniRef100 database is applied.

3.2.3 Secondary Structure and Solvent Accessibility

The most common protein secondary structure consists of alpha helices and beta sheets. The random coil is not a true secondary structure, but a class of conformations that indicate an absence of regular secondary structure. Long segments of alpha helices and beta sheets are generally stable and unlikely to be in the disordered state. Secondary structure information is therefore related to the distribution of DRs in the protein sequence. Many DR predictors have applied the prediction results of the secondary structure as a part of their features [Cheng et al., 2005b; Peng et al., 2006]. Secondary structure can either be retrieved from structure databases such as PDB or prediction results from secondary structure predictors. However, results of secondary structure prediction are generally used, as the secondary structure of a query sequence may be unknown.

All protein secondary structure predictors are based on the assumption that there is a correlation between the primary sequence and secondary structure [Mount, 2001]. Based on sequences with known secondary structures, relationships between sequences and secondary structures are examined. More specifically, how the type and locations of secondary structural elements are associated with amino acid sequences. Supervised machine learning approaches [Rost, 2001; Pollastri et al., 2002] are used to model the relationship and predict secondary structures in query sequences.

The solvent accessibility of a residue refers to the surface area exposed to the solvent surrounding the protein in its native form. However, the surface area is difficult to define, even when the structure is known. With a probe sphere, the accessible surface can be traced out from the centre of the sphere as it rolls over the protein. For sequences in the DSSP [Kabsch and Sander, 1983] or FSSP [Holm and Sander, 1996] databases, the solvent accessibility value of each residue can be extracted. This information has been used to build predictors [Cheng et al., 2005a; Sim et al., 2005; Chang et al., 2008] predicting the solvent accessibility of protein sequences.

The studies of solvent accessibility have shown that the process of protein folding is driven to maximal compactness by the solvent aversion of some residues [Chang et al., 2008]. Residues accessible to the solvent tend to be flexible and are likely to be disordered. In contrast, ordered regions are usually folded and solvent averse. Solvent accessibility, which plays an important role in tertiary structure prediction [Sim et al., 2005], has been used in DR prediction [Cheng et al., 2005b].

Table 3.4: Confusion matrix

		Actual value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

3.3 Performance Metrics

To estimate the prediction accuracy of a classifier, cross-validation is a reliable approach. By leaving out some training instances for testing, a classifier is developed on the remaining training instances and tested on the left out test instances. The error rate is estimated from the misclassified test instances. We adopt cross-validation results to evaluate the performance of our DR predictors.

The prediction of DRs can be assessed on a per-residue level or per-chain level. The Per-chain level is generally used to measure the accuracy of binary (global) disorder prediction while the per-residue level can be used to measure all kinds of disorder predictors. Assuming disordered and ordered residues belonging to the positive and negative classes, in Table 3.4, the confusion matrix which comprises *true positive* (TP, actual positive and predicted as positive), *false positive* (FP, actual negative but predicted as positive), *true negative* (TN, actual negative and predicted as negative) and *false negative* (FN, actual positive but predicted as negative) is used to evaluate the performance of disorder predictors.

Traditionally in machine learning, the performance of classification systems is measured with simple *overall accuracy*, $\frac{TP+TN}{TP+TN+FP+TN}$. However, overall accuracy can be misleading when one class dominates. Many measures have been proposed to measure the accuracy of both positive and negative classes.

The *true positive rate* defined as $\frac{TP}{TP+FN}$ is also called *recall* or *sensitivity*, this measurement represents how effective the predictor is at finding disordered residues, and the higher the recall then the better a predictor is. The *false positive rate* defined as $\frac{FP}{TN+FP}$, is calculated by $1 - specificity$ while *specificity* is $\frac{TN}{TN+FP}$. This false positive rate measures the mistakes made by the predictor in DR prediction, where the lower the false positive rate the better a predictor is. The *disorder precision* defined as $\frac{TP}{TP+FP}$ is the percentage of residues correctly predicted as disordered in relation to the number of residues predicted as disordered. More measurements calculated from the confusion matrix are shown below:

- AUC: the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the true positive rate against the false positive rate for the different parameters of a classifier. ROC curves have long been used in signal detection theory. They are also used extensively in medical and biological studies. ROC curves are especially useful for domains with skewed class distribution and unequal classification error costs, as each point of a ROC curve is defined by a pair of values for the true positive rate (sensitivity) and false positive rate (1-specificity). AUC has similar properties to the non-parametric Wilcoxon statistic in that a score of 50% represents a random classification and 100% represents a perfect classification.

- Balanced overall accuracy:

$$Bacc = \frac{sensitivity + specificity}{2} \tag{3.1}$$

- Sproduct:

$$Sproduct = sensitivity \times specificity \tag{3.2}$$

- Mathew's correlation functions (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{3.3}$$

- S_w:

$$S_w = \frac{w_{disorder} \times TP - w_{order} \times FP + w_{order} \times TN - w_{disorder} \times FN}{w_{disorder} \times (TP+FN) + w_{order} \times (TN+FP)} \tag{3.4}$$

where $w_{disorder}$ and w_{order} are the weights for disorder and order, respectively, that are inversely proportional to the number of residues in the disordered and ordered state. S_w is also called the *probability excess*.

Sproduct, S_w, AUC, and the ROC curve have been used in assessing the prediction of disordered residues in the CASP competition.

3.4 Comparing ROC Curves

Statistical comparison of the area under two ROC curves derived from the same dataset can be achieved by the z value. This quantity z is then referred to tables of the normal distribution for P-value, and values of z above a certain cutoff value (e.g., $z \geq 1.96$) are taken as evidence that the "true" ROC areas are different [Hanley and McNeil, 1983].

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \tag{3.5}$$

A_1 and SE_1 refer to the observed area (AUC) and estimated standard error of curve 1. The standard errors associated with these areas can be obtained from the variance of the Wilcoxon statistic [Hanley and McNeil, 1982]. r is the correlation coefficient between areas A_1 and A_2. To calculate r, two intermediate correlation coefficients corresponding to disordered and ordered classes are required. Intermediate coefficients are then converted into a correlation (r value) between A_1 and A_2 [Hanley and McNeil, 1983]. Each of the intermediate correlation coefficient can be calculated by the traditional Kendall tau(τ):

$$\tau = \frac{\sum_{i<j}(sgn(x_i - x_j)sgn(y_i - y_j))}{\sqrt{(T_0 - T_1)(T_0 - T_2)}} \tag{3.6}$$

where $T_0 = n(n-1)/2$, $T_1 = \sum_k t_k(t_k - 1)/2$, and $T_2 = \sum_l u_l(u_l - 1)/2$. The t_k is the number of tied x values in the kth group of tied x values, u_l is the number of tied y values in the lth group of tied y values, n is the number of observations, and $sgn(z)$ is defined as:

$$sgn(z) = \begin{cases} 1 \text{ (if } z > 0) \\ 0 \text{ (if } z = 0) \\ -1 \text{ (if } z < 0) \end{cases} \tag{3.7}$$

This statistical comparison of the area under the two ROC curves is actually an adjusted T test. In our experiments we have used multiple (> 50) different false positive rates to generate paired TP, TN, FP and FN values from two ROC curves. This approach has a similar effect as the method of Miller et al. [2008] where resampling is used to get samples for the T test.

Chapter 4

Prediction of Long Disordered Regions Using Decision Tree

It is well recognised that a protein sequence determines the structure, and presumably the sequence should determine lack of structure as well. Long DRs in sequences cause difficulties in protein structure determination by both X-Ray crystallography and NMR spectroscopy. Efficient prediction of these long DRs in IUPs by computational methods can provide valuable information in high-throughput protein structure characterisation and reveal useful information regarding protein functions [Oldfield et al., 2005b].

Predicting IUPs can be cast as the binary classification problem in machine learning and data mining as discussed in Chapter 2. Many predictors have been developed to predict DRs in proteins. Among existing long DR predictors, machine learning models including NN and SVM are widely used. But such prediction models are like black box learning systems which are often complex to interpret and understand from the biological perspective. In this chapter we present a system for predicting long DRs using reduced-AAC based on the decision tree learning model. The decision tree is one of the most widely used and practical methods for classification in machine learning [Mitchell, 1997]. Our predictor is trained on disordered proteins from DisProt [Vucetic et al., 2005] and ordered proteins from PDB-Select-25 [Hobohm and Sander, 1994] with well-defined crystal structures.

With the knowledge that DRs are enriched in charged and polar amino acids [Uversky et al., 2000; 2005], we have constructed decision trees based on reduced-AAC of four to five groups. Compact decision trees are easy to interpret and understand. Learnt trees can also be represented as sets of if-then rules to improve human readability. Based on these rules, a predictor has been developed that can achieve a recall of 80% at 13% false positive rate for predicting DRs in the ten-fold cross-validation test. At the same false positive rate, this recall is 6.7% higher than DISpro of Cheng et al. [2005b]. DISpro was reported to outperform DISOPRED2, VLXT, VL2 and VL3 on a set of proteins from the CASP5 competition. Moreover, the accuracy of our approach for extremely long DRs (of at least 100 residues) is 20% higher than the predictor VSL1 [Obradovic et al., 2005]. The use of reduced-AAC not only significantly improves learning and prediction efficiency, but also gives more accurate predictions than can be obtained by using AAC directly.

In this chapter, we study how decision trees can be applied for long DR prediction. In Section 4.1, material and methods of this tree-based approach are described. To alleviate the bias in training datasets, three independent predictors are built to determine the final results of prediction. We compare the accuracy of our predictor under different parameters in Section 4.2. One major contribution of decision tree based long DR predictors is understandability. Rules generated at the learning stage are explained and analysed. Given that amino acids have various physicochemical properties, we investigate reduced-AAC under different grouping methods in Section 4.3. Finally in Sections 4.4 and 4.5 we discuss and summarise our work.

4.1 Material and Methods

In this section, training datasets are presented first. The prediction model is then introduced and this is followed by an explanation of a decision tree classification algorithm.

4.1.1 Training Data

The ordered database is *ODB* built from PDB-Select-25 (the Oct.2004 version) [Hobohm and Sander, 1994] as described in Section 3.1.3. The disordered database is *LDDB1* built from DisProt (version 2.2) [Obradovic et al., 2003] as described in Section 3.1.1. There are 204 DRs in *LDDB1* and 366 structured segments in *ODB*.

4.1.2 Prediction Model

It is reported that AAC can be used to predict both disordered and ordered regions [Weathers et al., 2004; Coeytaux and Poupon, 2005; Peng et al., 2005; Romero et al., 2001; Dunker et al., 2002b]. Based on the hydrophobicity and polarity of amino acids shown in Table 2.1, we group 20 amino acids into four groups, the hydrophobic (H), positively charged (P),

Table 4.1: Reduced amino acids

Reduced AAC groups	Frequency	Residues
Positively charged (P)	F_P	Lys, Arg
Negatively charged (N)	F_N	Asp, Glu
Hydrophobic (H)	F_H	Trp, Phe, Tyr, Leu, Ile, Val, Met
Others (E)	F_E	Ala, Cys, Gly, His, Asn, Pro, Gln, Ser, Thr

negatively charged (N) and others (E) shown in Table 3.2. For ease of discussion, Table 3.2 is reproduced as Table 4.1.

Since aromatic residues are usually important for stabilising protein structures, long DRs usually have a low content of aromatic residues [Romero et al., 1997b]. We thus also divide hydrophobic residues into aromatic (Trp, Phe, Tyr) and aliphatic (Leu, Ile, Val, Met) groups and construct a prediction model based on five groups of amino acids. In principle, any grouping of amino acids into n groups ($n \leq 20$) can be used in our model.

Three windowing procedures are used: left-side, right-side and central windows [Romero et al., 1997b]. A central window predicts the residue at the centre of the window and includes the same number of residues on both sides. When a central window slides from the N-terminus of a protein to its C-terminus, there are some residues within the half window at both termini that are not covered, which is where left-side and right-side windows are applied. Left-side and right-side windows are half the size of the central window used. Instead of predicting the central residue, side windows predict for the terminal residues. Generally, windows should be much smaller than the length of a whole protein sequence. In our experiments we have tested different window sizes from 33 to 93.

We now use the central window to explain the windowing procedure; a similar procedure applies to side windows. A sequence of residues in a window can also be considered as a fragment. With the four groups of amino acids in Table 4.1, the reduced-AAC in a window is represented by four numbers. Generally with n groups of reduced amino acids, the reduced-AAC in a window is represented by n elements. So when a window of w residues slides along a sequence i, the content of the sequence is represented by $n \times (L_i - w + 1)$ elements, where L_i is the length of sequence i. As a result the disordered training segments are represented by $\sum_{i=1}^{204} n \times (L_i - w + 1)$ elements, denoted as D-M, and the ordered training segments by $\sum_{i=1}^{366} n \times (L_i - w + 1)$ elements, denoted as O-M.

Given the four groups of reduced-AAC in Table 4.1, the sum of F_H, F_N, F_P and F_E is one, so only three of them are independent. Therefore the ordered and disordered fragments

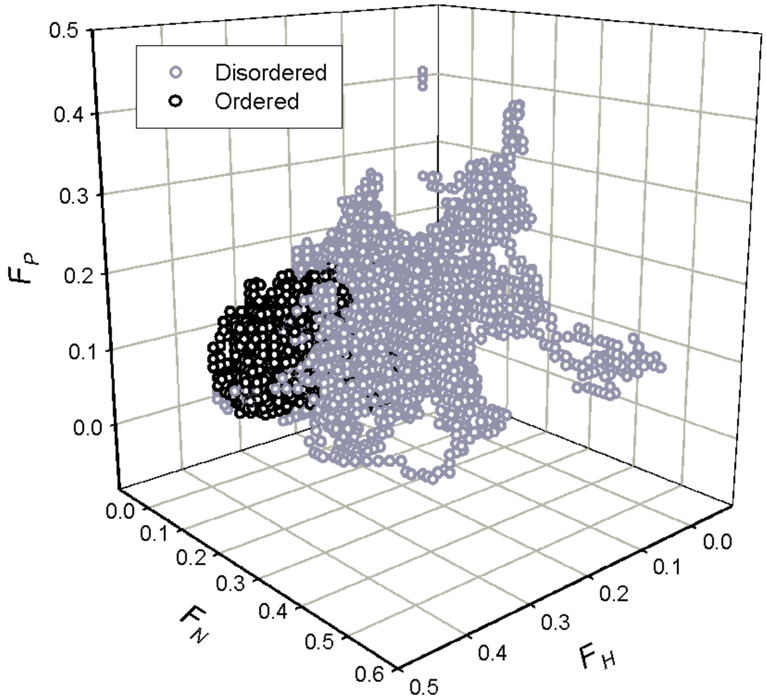


Figure 4.1: The representation of ordered and disordered fragments

can be represented in a three-dimensional space of F_H, F_N and F_P, which we define as a representative space (shown in Figure 4.1). It clearly shows that ordered and disordered fragments are located in separate subspaces, which indicates that they are predictable.

One difficulty of DR predictions is the shortage of disordered fragments for training. Imbalanced training sets make a predictor that focuses on achieving a higher overall accuracy regarding prediction of the majority class. In our study, the ordered training set is three times larger than the disordered training set and a predictor may focus on the accuracy of prediction of ordered residues to achieve high overall accuracy. Two approaches have been proposed in the literature to tackle this problem:

- Resampling balances the datasets by over-sampling the minority class, or under-sampling the majority class, or the combination of these two sampling methods. The main goal is to make the two training sets more balanced and so suit ordinary predictors.

- Cost modification gives a higher penalty for misclassification of minority class instances at the training stage. It balances the accuracy of prediction for two imbalanced datasets without modifying datasets.

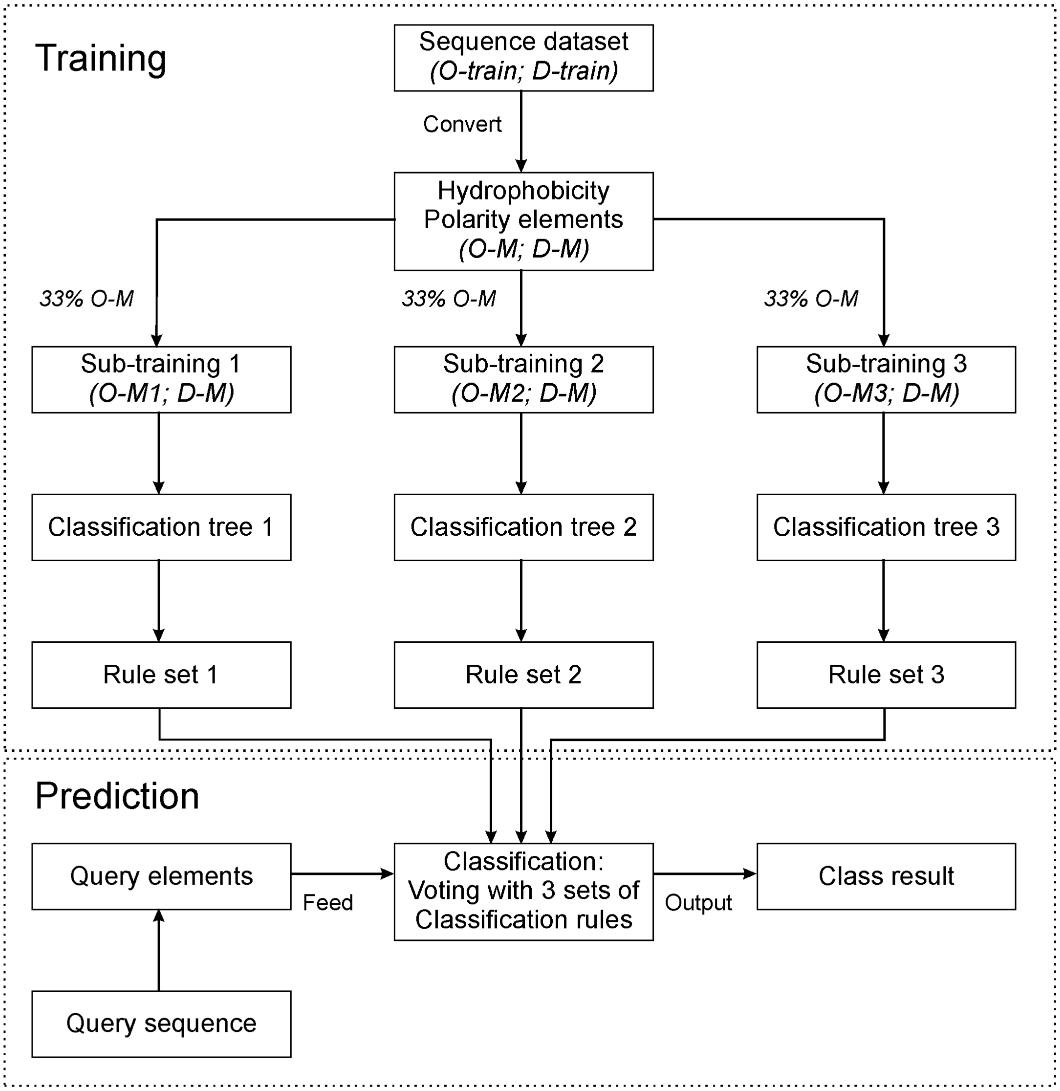


Figure 4.2: Development scheme of our long disorder predictor

To rectify the situation of over prediction favouring the ordered class, the ordered training set is divided into three sub-training sets $O\text{-}M1$, $O\text{-}M2$ and $O\text{-}M3$ from $O\text{-}M$. Each is trained with disordered set $D\text{-}M$ separately. The development scheme for our classification process is shown in Figure 4.2.

At the training stage, three distinct decision trees (details in Section 4.1.3) are constructed from the three training sets. Three sets of classification rules for disordered and ordered

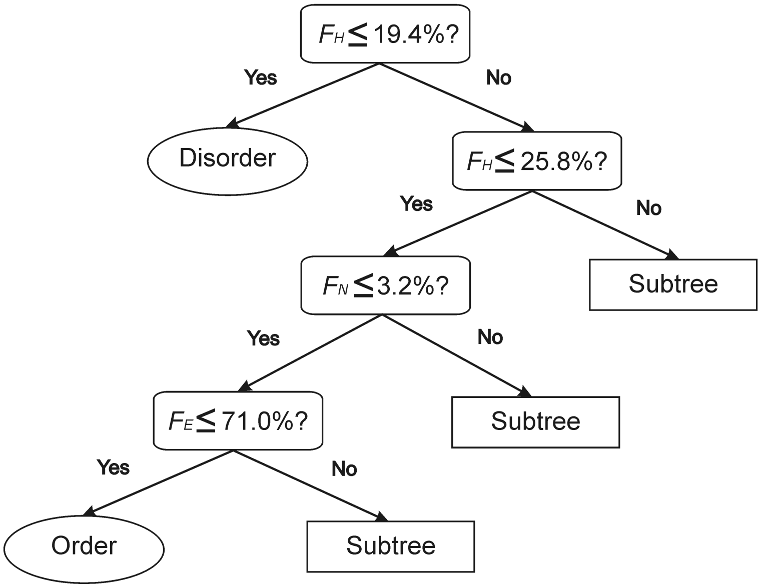


Figure 4.3: A sample decision tree

regions are then derived from the decision trees and each is assigned a confidence value. Each classification rule represents a hyperplane that best divides the ordered or disordered fragments in the representative space.

For a query sequence at the prediction stage, its corresponding reduced-AAC is first calculated for a given window along the sequence. The reduced AAC is then mapped to the representative space. If a reduced-AAC is mapped into the disordered subspace, it will be assigned a positive confidence value for the rule. If it is mapped into the ordered subspace, a negative weight that negates the confidence of the corresponding rule is assigned. Otherwise the reduced-AAC is assigned zero. By default, zero is the threshold for distinguishing disordered fragments while positive values indicate disorder and negative values indicate order. Considering that we have three groups of rules, the final status of disorder (or order) depends on the number of votes and the corresponding confidence from the rules.

4.1.3 The Classification Algorithm

In our representative space, each fragment is mapped to a point that is described by three groups of reduced-AAC. Fragments from the disordered training set *LDDB1* are tagged with the label D and all fragments from the ordered training set *ODB* with the label O.

Figure 4.3 provides a schematic sketch of a decision tree. With the test "$F_H \leq 19.4\%$?", the whole representative space is divided into two subspaces. In the space "$F_H \leq 19.4\%$" all fragments are labelled D and induction of the tree finishes. In the space "$F_H > 19.4\%$", the fragments are of mixed labels and they are further divided by the test "$F_H \leq 25.8\%$?". The decision tree of linear tests on all dimensions shown in Figure 4.3 is a close approximation of the boundary between ordered and disordered fragments shown in Figure 4.1. Constructing a decision tree for a set of reduced-AAC in the representative space follows the "divide and conquer" strategy. A space of a reduced-AAC with mixed labels is divided into subspaces by checking the composition of one reduced amino acid group. If a subspace consists of fragments of mixed labels, it is recursively divided by checking the composition of another reduced amino acid group. Such a process is repeated until when a space consists of reduced-AAC with a single label, either D or O, or a space becomes empty.

Using C4.5 [Quinlan, 1993], a popular decision tree learning system, a decision tree is constructed for the amino acid composition dataset. C4.5 employs information theory to decide the best test for dividing a subspace under consideration. Given a set T of D and O fragments the information content (entropy) for T is:

$$info(T) = -(\frac{|T_D|}{|T|} \times log_2(\frac{|T_D|}{|T|}) + \frac{|T_O|}{|T|} \times log_2(\frac{|T_O|}{|T|})) \tag{4.1}$$

After T has been partitioned into T_1 and T_2 following a test $F_i(i = H, N, P, E$ shown in Table 4.1), the information needed to classify T is:

$$info_{F_i}(T) = -(\frac{|T_1|}{|T|} \times info(T_1) + \frac{|T_2|}{|T|} \times info(T_2)) \tag{4.2}$$

The information gain $info(T) - info_{F_i}(T)$ measures the information that is gained by partitioning T with F_i. This gain is normalised by the information generated by the split of T into T_1 and T_2 to rectify the bias towards features with a large number of values.

$$split\ info(F_i) = -(\frac{|T_1|}{|T|} \times log_2(\frac{|T_1|}{|T|}) + \frac{|T_2|}{|T|} \times log_2(\frac{|T_2|}{|T|})) \tag{4.3}$$

Finally, the best test to divide a space is the one with the largest gain ratio

$$\frac{info(T) - info_{F_i}(T)}{split\ info(F_i)} \tag{4.4}$$

Every path from the root of an unpruned tree to a leaf gives one if-then rule. Each such rule is simplified by removing conditions that do not seem helpful for discriminating the nominated class from other classes, using a pessimistic estimate of the accuracy of the rule [Quinlan, 1993]. For each class in turn, all the simplified rules for that class are sifted to remove rules that do not contribute to the accuracy of the set of rules as a whole. The sets of rules for the classes are then ordered to minimise false positive errors and a default class is chosen. This process leads to a production rule predictor that is usually about as accurate as a pruned tree, but more understandable [Quinlan, 1993].

4.2 Results

A cross-validation test has been used in evaluating the accuracy of our predictor. With a two-class problem using positive (disorder) and negative (order) classes, the positive class is our main focus. The recall and precision for the disordered class as well as overall accuracy and ROC are used in this chapter to measure our predictor. These measures provide a clear description of the accuracy of our predictor.

Table 4.2 lists some representative disordered and ordered rules for the central window of 93 residues. These rules are ranked by confidence, which states the likelihood of order or disorder when a rule is satisfied. The first disordered rule indicates that, if the content of hydrophobic residues within a fragment is less than 19.4%, we are 100% sure that the central residue is disordered. As the window decreases to 73 and 53 residues, we still find this rule valid with a confidence higher than 99%, but the composition of hydrophobic residues drops to 16.4% and 15.1%, respectively.

It is clear from Table 4.2 that the combined composition of amino acids provides a confident prediction of disordered or ordered status. Disordered rule number four is different from previous findings. While previous findings show that a high content of charged residues is associated with disorder, this rule suggests that if the composition of hydrophobic (F_H) and charged (F_P and F_N) residues is less than 29%, then, with a probability of 99.8%, the central residue is disordered. This implies that, when there is a reasonable depletion of charged residues but a relative enrichment of hydrophobic residues, a fragment can also be disordered.

With the disordered rules shown in Table 4.2, the composition of hydrophobic residues (F_H) is always less than a threshold of 22.6%, except for rule number six. In contrast, six out of eight ordered rules have a composition of hydrophobic residues (F_H) $>$ 22.6%.

Table 4.2: Representative rules at a window of 93 amino acid residues

	Disordered				
	Rules	Confidence (%)		Rules	Confidence (%)
1.	$F_H \leq 19.4$	100.0	5.	$F_H \leq 22.6$, $F_N > 3.2$, $F_N \leq 6.4$	99.8
2.	$F_N > 11.8$, $F_P > 17.2$, $F_E > 39.8$	99.9	6.	$F_H \leq 33.3$, $F_P \leq 14.0$, $F_E \leq 33.3$	99.8
3.	$F_H \leq 20.4$, $F_P > 2.2$	99.9	7.	$F_N > 21.5$, $F_P > 11.8$	99.7
4.	$F_E > 71.0$	99.8	8.	$F_H \leq 22.6$, $F_N > 10.7$	99.7
	Ordered				
	Rules	Confidence (%)		Rules	Confidence (%)
1.	$F_H > 22.6$, $F_H \leq 25.8$, $F_N \leq 5.4$, $F_P \leq 11.8$, $F_E \leq 69.9$	98.9	5.	$F_H > 22.6$, $7.5 < F_N \leq 10.8$, $F_P \leq 11.8$	97.8
2.	$F_H > 23.7$, $F_N > 7.5$, $F_N \leq 14.0$, $7.5 < F_P \leq 8.6$, $F_E > 50.5$	98.6	6.	$F_H > 24.7$, $7.5 < F_N \leq 12.9$, $4.3 < F_P \leq 8.6$, $F_E \leq 58.1$	97.7
3.	$19.4 < F_H \leq 25.8$, $F_N \leq 3.2$, $F_E \leq 71.0$	98.2	7.	$F_H > 26.9$, $F_N \leq 10.8$, $F_P \leq 19.4$, $F_E \leq 43.0$	97.7
4.	$F_H > 24.7$, $8.6 < F_P \leq 11.8$, $F_E > 50.5$	97.9	8.	$7.5 < F_N \leq 16.1$, $F_P \leq 1.1$	97.6

Disordered rule number six states that the composition of charged amino acids $> 33.3\%$ and hydrophobic residues $\leq 33.3\%$ implies disorder. All these rules can be explained by the prevalence of hydrophilic and charged amino acids and the depletion of hydrophobic and aromatic amino acids in DRs [Uversky, 2002b; Dunker et al., 2001; Haynes and Iakoucheva, 2006].

Figure 4.4 gives the ten-fold cross-validation test results over four groups of reduced-AAC. With an increase in window size, the disordered recall, precision and overall accuracy improve steadily. The overall accuracy generated under window size 93 is significantly higher (P-value $< 10^{-6}$ in T test) than that generated by window size 53. For a window of 93 residues, the disordered precision is 89.0%, higher than the overall accuracy of 87.3%. This indicates that the predictor has slightly higher accuracy in predicting DRs than ordered regions, and the influence of imbalanced training datasets has been alleviated. To prove that three decision trees possess consistent prediction ability, we compared the prediction results before voting. They achieved overall accuracies of 84.5%, 85.3% and 87.0%. The average overall accuracy over these three predictors is 85.6%; voting improves the overall accuracy by 1.7%. This improvement is not statistically significant with a P-value = 0.35. However, by running ten-fold cross-validation 50 times, voting strategy consistently performed better than individual predictors.

The ROC curves for predictors that combine the left-side, right-side and central windows at sizes from 33 to 93 are plotted in Figure 4.5. All ROC curves in Figure 4.5 are based on ten-fold cross-validation tests and these curves are consistent with that of Figure 4.4 in which

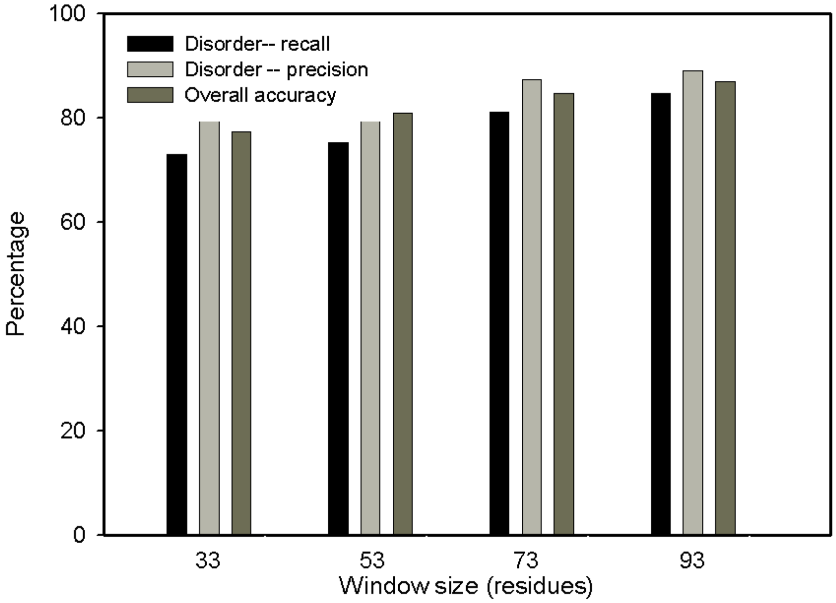


Figure 4.4: Results of ten-fold cross-validation test with four groups of reduced amino acid composition at different window sizes

accuracy of prediction from a larger window is higher. The window of 93 residues has the biggest area under the curve. When the false positive rate is 13%, disordered recall reaches 80% and the overall accuracy can reach 83.4%. In contrast, the small window of 33 residues for training is less accurate in predicting long DRs.

Figure 4.6 presents the result of the ten-fold cross-validation for AAC, and four and five groups of reduced-AAC, with the window of 93 residues. It is known that DRs are characterised by compositional bias toward aromatic and hydrophilic residues. Five groups of reduced-AAC are modified from four groups proposed in Table 4.1 by splitting the hydrophobic group further into aromatic (Phe, Trp and Tyr) and aliphatic (Ile, Leu, Met and Val) groups. Results in Figure 4.6 are obtained using the whole ordered training set *O-M* and disordered training set *D-M*. This experiment is specifically designed to evaluate the impact of imbalanced training datasets and compare the performance of different grouping strategies. All three predictors have an overall accuracy higher than 85%. In terms of disordered precision, grouping amino acids into four or five groups achieves much better accuracy than using AAC. Moreover, disordered recall and overall accuracy are also higher after grouping AAC. The disordered precision for four groups of reduced compositions is 86.3%, 13.3% higher than

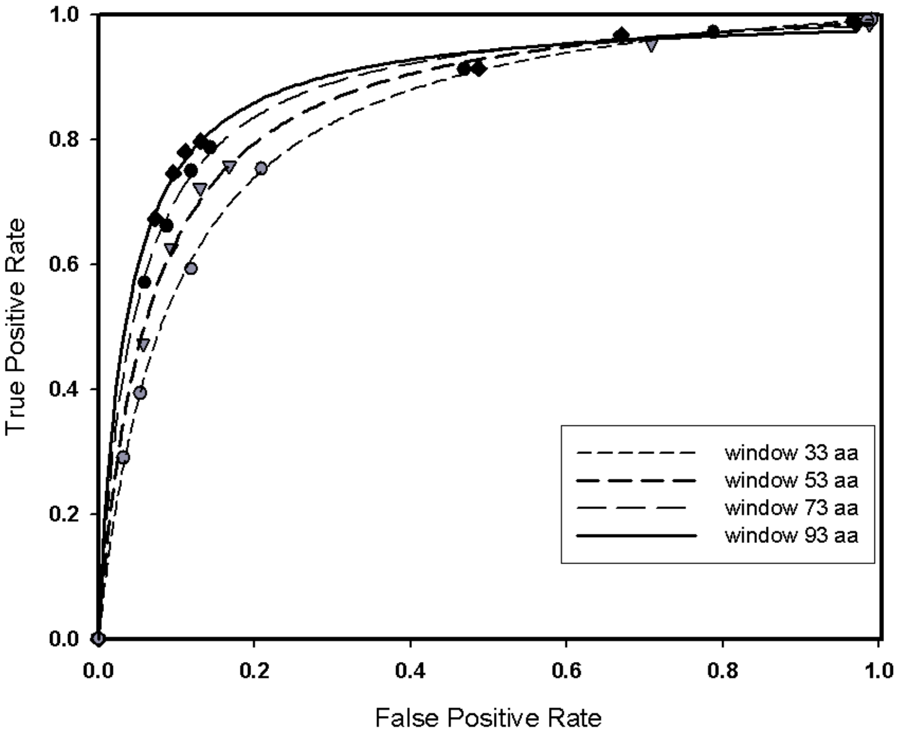


Figure 4.5: ROC curves from ten-fold cross-validation test under different window sizes

that for AAC. Moreover, four groups of reduced-AAC improve the training time by three fold. Both the four and five groups achieve very similar disordered recall and overall accuracy, indicating that separating aromatic and aliphatic residues in our hydrophobic group did not improve the accuracy of prediction. Figure 4.6 illustrates several important points when the predictor is trained by the lump sum of *O-M* and *D-M*. First, the precision for ordered regions is always higher than that for DRs. With four groups of reduced-AAC, the overall accuracy is 91.7% and the disordered precision is 86.3%. In other words, the ordered precision is much higher than 91.7%. In addition, while disordered precision is at a satisfactory level, the corresponding recall is low, implying that the predictor is too conservative to make a decision of disorder. Compared to Figure 4.4, balanced ordered and disordered training sets sacrifice the overall accuracy for higher disordered precision. The strategy of voting has lowered the overall accuracy by 4.4% but improved disordered recall and precision by 9.5% and 2.7% respectively.

The superior accuracy of reduced-AAC can be rationalised in biochemical terms, since mutation of amino acids with similar characteristics usually does not influence a protein's

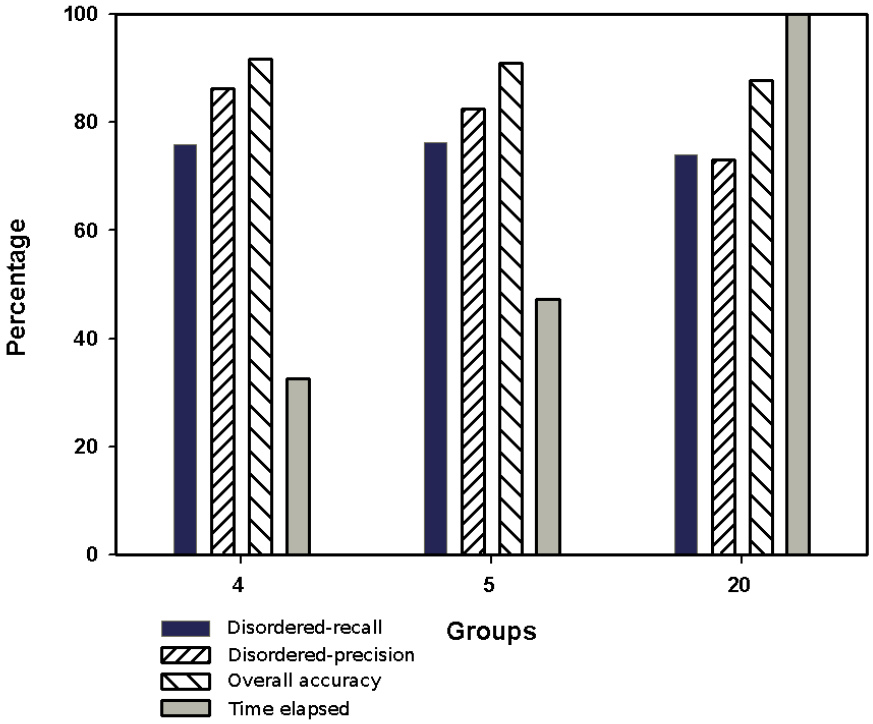


Figure 4.6: Comparison of the performance and time elapsed under different groups of the amino acid composition

three-dimensional structure. According to the Dayhoff mutation matrix, some amino acids in the same group show higher mutation rates compared with other amino acids [Dayhoff et al., 1978], such as residues Asp and Glu in the negatively charged group and Ala and Ser in other group.

4.3 Other Groups of Reduced Amino Acids

Apart from four groups of reduced amino acid proposed in Table 4.1 and five groups which split the hydrophobic group further into aromatic and aliphatic groups, we have examined two reduced amino acids as shown in Table 4.3 [Betts and Russell, 2003]. The corresponding ROC of cross-validation tests of these two groupings are shown in Figure 4.7. Reduced AAC including charged, and hydrophobic performs better than the other two groupings including polarity and size and is more related to the order or disorder status.

Table 4.3: Other groups of reduced amino acids

1. Polar, tiny, hydrophobic, others	
Reduced AAC groups	Residues
Polar	Asp, Glu, Lys, Asn, Gln, Arg
Tiny	Ala, Gly, Pro, Ser
Hydrophobic	Trp, Phe, Cys, Leu, Ile, Val, Met
Others	His, Thr, Tyr
2. Tiny, charged, hydrophobic, others	
Reduced AAC groups	Residues
Tiny	Ala, Gly, Pro, Ser
Charged	Lys, Arg, Asp, Glu
Hydrophobic	Trp, Phe, Cys, Leu, Ile, Val, Met
Others	His, Asn, Gln, Thr, Tyr

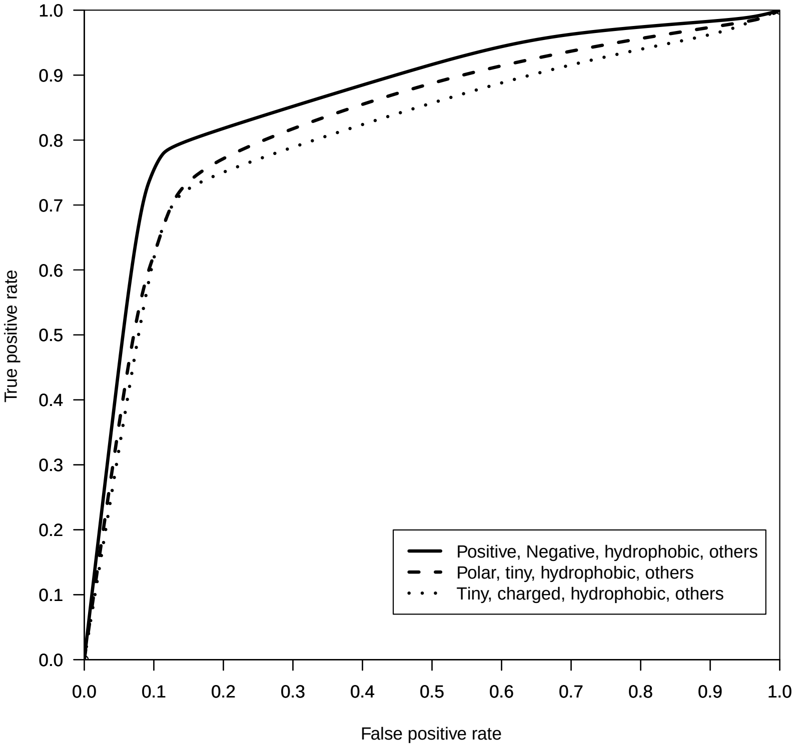


Figure 4.7: Comparing our four groups of reduced amino acids with other groups of reduced amino acids shown in Table 4.3

4.4 Discussions

In our prediction model, window size is a parameter affecting the accuracy of prediction. Our experiments have shown that larger windows produce a higher prediction accuracy. This may be partly explained by the fact that a large window can take long range information into account, whereas small windows only detect local protein information. Nevertheless, the accuracy cannot always be improved by increasing the window size. First, the larger the central window is, the more residues are not covered at the N and C termini. N and C side windows can be applied in these cases, but recall and precision rates drop significantly compared to central windows. Based on results of cross-validation, we illustrate this in Figure 4.8. Obviously, the accuracy of prediction and recall of internal regions are higher than those of N and C termini. Second, larger windows are insensitive to local structural information and are inapplicable to short sequences directly. We tested windows of up to 153 residues, and our results confirm this hypothesis.

Training datasets also contribute to high AUC values when a relatively larger window is applied. Given that the database contains only segments that are 100% ordered or disordered, a larger window makes the predictor similar to a global predictor. A predictor is less likely to misclassify the whole region into the opposite class. However, in “natural” distributions of order or disorder, ordered regions often coexist with DRs in the same sequence. In this case, the prediction is more challenging, as features derived from the window containing both ordered and disordered residues will not show “ideal” order or disorder properties that predictors have learnt. Therefore, predictors built from extremely large windows may no longer perform well as they focus on long range information, which overlooks the local information.

Some predictors reported in the literature have achieved very high precision in predicting medium to long DRs containing between 30 and 100 residues [Obradovic et al., 2005; Radivojac et al., 2003], but not all of them perform consistently well in predicting DRs containing more than 100 residues [Obradovic et al., 2005]. Our predictor is more accurate on these extremely long DRs. The optimum window size for our method is 93. We have also measured the performance of our predictor in predicting short DRs (< 30 residues). We apply rules generated for the default window of 93 residues to predict short query sequences by using small windows for prediction. For a prediction window of 23 residues, our system can achieve a disordered recall of only 60.6% and an overall accuracy of 78.6% in the cross-validation test. With increasing prediction window sizes, the disordered recall decreases and

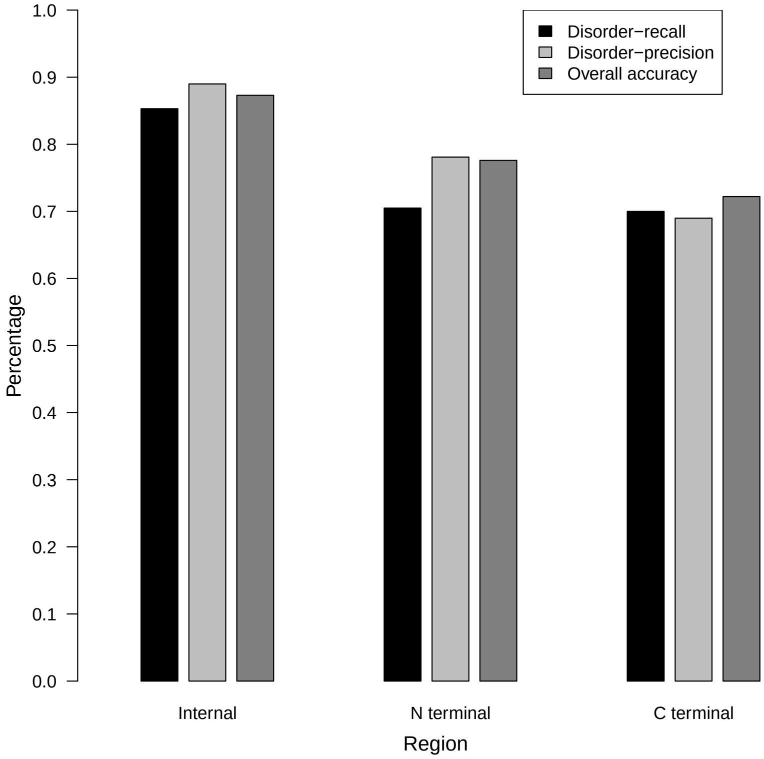


Figure 4.8: Comparison of accuracy of prediction between sequence internal and terminal regions

precision increases. We believe more sequence information needs to be considered in order to accurately predict short DRs.

Imbalanced training data tends to introduce bias into prediction models. Our approach which produces balanced training datasets and then applies the voting strategy to make the final prediction is similar in nature to the random forest model [Breiman, 2001b]. This approach helps relieve the problem of imbalanced training datasets and improves the robustness of our predictor.

The decision tree model can suffer from overfitting, which happens when “a tree grows so deep that it captures aberrations in data that harm the predictive power” [Quinlan, 1993]. Current techniques either stop growing the tree early before it overfits data or grow the full tree and then trim overfitting branches. In our prediction model, using reduced-AAC in fact reduces the number of parameters of the decision tree. Our experiments show that this not only reduces the number of rules from 1,100 to 150, but also improves the prediction accuracy. Reducing the number of input parameters can be an approach complementing current techniques in order to avoid overfitting.

The rules that describe the disordered state are simpler than those describing the ordered state. This is a result of the biased composition and lower sequence complexity of the sequences in the *LDDB1* dataset. Conversely, groups of reduced composition in the sequences of the *ODB* dataset are more uniform and sequence complexity is much higher. Ordered rules are generally more complicated and of lower confidence than the disordered rules. In Table 4.2, the composition of hydrophobic residues is one of the most important criteria in differentiating DRs from ordered regions. In 62.5% of disordered and 87.5% of ordered rules, the composition of hydrophobic residues is chosen as the first test to decide order/disorder status.

4.5 Summary

In this chapter we have presented a system for predicting long DRs in proteins based on decision trees and reduced amino acid composition. Concise rules based on biochemical properties of amino acid side chains are generated for prediction. Coarser information extracted from the composition of amino acids can not only improve the prediction accuracy but can also increase the learning efficiency. Specifically, four groups of reduced-AAC based on hydrophobicity and polarity reduce the set of classification rules by 85%, produce more succinct and understandable rules, and improve the prediction accuracy. In cross-validation tests, with four groups of reduced amino acid composition, our system can achieve a recall of 80% at a 13% false positive rate for predicting DRs, and the overall accuracy can reach 83.4%. This prediction accuracy is comparable to most, and better than some existing predictors. Moreover, four groups of reduced-AAC improve the training time by three folds and also significantly reduce the time for prediction, which is a great advantage for large-scale sequence analysis.

Compared with previously described predictors of DRs including DISpro and DISOPRED2, our predictor has achieved a higher accuracy of prediction. Our approach also suits analysis of large numbers of protein sequences.

Chapter 5

Prediction of Long Disordered Regions Using Random Forest

In this chapter, a new algorithm, IUPforest-L, is proposed for predicting long DRs based on the random forest learning model [Breiman, 2001b] and physicochemical properties extracted from the protein sequences using the AAindex (amino acid indices) database [Kawashima et al., 1999]. In ten-fold cross-validation tests, IUPforest-L can achieve an area of 89.5% under the ROC curve. Ten-fold cross-validation tests and blind tests demonstrate that IUPforest-L can achieve significantly higher accuracy than many existing algorithms in prediction of long DRs. Compared with existing disorder predictors, IUPforest-L has high prediction accuracy and is efficient for predicting long DRs in large-scale studies. The server of IUPforest-L can be accessed from (`http://dmg.cs.rmit.edu.au/IUPforest/IUPforest-L.php`).

5.1 Motivation

The long DR predictor based on the decision tree presented in Chapter 4 is our first test in length specific DR predictors. While the three tree-based predictors proposed have successfully alleviated the difficulty of learning caused by imbalanced training databases, we believe performance improvements can be gained using the voting strategy of the ensemble method. Generally, ensemble learning can achieve higher accuracy of prediction given that a group of independent predictors with an accuracy of over 50% vote for the final prediction results. A suitable machine learning model is random forest [Breiman, 2001b], which may comprise hundreds of trees built from bootstrap random samples.

5.2 Materials and Methods

In this section, we first introduce training and test datasets. To compare with other predictors, we apply several publicly available datasets for blind tests. We then describe how order/disorder related physicochemical properties (indices) are selected from the AAindex database. The random forest machine learning model is then described and we illustrate why it is suitable for the prediction of DRs. Finally, the model of IUPforest-L, our long DR predictor is described. The selected indices are also used for prediction of short DRs as described in Chapter 6.

5.2.1 Training and Test Datasets

To train IUPforest-L, we apply the disordered database *LDDB2* built from DisProt (version 3.6) [Obradovic et al., 2003] as described in Section 3.1.1. The ordered database *ODB* is built from PDB-Select-25 (the Oct.2004 version) and is described in Section 3.1.3. There are 352 DRs and 366 structured segments included in databases *LDDB2* and *ODB*.

To assess the prediction performance of IUPforest-L, five datasets are used for blind tests. The first one is constructed by Hirose et al. (*Hirose-ADS1*) as a blind test dataset of POODLE-L [Hirose et al., 2007]. It contains 53 ordered regions of at least 40 amino acids (11,431 amino acids in total) from the Protein Data Bank [Berman et al., 2000] and 63 DRs of at least 30 amino acids (8,700 amino acids in total) from DisProt (version 3.0). For an objective blind test, the homologous sequences of these 63 DRs are removed from the original training set, whereas the final IUPforest-L is still trained on the whole training set. The second test set (*Han-ADS1*) comprises 53 ordered regions in *Hirose-ADS1* and 33 long DRs (5,959 amino acids in total) from the latest DisProt (version4.8), after removing DRs homologous to those in DisProt (version3.6) using the CD-HIT algorithm with a threshold of 0.9 sequence identity [Li and Godzik, 2006]. The third test set (*Peng-DB*) is constructed based on the blind test dataset of VSL2 [Peng et al., 2006], where 56 long DRs of at least 30 amino acids (2,841 amino acids in total) and 1,965 ordered regions (318,431 amino acids in total) are used in the assessment. The CASP7 database described in Chapter 3, and the complete blind test dataset of VSL2 are our last two blind test datasets.

Table 5.1: AAindex VINM940102.

H	VINM940102								
D	Normalised flexibility parameters (B-values) for each residue surrounded by none rigid neighbours (Vihinen et al., 1994)								
R	LIT:2014123 PMID:8090708								
A	Vihinen, M., Torkkila, E. and Riikonen, P.								
T	Accuracy of protein flexibility predictions								
J	Proteins 19, 141-149 (1994)								
C	VINM940101 0.940			MIYS990104 0.922			PARS000101 0.917		
A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
1.315	1.310	1.380	1.372	1.196	1.342	1.376	1.382	1.279	1.241
1.234	1.367	1.269	1.247	1.342	1.381	1.324	1.186	1.199	1.235

5.2.2 AAindex

AAindex [Kawashima et al., 1999] (version 9.1)[1] includes 544 numerical indices representing various physicochemical and biochemical properties of amino acids. Each index is a set of 20 numerical values representing the different physicochemical and biological properties of 20 amino acids. We use the term AAindex to refer to a specific amino acid index. When we talk about the AAindex database we explicitly say so in the main text.

The AAindex VINM940102 is shown in Table 5.1. The first and second rows give the access number and a simple description of the AAindex. The third and fourth rows show the LITDB (Literature Database compiled by PRF and maintained by the Protein Research Foundation, Osaka) entry number and the author name of the AAindex. The next two rows list the publication for the AAindex and its source. The seventh row reveals other highly correlated indices and the last row illustrates the AAindex values. So AAindex VINM940102 represents normalised flexibility parameters for each residue surrounded by none rigid neighbours. Indices VINM940101, MIYS990104 and PARS000101 are all highly correlated with VINM940102 with correlation coefficients 0.94, 0.922 and 0.917 respectively.

Not all indices are related to the order/disorder status of the amino acid. We calculate the correlation coefficient between each AAindex and the order/disorder regions in DisProt (version 3.1).

$$Zscore = \frac{x - \mu}{\sigma} \tag{5.1}$$

Given that different AAindices are not at the same scale, they are first scaled using the

[1] `http://www.genome.jp/aaindex/`

Z score (Equation 5.1). In the equation, x is a raw score to be standardised, μ is the mean of the population (20 numbers) and σ is the standard deviation of the population.

For each protein sequence in DisProt 3.1, residues are replaced by the corresponding 20 Z score values. The sequence is then smoothed by the Savitzky-Golay filter [Press et al., 2002]. From the perspective of biology, the transition between adjacent residues should be gradual instead of sudden. The Savitzky-Golay filter here essentially performs a polynomial regression on the sequence to determine the smoothed value for each residue. The Savitzky-Golay filter has the advantage of preserving features of the distribution such as relative maximum score, minimum score and width of disordered or ordered regions, which are usually "flattened" by other smoothing techniques. Therefore, the Savitzky-Golay filter is selected as our smoothing function.

The original training protein sequences from DisProt (version 3.1) can be replaced by a sequence of numbers 1 and -1, where 1 corresponds to disordered residues and -1 corresponds to unannotated residues. The sequence of numbers is also smoothed by the Savitzky-Golay filter. The correlation coefficient can be calculated between these two sequences of numbers to select indices most related to order/disorder status. We have used the Pearson correlation coefficient $r_{sa,sb}$ [Mendenhall et al., 2003] as shown in Equation 5.2.

$$r_{sa,sb} = \frac{\sum_{i=1}^{n}(sa_i - \overline{sa})(sb_i - \overline{sb})}{(n-1)\sigma_{sa}\sigma_{sb}} \tag{5.2}$$

where sa is the sequence of AAindex numbers after smoothing and sb is the sequence of numbers 1 and -1 after smoothing. sa_i and sb_i are values of position i in sequence sa and sb. $\overline{sa}$ and $\overline{sb}$ are the mean values of sa and sb. σ_{sa} and σ_{sb} are the standard deviations of sa and sb and n is the total number of residues in the sequence.

For each of 544 AAindices, the correlation coefficient is calculated for all training sequences containing both disordered and unannotated residues. There are 359 sequences containing both disordered and unannotated residues in DisProt3.1. We list 20 indices most positively correlated to disorder in Table 5.2; and 20 indices most negatively correlated to disorder in Table 5.3. PCS, NCS and ACC in these tables are abbreviations for positively correlated sequences, negatively correlated sequences and average correlation coefficient.

These two tables illustrate many indices known to be related to DRs. In Table 5.2, indices VINM940101, VINM940102 and VINM940103 are normalised flexibility parameters under different conditions. DRs are more flexible than ordered regions and these indices

Table 5.2: Selected indices positively related to disorder

Name of AAindex	#PCS	#NCS	ACC	AAindex description
VINM940102	297	62	0.248	Normalised flexibility parameters (B-values) for each residue surrounded by none rigid neighbours
MIYS990102	297	62	0.241	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
MIYS990101	296	63	0.242	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
RACS770101	295	64	0.237	Average reduced distance for C-alpha
MIYS990104	295	64	0.245	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
MEIH800101	294	65	0.238	Average reduced distance for C-alpha
MIYS990103	294	65	0.244	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
OOBM770103	293	66	0.229	Long range non-bonded energy per atom
PARJ860101	293	66	0.233	HPLC parameter
BULH740101	292	67	0.223	Transfer free energy to surface
VINM940101	292	67	0.239	Normalised flexibility parameters (B-values), average
MUNV940103	292	67	0.231	Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices
MIYS990105	292	67	0.239	Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues
FASG890101	292	67	0.233	Hydrophobicity index
VINM940103	291	68	0.230	Normalised flexibility parameters (B-values) for each residue surrounded by one rigid neighbours
PUNT030102	290	69	0.212	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases
GUYH850102	290	69	0.243	Amino acid side-chain partition energies and distribution of residues in soluble proteins
CHOP780203	289	70	0.208	Normalised frequency of beta-turn
WOLS870101	289	70	0.215	Principal property value z1
PARS000101	289	70	0.245	p-Values of mesophilic proteins based on the distributions of B values

are all positively correlated with disorder. Similarly, indices MIYS990101, MIYS990102, MIYS990103, MIYS990104 and MIYS990105 are positively correlated with disorder. These indices are a self-consistent estimation of inter-residue protein contact energies with minor difference. Existing predictors IUPred and Ucon have adopted an approach based on contact energy to build their predictors. In Table 5.3, indices BASU050101, BASU050102 and BASU050103 indicate the eigenvectors of contact matrices and hydrophobicity profiles.

Table 5.3: Selected indices negatively related to disorder

Name of AAindex	#PCS	#NCS	ACC	AAindex description
BASU050101	302	57	-0.244	Interactivity scale obtained from the contact matrix
ZHOH040101	302	57	-0.243	The stability scale from the knowledge-based atom-atom potential
CIDH920103	300	59	-0.219	Normalised hydrophobicity scales for alpha+beta-proteins
NOZY710101	300	59	-0.232	Transfer energy, organic solvent/water
PONP930101	300	59	-0.239	Hydrophobicity scales
BASU050102	299	60	-0.255	Interactivity scale obtained by maximising the mean of correlation coefficient over single-domain globular proteins
BASU050103	299	60	-0.241	Interactivity scale obtained by maximising the mean of correlation coefficient over pairs of sequences sharing the TIM barrel fold
CIDH920104	297	62	-0.234	Normalised hydrophobicity scales for alpha/beta-proteins
CIDH920105	297	62	-0.237	Normalised average hydrophobicity scales
MANP780101	297	62	-0.235	Average surrounding hydrophobicity
VENT840101	297	62	-0.235	Hydrophobicity parameters and the bitter taste of L-amino acids
ZHOH040102	297	62	-0.226	The relative stability scale extracted from mutation experiments
ZHOH040103	297	62	-0.244	Quantifying the effect of burial of amino acid residues on protein stability
CIDH920101	296	63	-0.230	Normalised hydrophobicity scales for alpha-proteins
NISK860101	296	63	-0.248	Radial locations of amino acid residues in a globular protein: Correlation with the sequence
PTIO830102	296	63	-0.233	Beta-coil equilibrium constant
NADH010105	296	63	-0.227	Hydropathy scale based on self-information values in the two-state model
CIDH920102	295	64	-0.237	Normalised hydrophobicity scales for beta-proteins
LIFS790101	295	64	-0.238	Conformational preference for all beta-strands
PONP800101	295	64	-0.233	Surrounding hydrophobicity in folded form

Ordered regions are usually hydrophobic and DRs are hydrophilic. Indices ZHOH040101, ZHOH040102 and ZHOH040103 are negatively correlated with disorder. These indices either represent the relative stability scale or the buriability of amino acids. It is known that ordered regions are more stable than DRs and are likely to be buried.

Although these disordered related indices discussed above are discovered from different experiments under various physicochemical conditions, the correlation coefficient value among them can be very high (above 0.8). For example, the correlation coefficient between VINM940102 and VINM940104 is 0.922. Given a correlation threshold less than 0.8, only four features from the top 40 AAindices can be selected. They are VINM940102, BULH740101,

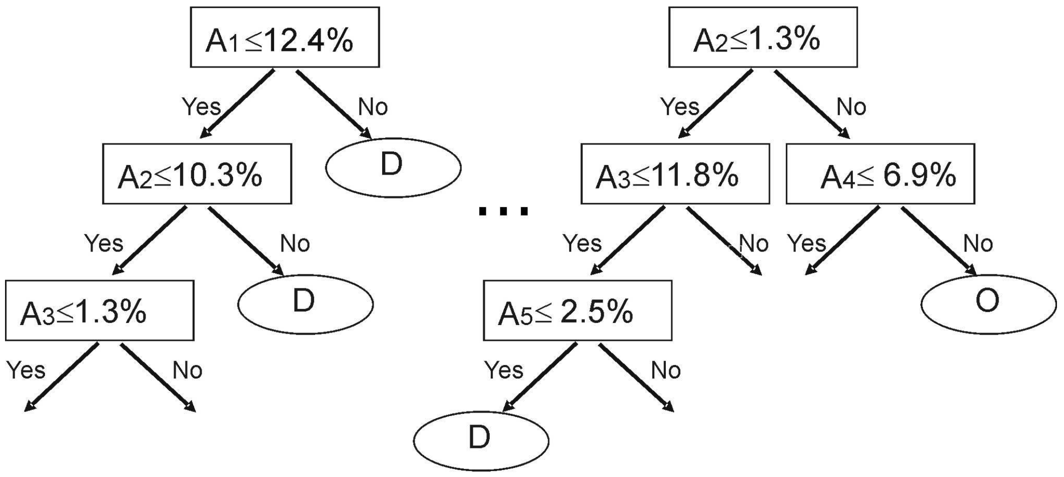


Figure 5.1: A sample random forest

PUNT030102 and CHOP780203 which correspond to the flexibility parameter, hydrophobicity scale, propensity scale and protein secondary structure. These features or their highly correlated counterparts have been adopted before for disorder prediction and have proven to be very effective [Li et al., 1999; Linding et al., 2003a; Hirose et al., 2007]. However it is the first time that a large number of AAindex features are selected and applied simultaneously to protein disorder prediction. We compared the predictor generated from these four features with the predictor generated from 40 features. The predictor based on 40 features achieved a higher accuracy of prediction with an improvement of the AUC value by more than 3%.

5.2.3 The Random Forest Machine Learning Model

In recent years, the random forest algorithm [Breiman, 2001b] has become widely used in the bioinformatics community for classification of microarray and other high-dimensional molecular data [Wu et al., 2003; Lee et al., 2003; Diaz-Uriarte and de Andres, 2006]. Random forest has been applied to protein protein interaction prediction [Qi et al., 2005] and classification of real and pseudo microRNA precursors [Jiang et al., 2007]. Notably, random forest predictors have been shown to have a predictive performance comparable to that of the best performing alternatives (including SVMs) for classification of microarray gene expression data [Diaz-Uriarte and de Andres, 2006].

A random forest is an ensemble of unpruned decision trees, an example of which is shown in Figure 5.1. In the decision tree on the left, the node at the root tests a feature, such as the first order autocorrelation function of the normalised flexibility parameters (see Section 5.2.4). If it is higher than a given threshold then the residue is in a disordered state (the right branch labelled D); otherwise, another input feature is tested and a set of other tests are further performed until a decision is made. A random forest can comprise hundreds of decision trees. Each tree is grown to full length using a bootstrap subset of the training dataset [Breiman, 2001b]. Bootstrapping is a resampling technique where a number of bootstrap training sets are drawn randomly from the original training set with replacement. The number of trees in the forest is adjustable. To predict an instance of unknown class label, each tree casts a unit prediction vote. The forest selects the prediction having the most votes over all the trees in the forest. Compared with the decision tree predictor, random forests have better accuracy of prediction, are more tolerant of noise and are less dependent on the training datasets [Dosztanyi et al., 2006].

5.2.4 Features

When a window of w residues slides along a sequence, six types of features are derived from residues within the window, as defined below.

1. Auto-correlation function of AAindices. Each residue in the training set is replaced with a value of the normalised AAindex, which is a set of 20 numerical values representing the physicochemical and biochemical properties of 20 amino acids chosen from the AAindex database. As such, a sequence of N amino acids in the training set is first transformed into a numerical sequence [Feng and Zhang, 2000; Bu et al., 1999], and denoted as: $P_1 P_2 \cdots P_i \cdots P_{i+w} \cdots P_N$.

   Then the sequence is smoothed with the Savitzky-Golay filter [Savitzky and Golay, 1964]. The Moreau-Broto autocorrelation function F_d of an AAindex is then calculated within a window, which is defined as Equation 5.3.

$$F_d = \frac{1}{w-d} \sum_{i=1}^{w-d} P_i \times P_{i+d}, (d = 1, 2, \cdots, w-1) \tag{5.3}$$

   where w is the window size, p_i and p_{i+d} are the AAindex values at positions i and $i+d$ respectively [Feng and Zhang, 2000; Bu et al., 1999]. For example, when $d = 1$, the numerical value for each residue (i) in the window is multiplied by the value of the next nearby residue (i+1) and F_1 is the average of these w-1 products. Similarly, F_2 is the average of the w-2 products generated from every other residue. The value d represents the order of the correlation and is tuned to optimise the prediction performance. The F_d (d=1, 2,$\cdots$, 15) for the 40 sets of indices (listed in Tables 5.2 and 5.3) is calculated and evaluated in training IUPforest-L.

2. The mean hydrophobicity, defined as the average value of Kyte and Doolittle's hydrophobicity index [Kyte and Dolittle, 1982] in the window.

3. The modified hydrophobic cluster value [Coeytaux and Poupon, 2005], calculated as the longest hydrophobic cluster in the window divided by the window size. A hydrophobic cluster is defined as a sequence without proline (P) and four (or more) continuous non-hydrophobic residues (ACDEGHKNQRST).

4. The mean net charge within the window and local mean net charge within a 13 amino acid fragment centred at the middle residue. Residues K and R are defined as +1; D and E are defined as -1; other residues are 0.

5. The mean contact number, defined as the mean expected number of contacts in the globular state of all residues within the window [Garbuzynskiy et al., 2004].

6. The composition of four reduced amino acid groups [Dosztanyi et al., 2006] and the Shannon's entropy (K2) of the AAC within the window. The Shannon entropy [Shannon, 1948] for the window is calculated as:

$$K2 = -\sum_{i=1}^{M} f_i log_2 f_i \tag{5.4}$$

where M is the number of different groups of residues in the window and f_i is the frequency of residue i.

5.2.5 IUPforest-L

A flow chart of IUPforest-L is shown in Figure 5.2. The sequence features are calculated when a window slides along a protein sequence.

At the training stage, six types of features listed above are calculated when a window of w amino acids slides from the N-terminal end to the C-terminal end of a protein sequence. Each window is tagged with a label of disorder (Positive or P) or order (Negative or N)

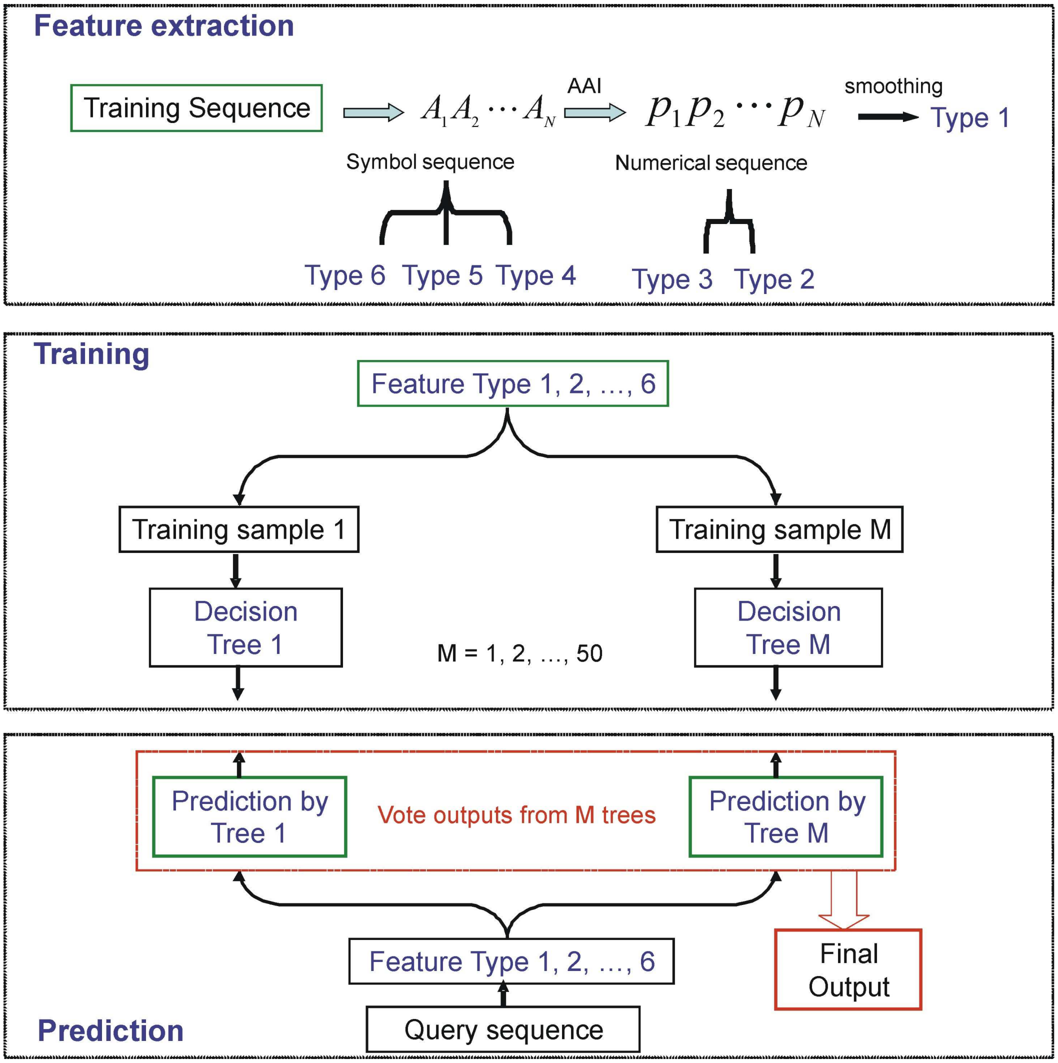


Figure 5.2: A flow chart of IUPforest-L.

according to the label of the central residue, and IUPforest-L models are trained from the six types of features and the prediction result could be obtained by each of the trees in the forest. The final score is the combination of outcomes from all trees by voting and after having been smoothed with the Savitzky-Golay filter [Savitzky and Golay, 1964]. A threshold that best classifies the ordered or disordered state of a residue can then be defined based on the scores and the optimal evaluated values in ten-fold cross-validation tests.

At the prediction stage, the features are first calculated when a window slides over a query sequence and then a probability score of a residue being disordered is assigned by each of the trees in the forest. The final score of IUPforest-L in the prediction is the combination of the outcomes from all trees by voting. A region is predicted as disordered only when 30 or more consecutive amino acid residues are predicted to be disordered.

5.2.6 Evaluations

To estimate the generalisation accuracy, ten-fold cross-validation tests are conducted, where in each fold 90% of the sequences in the training set are randomly used in training and the other 10% are used in the test. The process is repeated for the entire dataset and the final result is the average of the results from ten folds. In addition, independent tests are performed on *Hirose-ADS1* [Shimizu et al., 2007a], *Han-ADS1*, *Peng-DB* [Peng et al., 2006], the CASP7 targets and the blind test dataset of VSL2.

5.3 Results

5.3.1 Ten-fold Cross-Validation

The ROC curves of IUPforest-L in ten-fold cross-validation tests using a window of 31 amino acids are shown in Figure 5.3. With the type 1 features (the autocorrelation function of AAindices), a forest of more trees has better prediction accuracy. For example, the AUC increases by 2% when the number of trees is increased from 10 to 50. However, the prediction accuracy increases only modestly when the number of trees increases further from 50 to 100, while the training and prediction time increases significantly. Detailed test results on the time efficiency with number of trees from 10 to 300 are shown in Section 5.3.2.

The default setting of IUPforest-L is a forest of 50 trees for large-scale applications. With a forest of a fixed number of trees, the ROC curve trained with the autocorrelation function with a d value of between 1 and 15 almost overlaps with the ROC curve trained with d between 1 and 30. This result indicates that continuous correlations between nearby residues from 1 to 15 along the sequence could determine whether the fragment is involved in a long DR.

Figure 5.3 shows that training with either type 1 or the combination of type 2-6 features could reach the 70.5% or 70.0% true positive rate with a 10% false positive rate, while a combination of type 1-6 features could lead to a higher true positive rate of 76%, a MCC value of 0.67, a Sproduct value of 0.64 and an area of 89.5% under the ROC curve. This result indicates that type 1 and type 2-6 features have redundant, but complementary structural

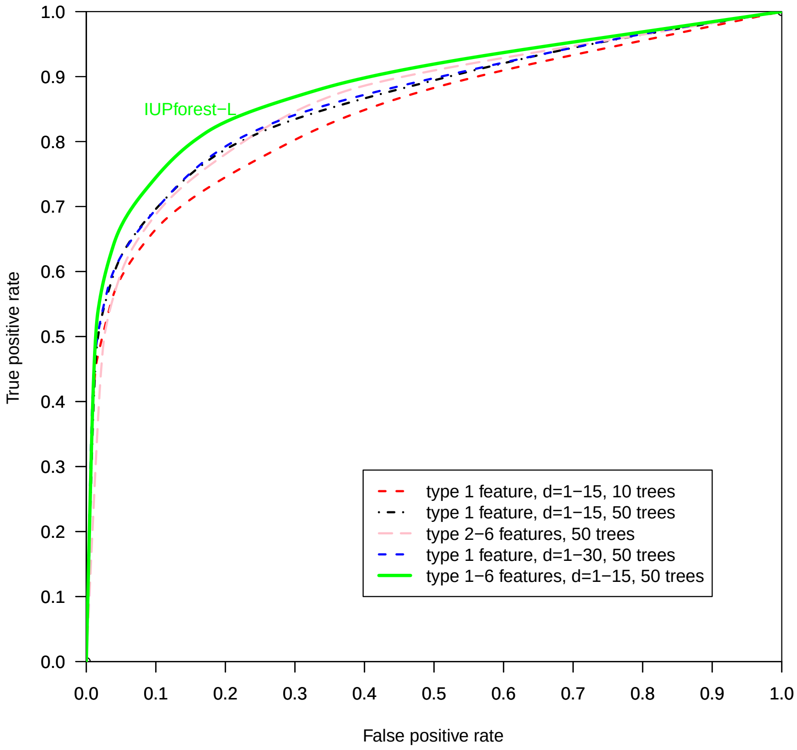


Figure 5.3: IUPforest-L ROC curves of ten-fold cross-validation tests.

information. Type 2-6 features generate only ten parameters in total within a given window, while type 1 features could generate hundreds of parameters that take into account both order information and physicochemical properties. It has been shown that the random forest model has no risk of overfitting with an increasing number of trees when the input parameters increase [Breiman, 2001b]. As such, using type 1 features to train the random forest could extract more sequence-structure information [Han et al., 2009a] and it is thus conjectured that better prediction accuracy could be achieved with the autocorrelation functions generated from AAindices combined with other features of type 2-6. The improvement of AUC from the red curve with a type 1 feature and 10 trees to the green curve with features 1-6 and 50 trees is statistically significant (P-value $< 10^{-4}$).

The window size and step size for sliding the window are additional parameters for tuning the performance of the IUPforest-L models. The window should be of a reasonable size so that the amino acid indices based correlation can be of significance within a reasonable training or test time. Training with small windows increases training time and can introduce

Table 5.4: Time for training IUPforest models under different numbers of trees.

Number of trees	Training time
10	28 mins 33 secs
50	120 mins 7 secs
100	186 mins 10 secs
200	360 mins 40 secs
300	system crash after 24 hours.

noise, whereas training with large windows can lose local information. Our results indicate that from a window size of 19 amino acids to 47 amino acids, the random forest gives more stable results on blind test set *Han-ADS1*, but the accuracy for ten-fold cross-validation test on the training set will drop with larger window size (details discussed in Section 5.3.3).

To batch predict long DRs, a window size of 31 amino acids is set in default to keep the balance between high efficiency and accuracy. The step size for sliding windows can also affect accuracy and overall time efficiency at both the training and test stage. If the step size is too small, when a window slides along a sequence, it will introduce redundancy between windows and prolong the time for training models. Our experiments in Section 5.3.3 show that a sliding step of 20 amino acids (default setting) produces models with stable sensitivity without significantly prolonging the training process.

5.3.2 The Number of Trees and Time Efficiency

It has been reported that with a larger number of trees in a random forest, rather than overfitting, the prediction error for the forest converges to a limit [Breiman, 2001b]. The number of trees for which a forest converges depends on the application. Ten-fold cross-validation tests on the training data are conducted to examine the performance of IUPforest under different numbers of trees. The results indicate that prediction accuracy can be consistently improved with an increase in the number of trees up to 50 without significantly increasing training time. Prediction accuracy is only modestly improved with the number of trees growing from 50 to 100, while the time for training forests increases significantly as shown in Table 5.4.

The timing experiments are carried out on a computer with the following specifications:

Processor:	Intel(R) Core(TM)2 Duo CPU E6550 @ 2.33GHz
RAM:	2G
Hard disk drive:	Western Digital WDC WD1600AAJS-60WAA0 (160GB)
Operating system:	Ubuntu GNU/Linux 8.04 (kernel: 2.6.24-24-generic)

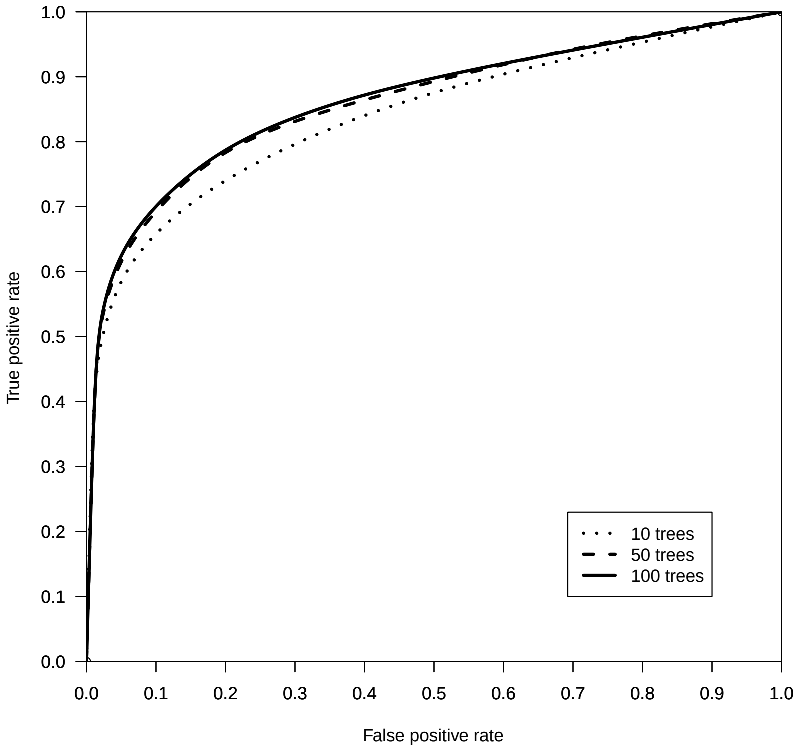


Figure 5.4: The prediction accuracy of IUPforest under different number of trees in ten-fold cross-validation tests.

The ten-fold cross-validation test accuracy of IUPforest under different numbers of trees is shown in Figure 5.4. Training features are type 1 features with d =1,2,...,15. Fifty trees are the default setting for IUPforest. When the number of the trees increases from 10 to 50, the increase in AUC is not significant with a P-value = 0.18. But the difference in the true positive rate is apparent when the false positive rate is less than 40%. The AUC difference is trivial when the number of the trees increases from 50 to 100.

5.3.3 The Windows for Training IUPforest-L

Training with small windows increases training time and can introduce noise, whereas training with large windows can lose local information. Figure 5.5 shows the ROC curves for ten-fold cross-validation tests on the training set with windows of 31 amino acids and 41 amino acids, and blind tests on *Han-ADS1* with the IUPforest models trained with different window sizes. Training features are type 1 features with d =1,2,...,15. It can be seen that although

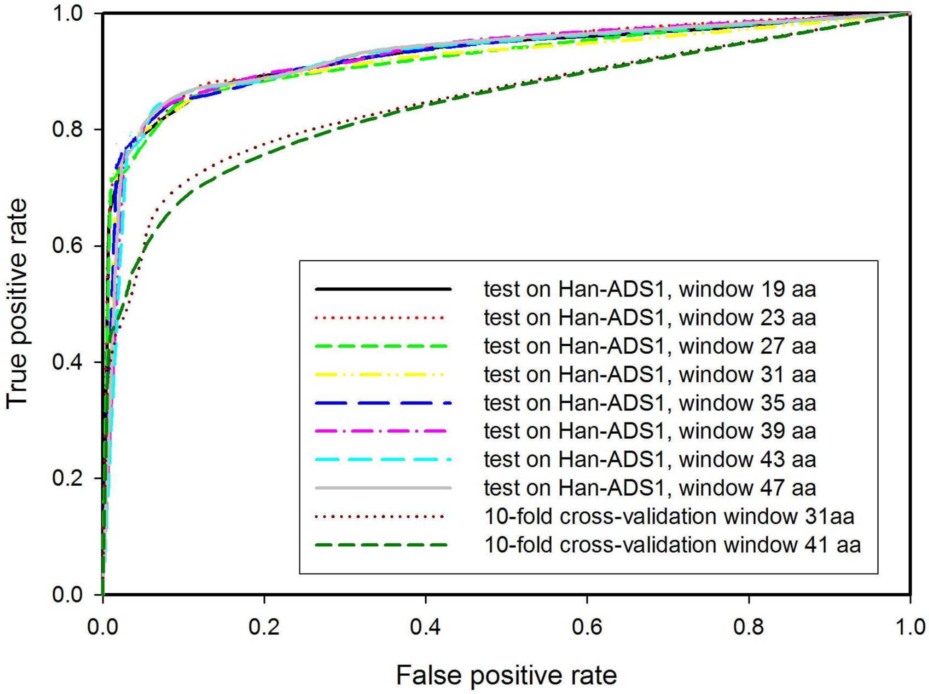


Figure 5.5: ROC curves under different window sizes.

independent test results on *Han-ADS1* are stable between windows of 19 amino acids to 47 amino acids, ten-fold cross-validation test accuracy drops with larger window size. This result is different from that observed in Chapter 4, in which larger windows usually present better accuracy of prediction. This may be due to the different features and machine learning models applied. As a result, 31 amino acids is set as the default window size for large-scale prediction to keep the balance between high efficiency and accuracy.

The step size for sliding windows can also affect the accuracy and overall time efficiency at both the training and test stage. If the step size is too small, when a window slides along a sequence, it will introduce redundancy between windows and prolong the time for training models. According to the ten-fold cross-validation test results with sliding steps of one amino acid and twenty amino acids for an IUPforest model defined by type 1 features of only five sets of AAindices, when a sliding step increases from one amino acid to twenty amino acids, the AUC will drop by about 3%, but the time efficiency will increase four fold. To ensure an efficient large-scale application, a sliding step of 20 amino acids is the default setting.

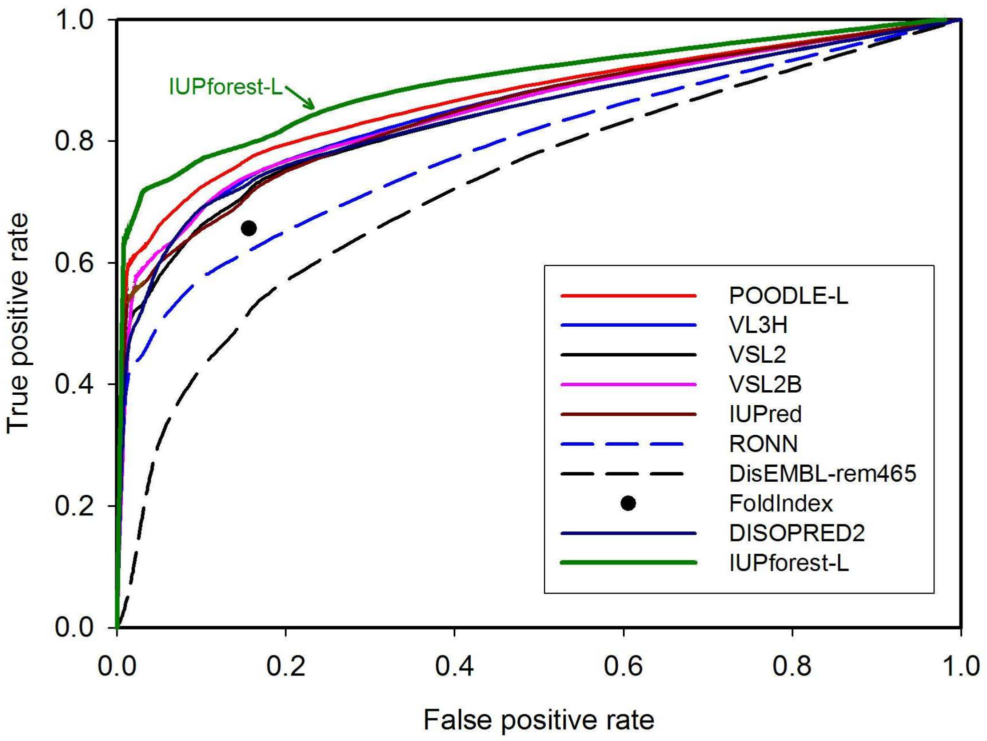


Figure 5.6: ROC curves of IUPforest-L and other predictors on the test set Hirose-ADS1*.*

5.3.4 Blind Tests

Figure 5.6 depicts the ROC curves on the blind test dataset *Hirose-ADS1* for IUPforest-L and nine other publicly available predictors, including the most recently developed POODLE-L [Hirose et al., 2007] and the well-established predictor VSL2 [Peng et al., 2006].

IUPforest-L outperforms all other predictors in terms of the AUC in predicting long DRs. At low false positive rates (<10%), IUPforest-L achieves the highest sensitivity among all the predictors. The improvement in the AUC value is significant (P-value $< 10^{-5}$) for all predictors except POODLE-L. In terms of other performance measures listed in Table 5.5, IUPforest-L is also comparable to or better than other predictors.

Figure 5.7 and Table 5.6 show the results of comparisons of IUPforest-L with POODLE-L and other predictors on *Han-ADS1.* It can be seen that IUPforest-L always performs better than most of them. Compared to predictors RONN and DisEMBL-rem465, the improvement in the AUC value is statistically significant (P-value $< 10^{-6}$). The AUC values of IUPforest-L are also marginally larger than VSL2B, IUPred and POODLE-L. Figure 5.8 and Table 5.7 show the results of comparisons of IUPforest-L with POODLE-L and other predictors on

Table 5.5: Comparison of IUPforest-L with other predictors on the test set Hirose-ADS1

Measure(%)	Sensitivity	Specificity	MCC	Spro	Bacc	Sw	AUC
IUPforest-L	72.0	93.4	72.3	68.7	84.3	68.7	90.0
DisEMBL	24.5	96.4	31.2	23.6	60.5	20.9	73.3
DISOPRED2	63.5	93.9	61.6	59.6	78.7	57.4	84.8
FoldIndex	62.2	84.4	48.1	52.5	73.3	46.6	N/A
FoldUnfold	59.8	95.9	61.4	57.4	77.9	55.7	N/A
IUPred	59.5	95.6	60.7	56.9	77.6	55.1	85.3
POODLE-L	66.9	94.9	65.8	63.5	80.9	61.8	87.3
RONN	62.8	83.7	47.8	52.5	73.3	46.5	79.7
Spritz (long)	16.5	92.5	13.9	15.2	54.5	-2.65	N/A
VL3H	73.4	85.8	60.0	63.0	79.6	59.2	85.6
VSL2	75.5	79.4	54.7	59.9	77.5	54.9	84.4
VSL2B	60.9	83.0	60.9	59.1	79.6	59.1	85.4

Table 5.6: Comparison of IUPforest-L with other predictors on the test set Han-ADS1

Measure(%)	Sensitivity	Specificity	MCC	Spro	Bacc	Sw	AUC
IUPforest-L	87.5	94.4	82.3	82.6	91.0	82.1	94.3
DisEMBL	43.9	93.2	44.4	40.9	68.5	37.8	76.5
DISOPRED2	63.1	96.2	65.8	60.7	79.7	60.2	90.0
FoldIndex	67.9	84.4	52.5	57.3	76.2	52.7	N/A
FoldUnfold	85.5	87.7	71.0	75.0	86.6	73.4	N/A
IUPred	71.4	95.6	71.4	68.3	83.5	67.6	92.0
POODLE-L	82.2	94.9	78.8	78.0	88.6	77.4	94.0
RONN	54.9	96.7	60.2	53.1	75.8	52.6	86.6
Spritz (long)	62.1	96.7	65.8	60.0	79.4	59.3	N/A
VL3H	66.4	96.1	68.3	63.8	81.3	63.2	94.5
VSL2	87.1	89.3	75.1	77.8	88.2	76.4	84.4
VSL2B	70.4	94.8	69.3	66.8	82.6	65.8	91.0

Peng-DB. It can be seen again that IUPforest-L always performs better than most of them. The AUC value of IUPforest-L is significantly larger (P-value < 0.03) than that of IUPred, DisEMBL-rem465 and VSL2B. The blind test on *Peng-DB* is different from *Hirose-ADS1* and *Han-ADS1*. Each sequence in *Peng-DB* may contain both ordered and disordered residues, while sequences in *Hirose-ADS1* and *Han-ADS1* are completely disordered or ordered. The blind test results on *Hirose-ADS1*, *Han-ADS1* and *Peng-DB* reveal that IUPforest-L performs consistently better than other DR predictors on sequences completely disordered or ordered as well as on sequences with DRs.

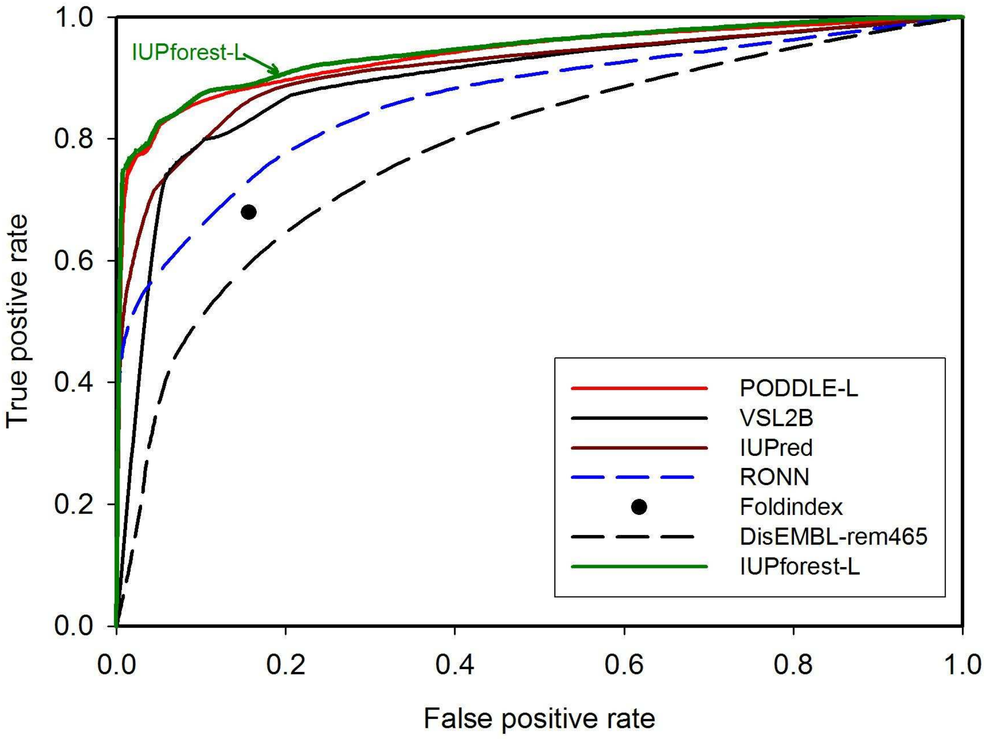


Figure 5.7: ROC curves of IUPforest-L and other predictors on the test set Han-ADS1.

Table 5.7: Comparison of IUPforest-L with other predictors on the test set Peng-DB

Measure(%)	Sensitivity	Specificity	MCC	Spro	Bacc	Sw	AUC
IUPforest-L	39.1	98.3	24.8	38.4	68.7	41.0	83.2
DisEMBL	43.2	89.9	10.1	38.8	66.5	33.1	73.7
DISOPRED2	40.9	96.7	19.3	39.5	68.8	39.6	80.0
FoldIndex	49.1	84.4	85.8	41.4	66.7	35.7	N/A
FoldUnfold	37.9	90.9	31.0	34.5	64.4	19.0	N/A
IUPred	43.2	92.8	12.8	40.1	68.0	36.1	75.2
POODLE-L	39.9	96.8	18.6	38.9	68.3	36.7	79.8
RONN	42.3	96.3	18.3	40.7	69.3	38.6	79.2
Spritz (long)	23.1	96.9	18.6	22.4	60.0	-19	N/A
VL3H	40.2	96.8	11.8	38.9	68.5	65.6	91.3
VSL2	58.2	92.9	14.6	54.1	75.6	61.0	85.4
VSL2B	42.2	95.7	16.9	40.4	69.0	38.0	78.9

5.4 The IUPforest-L Server

We have developed a public server to provide an online disorder prediction service. The home page is: `http://dmg.cs.rmit.edu.au/IUPforest/IUPforest-L.php` as shown in Fig-

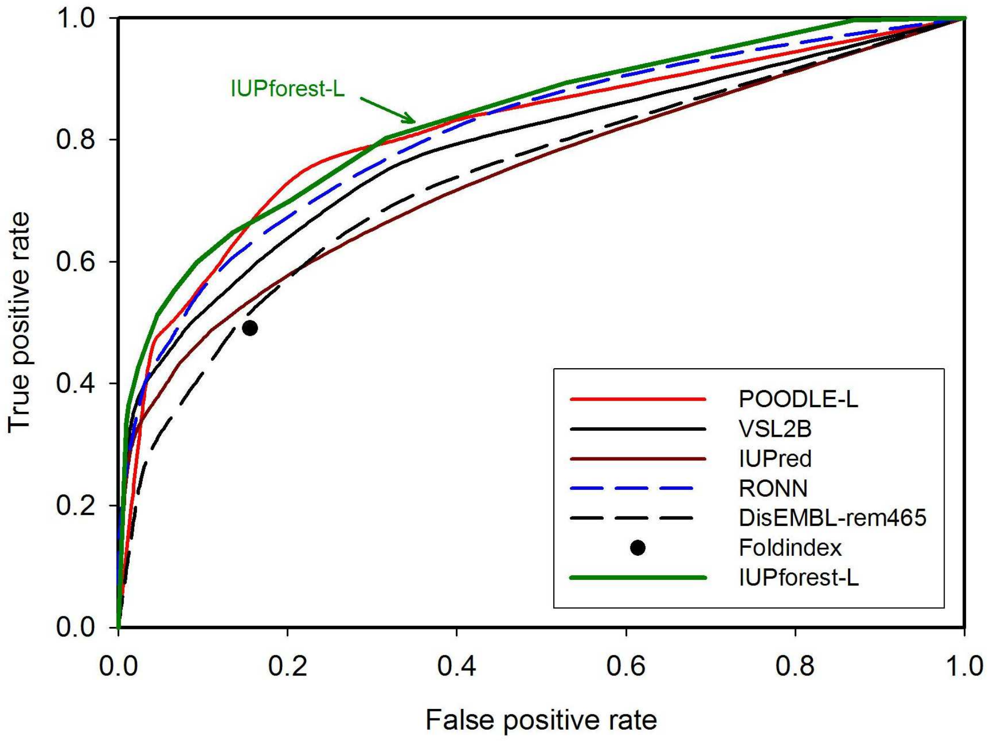
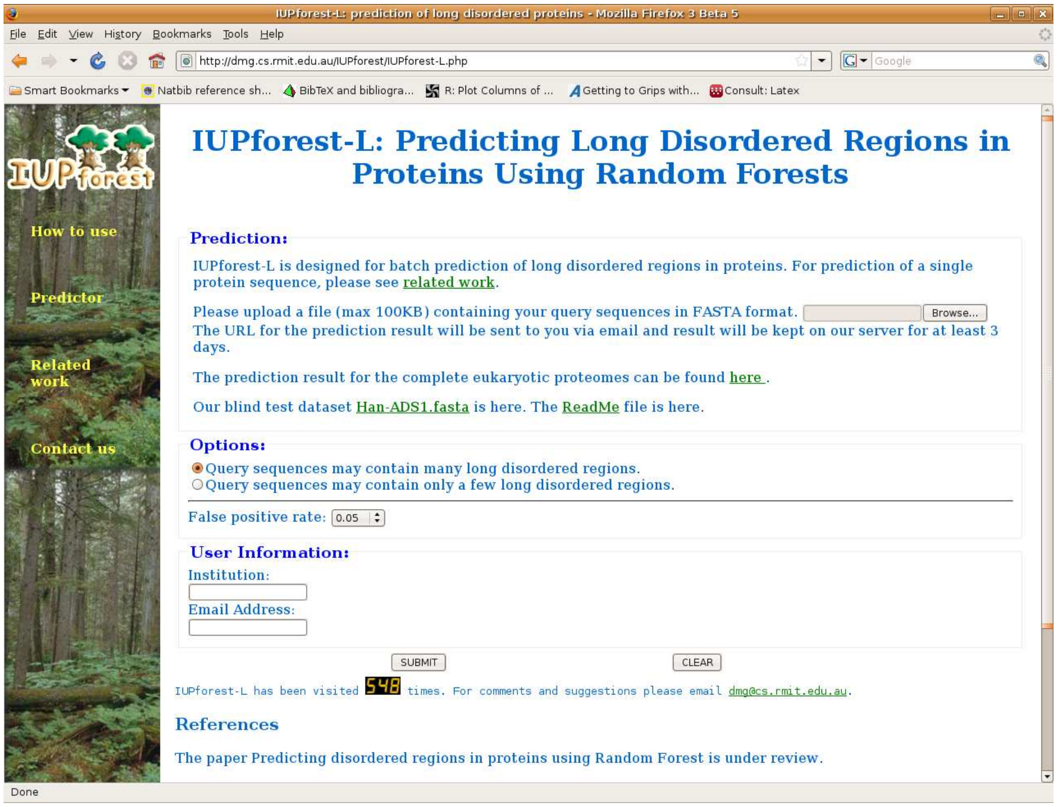


Figure 5.8: ROC curves of IUPforest-L and other predictors on the test set Peng-DB*.*

ure 5.9. Users upload a file with query sequences in FASTA format. At the server side, we check the format of these sequences and predict them using IUPforest-L. Results of predictions are stored on our server. As soon as the prediction is complete, we send an email containing the URL for the results of the prediction. Different cutoff values for disorder can be selected, which lead to recall at various false positive rates.

We have predicted the complete 62 eukaryotic proteomes using IUPforest-L and stored the results (`http://dmg.cs.rmit.edu.au/IUPforest/Eukaryota-L.php`). The page is shown in Figure 5.10. According to our results of prediction, Toxoplasma gondii is the most long DR populated proteome from among the 62 proteomes. Among 492 sequences and 378,364 residues in this proteome, 1,548 long DRs have been found from 348 sequences. In other words, 70.7% of the total sequences contain long DRs. The total predicted disordered residues in long DRs is 25.3% of the total residues.

In contrast, Guillardia theta is the proteome with the least long DRs. From 598 sequences, only 28 query sequences were predicted containing long DRs. In other words, 4.7% of the total sequences contain long DRs and disordered residues on long DRs are 0.7% of the total



IUPforest-L: Predicting Long Disordered Regions in Proteins Using Random Forests

Prediction:

IUPforest-L is designed for batch prediction of long disordered regions in proteins. For prediction of a single protein sequence, please see related work.

Please upload a file (max 100KB) containing your query sequences in FASTA format.
The URL for the prediction result will be sent to you via email and result will be kept on our server for at least 3 days.

The prediction result for the complete eukaryotic proteomes can be found here .

Our blind test dataset Han-ADS1.fasta is here. The ReadMe file is here.

Options:

● Query sequences may contain many long disordered regions.
○ Query sequences may contain only a few long disordered regions.

False positive rate: 0.05

User Information:

Institution:

Email Address:

IUPforest-L has been visited 548 times. For comments and suggestions please email dmg@cs.rmit.edu.au.

References

The paper Predicting disordered regions in proteins using Random Forest is under review.

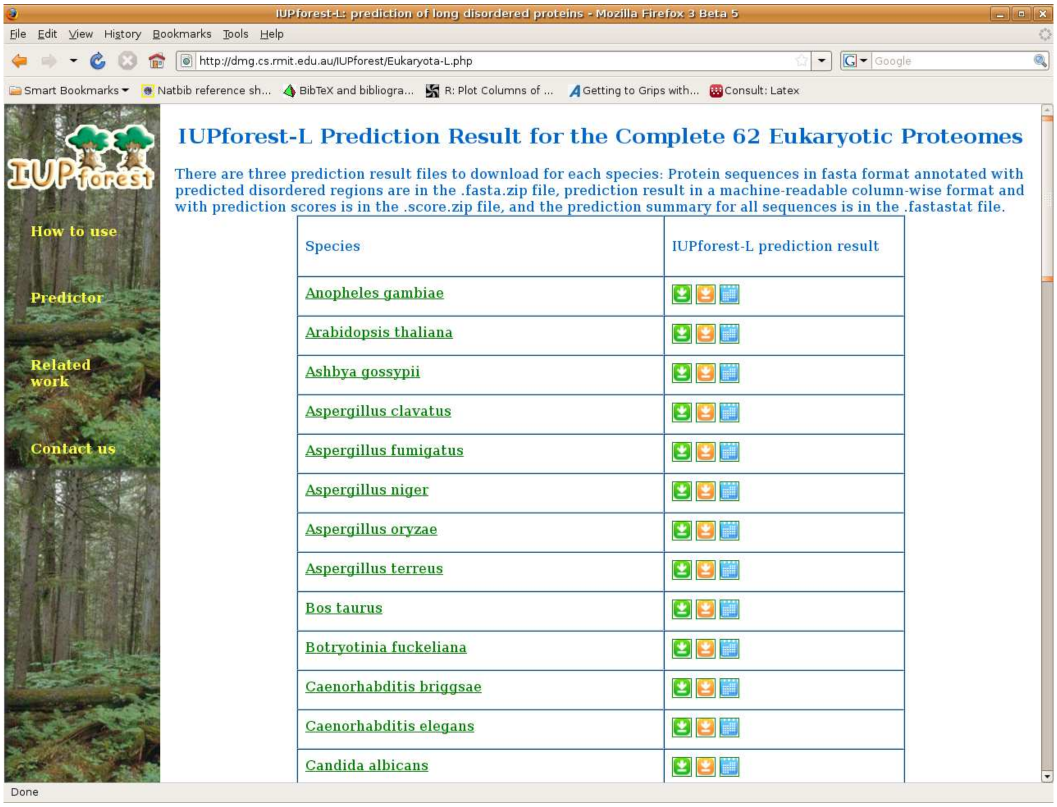
Figure 5.9: IUPforest-L server

residues.

5.5 Evaluation on Short Disordered Database

Peng et al. [2005] suggest that short DRs have different characteristics from long DRs. Predictors that perform well at predicting long DRs can perform considerably worse on DRs of less than 30 consecutive residues. We apply IUPforest-L to predict CASP7 targets to measure the performance of IUPforest-L on short DRs. The accuracy of prediction is shown in Figure 5.11.

The dotted line represents the performance of IUPforest-L on all DRs of CASP7. The solid line in Figure 5.11 is the accuracy of prediction on long DRs (>30 amino acids) in CASP7 (CASP7 contains mainly short DRs and has only four long DRs). We also apply IUPforest-L to predict the blind test dataset of VSL2 and produce the ROCs, as shown in Figure 5.12.



IUPforest-L Prediction Result for the Complete 62 Eukaryotic Proteomes

There are three prediction result files to download for each species: Protein sequences in fasta format annotated with predicted disordered regions are in the .fasta.zip file, prediction result in a machine-readable column-wise format and with prediction scores is in the .score.zip file, and the prediction summary for all sequences is in the .fastastat file.

Species	IUPforest-L prediction result
Anopheles gambiae	⬇ ⬇ ▦
Arabidopsis thaliana	⬇ ⬇ ▦
Ashbya gossypii	⬇ ⬇ ▦
Aspergillus clavatus	⬇ ⬇ ▦
Aspergillus fumigatus	⬇ ⬇ ▦
Aspergillus niger	⬇ ⬇ ▦
Aspergillus oryzae	⬇ ⬇ ▦
Aspergillus terreus	⬇ ⬇ ▦
Bos taurus	⬇ ⬇ ▦
Botryotinia fuckeliana	⬇ ⬇ ▦
Caenorhabditis briggsae	⬇ ⬇ ▦
Caenorhabditis elegans	⬇ ⬇ ▦
Candida albicans	⬇ ⬇ ▦

Figure 5.10: IUPforest-L prediction result for the complete 62 eukaryotic proteomes

From our experimental results in Figures 5.11 and 5.12, we notice that IUPforest-L is very successful at predicting long DRs; however, the performance of short DR prediction is poor, which reduces the overall DR prediction accuracy significantly. The prediction of short DRs is therefore different from long DR prediction and the predictor has to be very sensitive to pinpoint small segments of DRs. For IUPforest-L, the poor performance on short DRs of CASP7 and the blind test dataset of VSL2 can be attributed to long range MOREAU-BROTO correlation (a larger d value) not providing enough information for short DR prediction.

5.6 Discussion

Protein structures are stabilised by numerous intramolecular interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bonds. The autocorrelation function tests whether the physicochemical property of one residue is independent to that of neighbouring

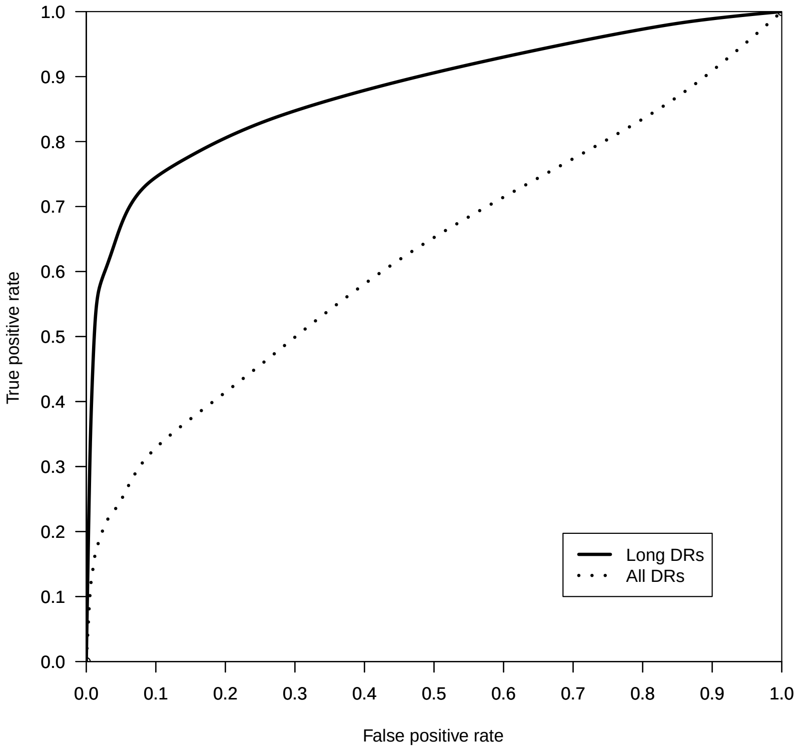


Figure 5.11: ROC curves of IUPforest-L on the test set CASP7.

residues. A group of residues involved in ordered structure close to other groups of residues in space will be dynamically constrained by the backbone or side chain interactions from these residues, and hence the residues in both groups will show higher density in the contact map or have higher pairwise correlation. Conversely, a repetitive sequence of amino acids can also produce a significant positive correlation for all physicochemical properties. Therefore, residues within a fragment exhibiting a higher autocorrelation may either be structurally constrained, or have low sequence complexity. The random forest learning model employed by the IUPforest-L disorder predictor combines the complementary contributions from the autocorrelation function (type 1 feature) and other types of features, so that structural information is extracted with a high degree of prediction accuracy.

The random forest model is an ensemble learning model and is known to be more robust to noise than many non-ensemble learning models. However, as a predictor based on the random forest needs to load many decision trees into memory, it is relatively slow for a forest to predict a single instance at a time. As a result, the current web server of IUPforest-L

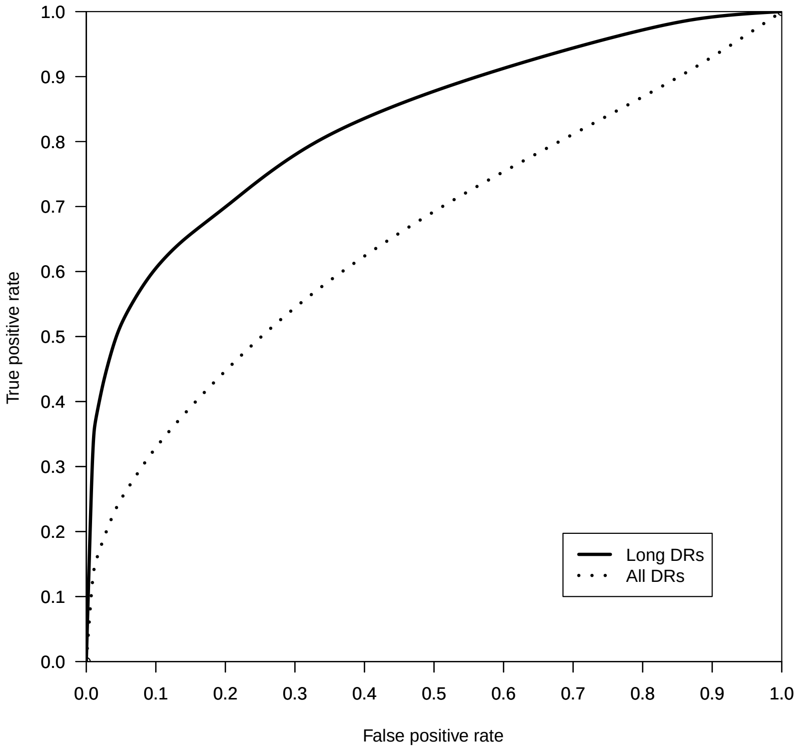


Figure 5.12: ROC curves of IUPforest-L on the blind test dataset of VSL2

is better suited to batch prediction of a large number of protein sequences, which provides a useful alternative tool in large-scale analysis of long DRs in proteomics. As an initial application, we have provided a server, IUPforest-L, for batch protein sequence analysis with the output of an overall summary and details for each sequence. For convenience in proteomic comparisons, the prediction results for 62 eukaryotes linked to the European Bioinformatics Institute are also pre-calculated and can be downloaded from the server.

5.7 Summary

IUP studies are important because DRs are common and functionally important in proteins. The new features, the autocorrelation functions of AAindices within a protein fragment, reflect both residues' contact information and sequence complexity. The random forest model based on this new type of feature and other physicochemical features could effectively detect long DRs in proteins. As a result, a new predictor, IUPforest-L, is developed to predict long

DRs in proteins. Its high accuracy and high efficiency make it a useful tool in large-scale protein sequence analysis.

Chapter 6

Prediction of Short Disordered Regions by Wavelet Transform

Compared to long DRs, short DRs have different AAC and physicochemical properties. Predicting short DRs (< 30 residues) is therefore a different problem in IUP prediction [Romero et al., 1997b].

6.1 Motivation

PSSM and prediction results of secondary structure are widely used in existing short DR predictors [Ward et al., 2004; Cheng et al., 2005b; Peng et al., 2006]. All top five predictors (ISTZORAN, CBRC-DR, Fais, DISOPRED and DISpro) in the CASP7 competition have used PSSM as part of their features. However, a reliable PSSM is generated from protein sequence alignment, which is a time consuming process, so predictors based on PSSM during the training/prediction procedure are not very suitable for large scale analysis of sequences. Many successful secondary structure (SS) predictors and solvent accessibility (SA) predictors use PSSM during prediction and rely on alignment as well. Short DR predictors adopting PSSM based SS or SA prediction results as features can also be less efficient at prediction.

Recently, a predictor of ensemble learning metaPrDOS [Ishida and Kinoshita, 2008] has been developed. It applies results of prediction of seven successful DR predictors including PrDOS, DISOPRED2, DisEMBL, VSL2P, DISpro, IUpred and POODLE-S as features. The SVM model learns from results of these predictors and predicts final order/disorder status. metaPrDOS is more accurate than all seven predictors at the expense of slower speed of prediction. For our test run of the metaPrDOS online service at `http://prdos.hgc.jp/`

`cgi-bin/meta/top.cgi`, depending on the length of query proteins and the condition of each component server, metaPrDOS on average takes from 5 to 10 minutes to predict one protein sequence. Occasionally, some component servers did not reply to a query within an hour, hence the time of prediction of a single sequence may be over one hour [Ishida and Kinoshita, 2008]. So in addition to accuracy, prediction efficiency is an important issue to tackle in short DR prediction.

A feature group which can predict short DRs accurately and avoid the alignment process would be very useful. In this chapter, we propose the wavelet transform (WT) as a new strategy for short DR prediction. The wavelet transform allows researchers to easily study the overall physicochemical property with any window length. However, some existing short DR predictors [Peng et al., 2006; Ishida and Kinoshita, 2007] have to choose an optimal window size after the cross-validation test. The selected window size can be subjective, given that various training datasets are used for different predictors.

To the best of our knowledge, this is the first attempt at applying WT to DR prediction. A protein sequence is first substituted by a selected AAindex and the output is called an AAindex profile. Recall in Chapter 5 the term AAindex refers to a specific amino acid index. When we talk about the AAindex database, we explicitly say so in the text. After WTs of AAindex profiles, if DRs in the sequences correlate with wave crests or troughs of a wave, these DRs will be uncovered. This novel approach is more straightforward than most short DR predictors as machine learning models are not involved. Our cross-validation results show that a single WT can achieve an AUC value of 74.1% in predicting DRs in internal regions of sequences. This result is comparable to a SVM model based short DR predictor trained by AAC (an AUC value of 74%).

WT provides the opportunity for predicting DRs at different levels of detail (scales). In this chapter, we use the term “WT scale (s)” to refer to WTs at different scales. A predictor that can consider different WT scales should achieve higher accuracy of prediction than a predictor based on a single WT. We combine a SVM machine learning model with WTs over different WT scales, AAindices and wavelets. This predictor, called IUPwavelet, gives more accurate results than the application of a single WT in short DR prediction. Our cross-validation test results show that the AUC values of internal regions and complete sequences can reach 79.1% and 85.5% respectively. To compare this with other features used in short DR prediction, we built predictors trained from different feature groups including AACs, physicochemical properties, normalised AAindex score [Han et al., 2009a], prediction results of secondary structure and solvent accessibility and PSSM. The performance of WT of AAindices is superior to all non-alignment based feature groups in short DR prediction. It is only less accurate than the alignment based feature group of PSSM. After taking AAC and terminal status (the percentage of space in windows) into consideration, IUPwavelet achieves an AUC value of 80.4% on CASP7 targets and is only less accurate than the top two server predictors in the competition. IUPwavelet still has the advantage of fast feature generation and it avoids the alignment process. Our time efficiency experiments show that IUPwavelet is at least 50 times faster than many existing predictors including DISpro, DISOPRED2 and VSL2.

The predictor IUPwavelet-II is proposed by combining WTs with alignment based features. Cross validation test results show that the AUC value of IUPwavelet-II can reach 88.7%. The blind test result on CASP7 is 84.6%, higher than all server predictors in the CASP7 competition.

In Section 6.2, our initial attempt at short DR prediction without alignment is described. Then we describe the background and theory of WT in Section 6.3. In Section 6.4, we apply a single WT for short DR prediction. The correlation between WTs and the order/disorder status of protein sequences is measured. In Section 6.5, we propose a three-stage learning model based on WTs and the SVM machine learning model. Greedy feature selection has been applied to find a suitable combination of WTs, which are used to build our short DR predictor IUPwavelet. In Section 6.6, selected WTs are combined with some alignment based feature groups to build predictor IUPwavelet-II, which achieves a higher accuracy of prediction. The results of cross-validation tests are reported at the end of this section. In Section 6.7, we discuss related work that applies WT to bioinformatics, especially secondary structure prediction problem. Finally, we summarise our work in Section 6.8.

6.2 An Initial Attempt

The DRaai-S predictor [Han et al., 2009a] is our initial attempt to predict short DRs without alignment. DRaai-S is based on AAindices and a random forest model. All sequences in the database DisProt3.6 are used to train DRaai-S. Each amino acid sequence in the training set is replaced with numerical sequences by the five sets of AAindices (selected from Table 5.2 and 5.3 with a correlation coefficient of less than 0.8 among them) and smoothed using the Savitzky-Golay filter with a window of 17 amino acids. Then the smoothed vectors are directly used as input parameters to develop the DRaai-S model. DRaai-S is used with a setting of ten trees and a smoothing window of 17 amino acids. Each residue in the sequences

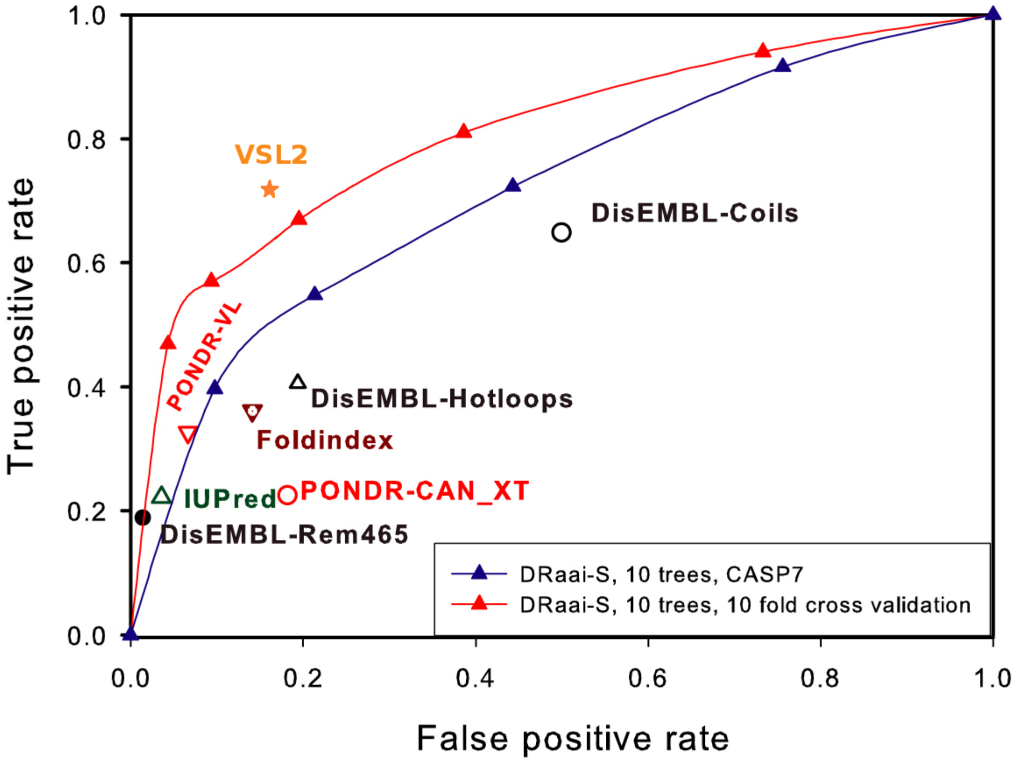


Figure 6.1: Ten-fold cross-validation and blind tests of DRaai-S.

corresponds to five AAindex features. To predict the disorder of a query sequence, the sequence is transformed similarly to five smoothed vectors, which are input to the DRaai-S model to predict the disorder/order status of each residue.

Figure 6.1 shows the ROC curves of DRaai-S under a ten-fold cross-validation test on the DisProt sequences and a blind test on CASP7 targets. The area under the ROC of DRaai-S for the ten-fold cross-validation test is 81.2%, while it is 72.2% when used to predict the CASP7 targets. All independent points in the figure are results on CASP7 targets obtained from the respective online predictors with their default settings. Table 6.1 describes the performance of DRaai-S on CASP7 in comparison with other predictors. The results in Figure 6.1 and Table 6.1 demonstrate that DRaai-S can achieve comparable or more accurate prediction than some published algorithms.

In summary, by using the simple AAindex information, DRaai-S has shown better performance than many well developed published algorithms. DRaai-S has the potential to be

Table 6.1: Performance of DRaai-S and other predictors on the CASP7 targets

Measure(%)	Sensitivity	Specificity	MCC	Spro	Sw
DRaai-S	0.55	0.79	0.19	0.43	0.34
DisEMBL(Coil)	0.65	0.50	0.07	0.33	0.15
DisEMBL(Rem465)	0.19	0.99	0.27	0.19	0.18
DisEMBL(Hot Loop)	0.41	0.81	0.12	0.33	0.21
FoldIndex	0.36	0.86	0.14	0.31	0.22
IUPred	0.22	0.96	0.21	0.21	0.19
PONDR(CANXT)	0.23	0.82	0.03	0.18	0.05
PONDR(VL)	0.33	0.93	0.23	0.30	0.26
PONDR(VLXT)	0.46	0.79	0.14	0.36	0.25
PONDR(XL)	0.30	0.72	0.01	0.22	0.02
VSL2	0.73	0.85	0.33	0.61	0.58

further improved by adjusting the sets of AAindices and the method to transform AAindices. Compared to top ranked server predictors in the CASP7 competition with AUC values around 83%, the prediction accuracy of DRaai-S is low.

6.3 Wavelet Transform

In this section we will initially introduce the background knowledge of the WT and then the theory of WT.

The WT originated in the mid-eighties and has been applied to many types of signal processing and compression problems in communication systems, biomedical imaging, radar, air acoustics, theoretical mathematics, control systems and other areas [Young, 1992; Stark, 2005]. By definition, “wavelets are functions that can be used to efficiently describe a signal by breaking it down into its components at different WT scales (or frequency bands) and following their evolution in the time domain” [Liò and Vannucci, 2000].

The main idea of WT is to represent a signal as a superposition of a set of such wavelets or basis functions. This results in wavelets being local both in frequency/scale and time, a property which is not shared by other families of functions, such as the Fourier basis. For a simplified example, WT is like a tuning fork physically resonating with sound waves of its specific tuning frequency. So given an unknown signal containing the frequency of disorder, WTs which resonate with disorder will be able to detect the position of DRs.

6.3.1 Wavelet Transform Theory

Wavelet analysis uses basis functions localised in time and frequency to represent non-stationary signals. The idea of having a basis function is that any signal can be described as a weighted sum of a family of functions. These functions are the basis functions of the transform. Transforms are often just changes from one basis function to another, although other transform types exist.

Different from sine waves, a wavelet tends to be irregular and asymmetric. The definition of a continuous WT (CWT) is:

$$CWT_x^{\psi}(\tau, s) = \int x(t) \cdot \psi_{\tau,s}^{*}(t) dt \tag{6.1}$$

The CWT function can be thought of as the inner product of the test signal $x(t)$ (t is time) with the basis functions $\psi_{\tau,s}(t)$ (also called wavelets). The range of CWT outputs is around 0 and not limited to [-1...1]. The $*$ denotes complex conjugation[1]. Wavelets are generated from a single basic $\psi(t)$ (the so called mother wavelet) by scaling and translation:

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{s}} \psi \left(\frac{t-\tau}{s} \right) \tag{6.2}$$

The variable s is a WT scale, a positive value. A high WT scale corresponds to low frequency while a low scale corresponds to high frequency. Therefore the smaller the WT scale s, the more "compressed" the wavelet. Scaling a wavelet simply means stretching (or compressing) it. The variable τ represents translation, which indicates where the mother wavelet is located. The factor $\frac{1}{\sqrt{s}}$ is for energy normalisation across different WT scales. So the transformed signal will have the same energy at every WT scale.

This definition of CWT shows that the wavelet analysis is a measure of similarity between a wavelet and the signal itself. Here the similarity is in the sense of similar frequency content. The calculated CWT refers to the closeness of the signal to the wavelet at the current WT scale.

This clarifies the correlation of a signal with a wavelet at a certain WT scale. If the signal has a major component of the frequency corresponding to the current WT scale, then the wavelet at the current WT scale will be similar or close to the signal at the particular location where this frequency component occurs. Therefore, the CWT computed at this point in the

[1] A complex conjugation is most simply defined as one complex number having the same real and imaginary parts to another complex number, but with imaginary part different in sign.

time-scale plane will be a relatively large number.

A major strength of wavelet analysis [Goswami and Chan, 1999] is its capability to represent signals at several levels of resolution called multiresolution analysis (analysis at different levels of WT scales). WTs are designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. This approach makes sense especially when the signal at hand has high frequency components for short durations and low frequency components for long durations, which is the case in most biological signals. Using the powerful multiresolution analysis, we can represent a signal by a finite sum of components at different resolutions so that each component can be processed adaptively based on the objectives of the application.

There are many families of wavelets including Haar, Mexican Hat, Coiflets, Biorthogonal, reverse Biorthogonal and Gaussian. These families are different from each other by mother wavelets $\psi(t)$. For example the basis function of the simplest Haar wavelet is defined as:

$$\psi(x) = \begin{cases} 1 & : \quad 0 \le x < 0.5 \\ -1 & : \quad 0.5 < x \le 1 \\ 0 & : \quad otherwise. \end{cases}$$

While the basis function of the Mexican Hat wavelet is much more complex: $\psi(x) = \left(\frac{2}{\sqrt{3}}\pi^{-1/4}\right)(1-x^2)e^{-x^2/2}$ and proportional to the second derivative function of the Gaussian probability density function. Different wavelets have different shapes and properties. As shown in Figure 6.2, the Haar wavelet on the left is asymmetric, orthogonal and biorthogonal while the Mexican Hat wavelet on the right is neither orthogonal nor biorthogonal but symmetric.

A very important fact about wavelets is that the transform itself poses no restrictions whatsoever on the form of the mother wavelet. This is an important difference to other transforms. One can thus design wavelets to suit the requirements of the situation.

The AAindex profile for sequence 1h5wA (PDB code) using AAindex BASU050102 is plotted in Figure 6.3. The original AAindex BASU050102 has been transformed with a Z score to generate this profile. For ease of discussion, the curve is scaled to [0...1] and, given that BASU050102 is negatively correlated with DRs, each point in the profile is the 1's complement of the scaled Z score. In Figure 6.3, the fluctuation of the curve seems random and does not relate to the annotation of DRs at the top of the figure.

Figure 6.4 shows the AAindex profile after WT (wavelet rbio2.2 and WT scale 64) for sequence 1h5wA. Peaks of the curve are much more correlated with short DRs than the AAindex profile in Figure 6.3. So after wavelet transformation, the curve is more correlated

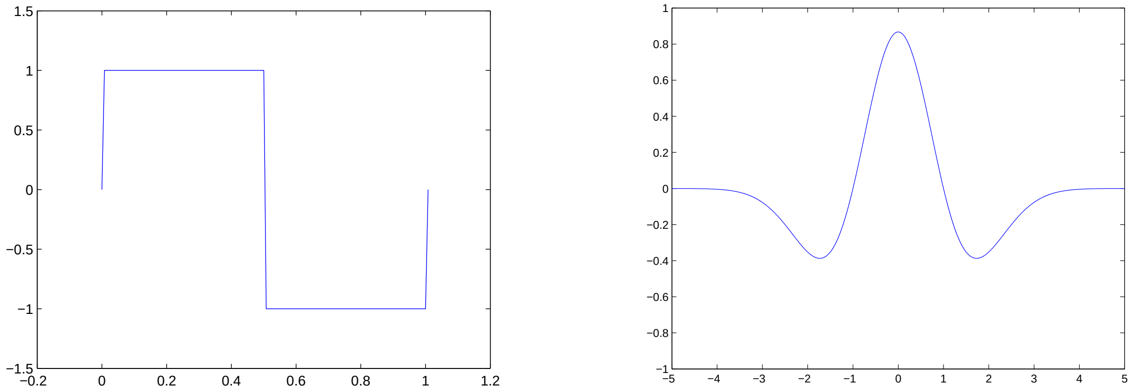


Figure 6.2: The Haar wavelet and Mexican Hat wavelet

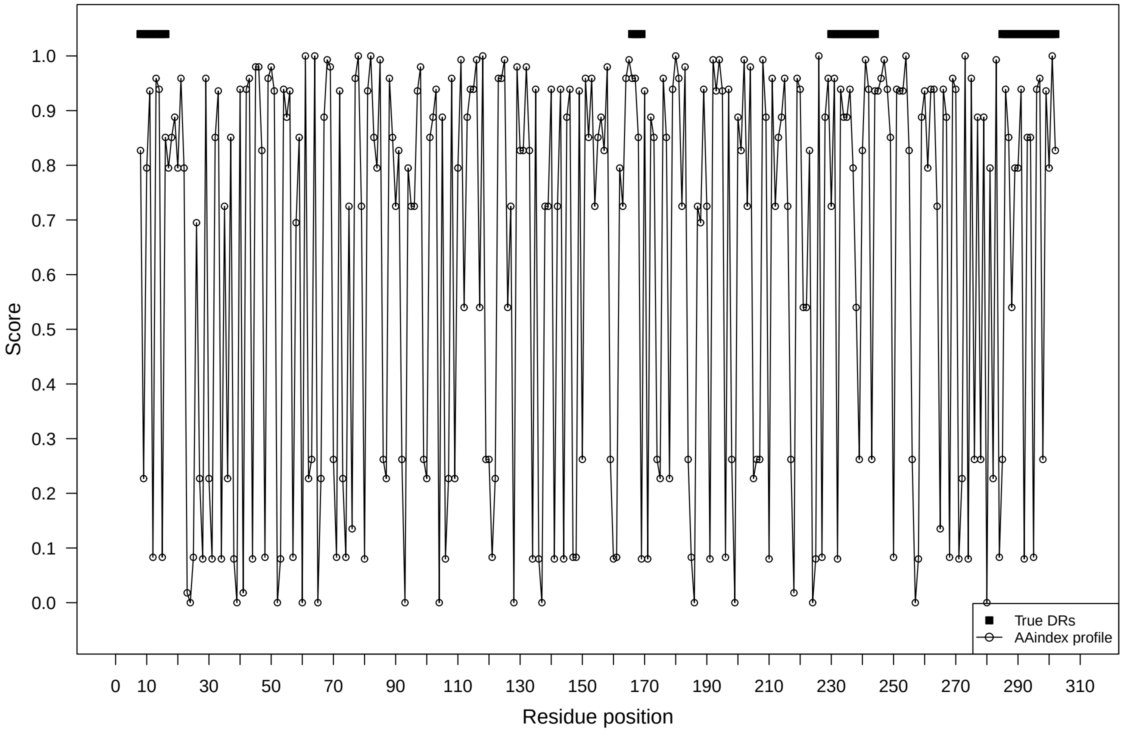


Figure 6.3: AAindex BASU050102 profile of sequence 1h5wA. Black bars indicate DRs.

with the distribution of DRs in the protein sequence. For ease of discussion, from hereafter we define a WT as a transformation of a specific AAindex profile by a specific wavelet function

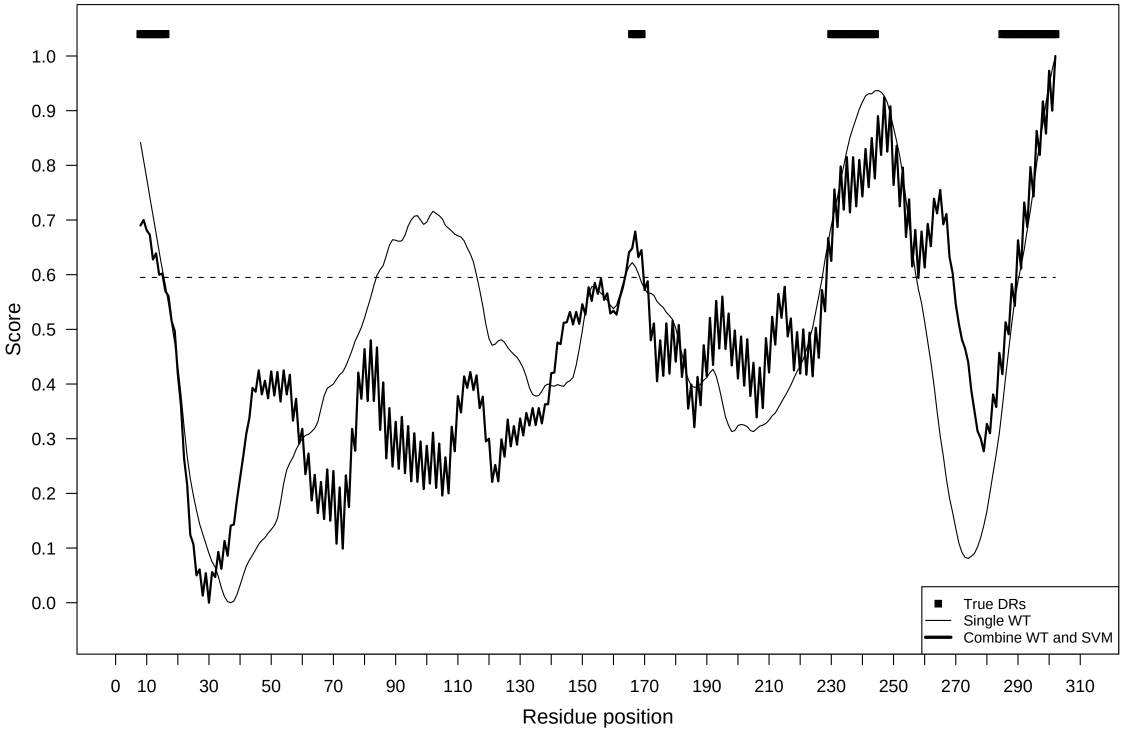


Figure 6.4: AAindex BASU050102 profile of sequence 1h5wA after WTs. The fine-line curve is generated by the wavelet rbio2.2 at WT scale 64. The bold-line curve is built from combination of several wavelet transforms. Black bars indicate DRs.

at a certain WT scale.

6.4 Single Wavelet Transform Based Short DR Prediction

According to Figures 6.3 and 6.4, a single WT may be applied to predict DRs. However, WT selection based on one IUP is not reliable. To find the most appropriate WT for short DR prediction, we have applied WTs to all sequences in the database *SDB*. Recall from Section 3.1.2 that *SDB* is a set of protein chains derived from PDB, and residues are annotated as disordered if corresponding coordinates are missing in PDB files. Protein chains that have DRs equal to or less than three residues are removed. Chains of less than 30 residues in length or that have a resolution coarser than 2.5 Å are filtered out. Homologous chains are also removed and the final dataset contains 13,831 disordered residues (6.45% of total residues).

To determine an appropriate AAindex, wavelet and WT scale, it is important to measure the correlation between WT and the order/disorder status. From the fine-line curve of one WT and the annotation of order/disorder in Figure 6.4, a confusion matrix can be created given a cutoff value between 0 and 1. Residues with scores above a cutoff value are predicted as disordered, otherwise ordered. Assuming a series of cutoff values exist, there are many confusion matrices from which a ROC curve can be drawn. The AUC value is therefore used to measure the correlation between a WT and the order/disorder status. An AUC value above 0.5 represents a WT that is positively correlated to DRs in protein sequences and a value below 0.5 represents a negative correlation.

Given a WT is performed on an AAindex profile, first, we need to select indices in the AAindex database. Tables 5.2 and 5.3 list 20 selected AAindices in the AAindex database that are most correlated (either positively and negatively) with order/disorder. To generate order/disorder related AAindex profiles, here we consider the top ten AAindices from each of two tables and they are re-produced as Table 6.2. PCS, NCS and ACC in this table are abbreviations for positively correlated sequences, negatively correlated sequences and the average correlation coefficient. One protein sequence can be converted by any of these AAindices in Table 6.2.

To analyse a signal's different details, we have used seven different WT scales including 2, 4, 8, 16, 32, 64 and 128 as too large WT scales omit important details in a signal while small scales may distort a signal by noise. From 15 wavelet families shown in Table 6.3, 45 wavelets in Table 6.4 are selected. Wavelets from the same family are separated from each other by numbers after their family name. These numbers represent different orders in mother wavelet function.

Given 20 AAindices, 45 wavelet functions and seven WT scales, there are 6,300 (20×45×7) different WTs. Obviously not all of them can be used for short DR prediction, and we find appropriate WTs according to their AUC values. From all 6,300 WTs, the most highly ranked WTs have applied the AAindex BASU050102. BASU050102 is the interactivity scale obtained by maximising the mean of the correlation coefficient over single-domain globular proteins. Given the AAindex BASU050102, Table 6.5 lists the largest and smallest AUC values along with their corresponding wavelets and WT scales. For fair comparison with features involving sliding windows, our evaluation always ignores seven residues at the N and C termini unless otherwise specified. DRs with three or less residues are excluded from evaluation.

In Table 6.5 all AUC values are less than 0.5, and high scores of the WT are positively

Table 6.2: Selected AAindices positively/negatively related to disorder derived from Tables 5.2 and 5.3

Name of AAindex	#PCS	#NCS	ACC
VINM940102	297	62	0.248
MIYS990102	297	62	0.241
MIYS990101	296	63	0.242
RACS770101	295	64	0.237
MIYS990104	295	64	0.245
MEIH800101	294	65	0.238
MIYS990103	294	65	0.244
OOBM770103	293	66	0.229
PARJ860101	293	66	0.233
BULH740101	292	67	0.223
BASU050101	302	57	-0.244
ZHOH040101	302	57	-0.243
CIDH920103	300	59	-0.219
NOZY710101	300	59	-0.232
PONP930101	300	59	-0.239
BASU050102	299	60	-0.255
BASU050103	299	60	-0.241
CIDH920104	297	62	-0.234
CIDH920105	297	62	-0.237
MANP780101	297	62	-0.235

Table 6.3: 15 wavelet families

Short name	Family name
haar	Haar wavelet
db	Daubechies wavelets
sym	Symlets
coif	Coiflets
bior	Biorthogonal wavelets
rbio	Reverse biorthogonal wavelets
meyr	Meyer wavelet
dmey	Discrete approximation of Meyer wavelet
gaus	Gaussian wavelets
mexh	Mexican hat wavelet
morl	Morlet wavelet
cgau	Complex Gaussian wavelets
shan	Shannon wavelets
fbsp	Frequency B-Spline wavelets
cmor	Complex Morlet wavelets

Table 6.4: 45 selected wavelets

bior1.1	bior1.5	bior2.6	bior3.3	bior4.4
cgau1	cgau3	cgau5	cmor1-1.5	cmor1-0.5
cmor1-0.1	coif1	coif3	coif5	db2
db6	db10	db25	db45	fbsp1-1-1.5
fbsp1-1-0.5	fbsp2-1-1	fbsp2-1-0.1	gaus2	gaus4
gaus6	gaus8	rbio1.3	rbio2.2	rbio2.8
rbio3.5	rbio5.5	shan1-1.5	shan1-0.5	shan1-0.1
shan2-3	sym2	sym6	sym10	sym20
Haar	morl	mexh	meyr	dmey

Table 6.5: Best performing AUC results from AAindex BASU050102

Wavelet	WT Scale	AUC
rbio2.2	64	0.259
coif1	128	0.265
coif1	64	0.268
mexh	32	0.275
rbio2.2	128	0.275
gaus2	32	0.276
rbio2.2	32	0.279
mexh	16	0.279

correlated to structured regions in protein sequences and are negatively correlated with DRs. With the 1-AUC values in Table 6.5, wavelet families mexh, rbio, gaus and coif achieve AUC values over 72%. The accuracy of prediction can reach 74.1% with the wavelet rbio2.2 and a WT scale of 64. The wavelet rbio2.2 is shown in Figure 6.5. This wavelet belongs to the family of reverse biorthogonal wavelets and is symmetric, but not orthogonal.

Even though a machine learning process is not involved, the AUC result of this single WT is already reasonably high. For comparison purposes, we built a SVM based short DR predictor trained on AACs and its AUC value is 74%.

Figure 6.6 compares the accuracy of short DR prediction by using a AAindex profile and a single WT based on that AAindex profile. All 20 AAindices in Table 6.2 have been applied and the wavelet adopted is rbio2.2. The first grey bar demonstrates that when we substitute all sequences in dataset *SDB* to the physicochemical property of AAindex VINM940102, the AUC value derived is 57%. This AUC value increases to 72% (the first black bar) after WT (rbio2.2) is applied.

In general, Figure 6.6 shows that when AAindex profiles are applied for prediction, AUC values are always between 50% and 60% for all 20 AAindices. It is not surprising, as AAindex profiles are just a substitution of amino acids to their corresponding physicochemical properties. The direct application of physicochemical properties cannot help the prediction

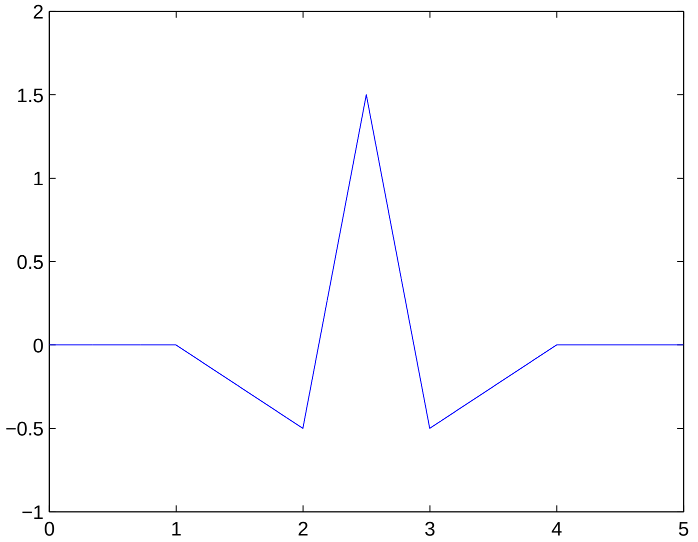


Figure 6.5: The rbio2.2 wavelet

of short DRs. However, after WTs with appropriate WT scales are applied, AUC values increase dramatically to around 70% over all 20 AAindices. The improvement in the AUC is more than 10% for all AAindices. It is also interesting that those AUC values of AAindex profiles from all 20 AAindices are quite similar. It is due to the correlation among these AAindices.

It is clear that the single WT can be used for short DR prediction. With an appropriate AAindex, wavelet and WT scale, the accuracy of prediction is comparable with some machine learning based short DR predictors. In cross-validation tests, the AUC value can reach 74.1%.

6.5 Combining Wavelet Transforms and SVM for Efficient Short DR Prediction

Results of the experiments in the previous section are encouraging; however they also reveal that short DR prediction is a complicated problem. A single WT might not provide enough information to build an accurate predictor. However, a combination of multiple WTs may provide comprehensive information regarding DRs in protein sequences. To build an accurate short DR predictor, our model comprises three major stages as shown in Figure 6.7.

At stage one, a WT is applied to each sequence profile in our training database *SDB*. From 6,300 WTs that have been considered potentially related to short DRs, the combination

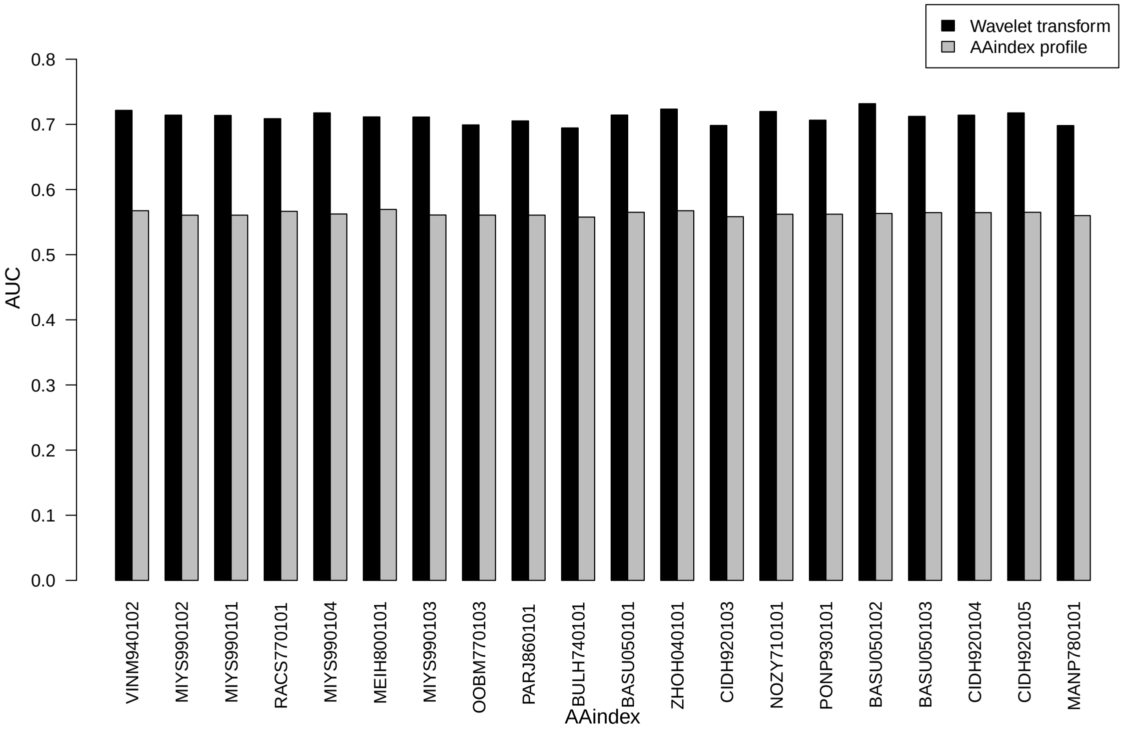


Figure 6.6: Comparison between AAindex profile and the best performing WT of the profile.

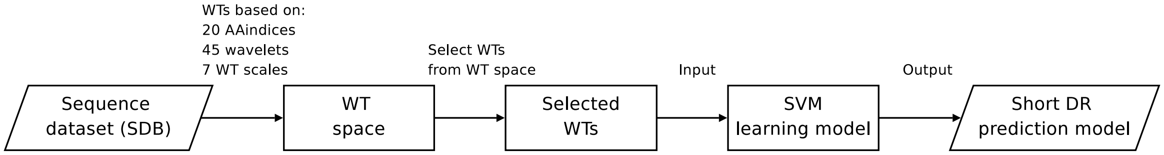


Figure 6.7: Our three-stage learning model

of WTs that can produce the most accurate DR prediction should be selected. However, the selection of optimal WTs requires an exhaustive search of all possible subsets of 6,300 WTs, which is computationally expensive. Therefore, at stage two, we propose a greedy search based WT selection process as described in Section 6.5.1. In the last stage, the WTs selected are input to the SVM learning model.

6.5.1 Greedy Selection of Wavelet Transforms

The greedy algorithm of forward/backward feature selection adds one feature to (or deletes one feature from) the current feature space if it brings the most extra information. In our scenario, features are WTs and the extra information is measured by the improvement of the AUC value from a ten-fold cross-validation test. The SVM machine learning model is used to build the predictor by learning from multiple WTs. Once a feature is added into (or deleted from) the current feature space, all features in the current feature space are used to rebuild the predictor. It is an iterative process until any further addition (or deletion) does not increase the accuracy of the predictor. However, as introduced before, due to the large WT space (6,300 WTs), applying forward/backward feature selection directly is too time consuming for us.

Here we propose a greedy search based WT selection approach by assuming that the AAindex, wavelet and WT scale are three independent dimensions. Therefore all 6,300 WTs can be cast to a three-dimensional space. To search for the optimal WT combination (a combination of WTs that achieves the highest AUC value), the classic greedy algorithm of forward feature selection needs to do a stepwise test from 1 to 6,300. Our greedy selection algorithm of WT has simplified the goal of finding the optimal WT combination and aims to explore the small three dimensional space that contains many optimal WTs. It is still a greedy algorithm because forward/backward WT selection is carried out in each dimension given the other two dimensions are fixed. In each step, instead of one WT, a set of WTs of one dimension is added to (or deleted from) the current WT subset if it brings the largest improvement to the AUC. When the tuning of one dimension produces the highest AUC value, this dimension is fixed and we start to tune another dimension from the two previously fixed dimensions. Our greedy search based WT selection approach ends when all three dimensions have been fixed and the small three dimensional space is determined.

For ease of explanation, Figure 6.8 illustrates the space we try to find in a sample two dimensional space. The X axis of Figure 6.8 is the WT scale and the Y axis represents the AAindex. A complete set of optimal WTs are represented by circles. Figure 6.8 (I) demonstrates the start of forward WT selection along the AAindex dimension. WTs under consideration are represented by a dashed line and when the AAindex is 5, the number of WTs is seven. These seven WTs are used to build a SVM based DR predictor and the AUC from a ten-fold cross-validation test can then be calculated. Then AAindex 5 is combined with the other 19 AAindices one by one and we find the AUC values of 19 paired AAindices.

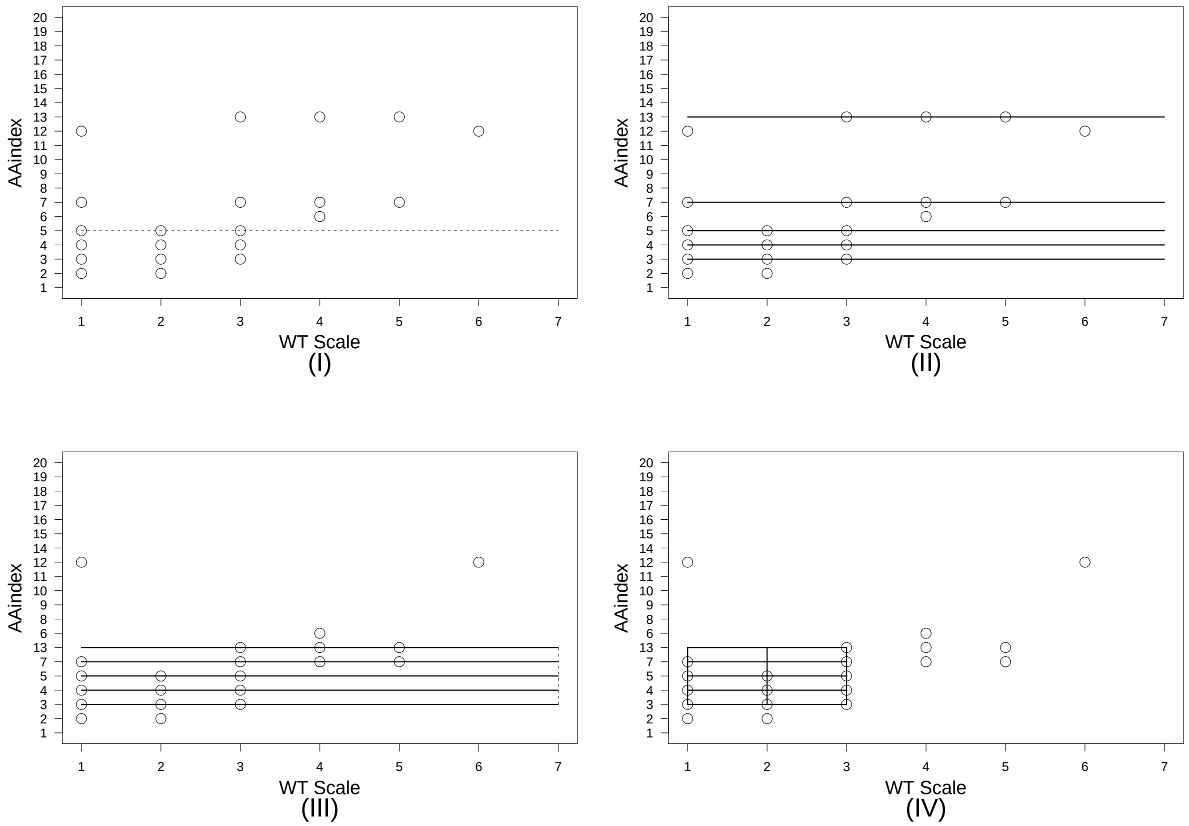


Figure 6.8: Illustration of greedy WT selection in a sample two dimensional space

If the best performing pair of 14 WTs (2×7) achieves an AUC value larger than that from AAindex 5, it is included in the current feature space and is used for forward WT selection during the next round. This is an iterative process until the addition of a new AAindex can not improve the accuracy of prediction further. In Figure 6.8 (II), the bold horizontal lines demonstrate that the combination of AAindices 3, 4, 5, 7 and 13 are selected as the best combination on the AAindex dimension. In Figure 6.8 (III), the backward feature selection is applied to WT scale dimension. The order of the AAindices (Y axis) has been changed to bring the selected five AAindices together. First, the value seven in WT scale dimension is removed and then we calculate the AUC result of the six WT scales left. Then the value six in the WT scale dimension is removed and all the other five values are kept to calculate the AUC value. At the end of this round, we will remove one WT scale value, whose deletion leads to the biggest increase in the prediction accuracy. The deletion in this dimension will terminate if the removal of any WT scale value leads to a decrease in the prediction accuracy. As demonstrated in Figure 6.8 (IV), the final rectangle selected should include many important WTs of the optimal WT combination.

Selected Wavelet Transforms After Greedy Selection

Determining the starting point of each of the three dimensions is the initial step in our greedy WT selection. According to Table 6.2, BASU050102 is the AAindex of the highest absolute value in terms of average correlation coefficient with order/disorder status in the AAindex database. In Figure 6.6, the WT based on AAindex BASU050102 achieves the highest AUC value from among all 20 AAindices. AAindex BASU050102 is therefore set as the start point in the AAindex dimension. According to Table 6.5, the WT based on AAindex BASU050102 and wavelet rbio2.2 achieves an AUC value of 74.1% (1−25.9%), which is higher than AUC values generated from other wavelets. The wavelet rbio2.2 is set as the start point in the wavelet dimension. So we determine that the AAindex BASU050102, all seven WT scales and the wavelet rbio2.2 are the start point for three dimensions in WT selection. The reason that all of the values in the WT scale dimension are applied is that this dimension contains the least number of values in three dimensions and the calculation overhead is lower even if all values are included.

In the first step of greedy selection of the WT, we fix the WT scale and wavelet dimensions and carry out greedy selection in the AAindex dimension. The increment of the AAindex leads to higher accuracy of prediction when the number of the AAindex is less than 12. The highest AUC value of 78.3% is achieved when the number of the AAindex is 11. These AAindices include VINM940102, MIYS990102, RACS770101, MIYS990104, MIYS990103, PARJ860101, BULH740101, BASU050101, ZHOH040101, NOZY710101 and BASU050102. The greedy selection for the AAindex dimension is completed.

In step two, the backward WT selection is carried out to find the best WT scale combination from seven scale values along the scale dimension. The main point is to remove redundant information because all values in this dimension are involved. According to experimental results after removing WT scale two and four, the accuracy of prediction is almost kept the same and the AUC value is 78.4%. So selected WT scales include 8, 16, 32, 64 and 128.

In the last step, we select the best combination of wavelets. The AAindex and WT scale dimensions have already been fixed. The results demonstrate that when the number of wavelets was increased to four, we achieved the best result of prediction with an AUC value

Table 6.6: AUC results of the best performing N WTs

N	20	50	100	150	200
AUC	77.01%	77.38%	77.64%	77.84%	77.76%

of 79.1%. These wavelets include rbio2.2, sym20, mexh and dmey.

Our final selected WT space is defined by 11 AAindices, 5 WT scales and 4 wavelet functions. The number of WTs in this space is $11 \times 5 \times 4 = 220$. We call the predictor trained from these selected WTs the IUPwavelet. The bold-line curve of Figure 6.4 corresponds to a blind test result on sequence 1h5wA by the IUPwavelet. Compared to the thin curve of the single WT, the application of multiple WTs in short DR prediction is more successful. That is, the peaks that cross the dotted line are more related to DRs.

The Best N WT Combination

The WTs selected using our greedy approach achieved an AUC value of 79.1%. We also apply the best N approach which selects the best performing N WTs from the 6,300 WT space according to their AUC values. The top N WTs with the highest AUC values are selected and fed into the SVM learning model for training and prediction purposes. With a linear kernel of SVM, ten-fold cross-validation results are shown in Table 6.6. When the number of WTs is 150, the predictor achieves the highest AUC value of 77.8%. However, this AUC value is lower than that of the predictor built from the WTs selected from the greedy selection strategy.

6.5.2 Other Feature Groups

There are many features proposed for short DR prediction. To compare the prediction accuracy between our selected WTs and other feature groups, we list five non-alignment based feature groups and four alignment based feature groups as follows:

- Group 1. Amino acid composition based features
- Group 2. AAindex based features, normalised Moreau-Broto autocorrelation
- Group 3. AAindex based features, average physicochemical property
- Group 4. AAindex based features, AAindex smoothed by the Savitzky-Golay filter

- Group 5. Physicochemical properties in a sliding window
- Group 6. Wavelet transform of the AAindex
- Group 7. Prediction results of Secondary Structure (SS) and Solvent Accessibility (SA)
- Group 8. PSSM features derived without a window
- Group 9. PSSM features derived from the average value within window
- Group 10. PSSM features derived from the concatenated PSSM values in a window

The first five feature groups have been used or introduced in previous chapters. The first feature group has been introduced in Chapter 3. Feature Groups 2, 3 and 4 are all AAindex-based features. Groups 2 and 3 are part of the training features of IUPforest-L as introduced in Chapter 5. Feature Group 4 is the smoothed AAindex scores using the Savitzky-Golay filter which has been described in Section 6.2. Feature Group 5 includes seven physicochemical properties: mean hydrophobicity, hydrophobic cluster, mean net charge, charge cluster, Shannon entropy, average number of contact and foldIndex calculated in each sliding window. These properties have also been used in IUPforest-L in Chapter 5. The sixth feature group includes all 220 WTs selected from the three dimensions. All the final four features are alignment based, which include three groups calculated from the PSSM.

Group 2 is calculated from the AAindex and uses normalised Moreau-Broto autocorrelation descriptors in a sliding window. We have applied a small number of AAindices. In Tables 5.2 and 5.3, the correlation among the top 40 disorder related AAindices is very high (above 0.8). To alleviate learning from repeating information, four features with a correlation of less than 0.8 are chosen from the top 40 AAindices and these are shown in Table 6.7. The d value of the normalised Moreau-Broto autocorrelation of Equation 5.3 is set to five, as the size of window for short DR prediction is generally small. We calculated 20 (4 AAindices $\times$ $5d$) autocorrelation values for the central residue of each window.

Group 3 is calculated from the AAindex average physicochemical property in a sliding window. We used the average physicochemical property in our long DR predictor IUPforest-L in Chapter 5. In this chapter, the 20 most highly ranked ordered and disordered AAindices shown in Table 6.2 are applied to calculate the average physicochemical property in a sliding window. Each window is represented by 20 averaged physicochemical properties.

Group 4 is calculated from the smoothed AAindex profiles using the Savitzky-Golay filter. The Savitzky-Golay filter has been used to clean the data of high frequency noise. We first

Table 6.7: Selected AAindices of feature Group 2

Name of AAindex	Positively correlated	Negatively correlated	AVG correlation coefficient	AAindex description
VINM940102	297	62	0.248	Normalised flexibility parameters (B-values) for each residue surrounded by none rigid neighbours
BULH740101	292	67	0.223	Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues
PUNT030102	290	69	0.212	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases
CHOP780203	289	70	0.208	Normalised frequency of beta-turn

selected one normalised AAindex to substitute the protein sequence and output the profile. The Savitzky-Golay filter is then applied to smooth the profile. This filter can optimise the original profile and make it more related to the distribution of DRs [Han et al., 2009a]. We have used all 20 AAindices in Table 6.2.

Group 7 consists of the prediction results of secondary structure and solvent accessibility. We have used the prediction results of three successful secondary structure predictors including PSIPRED [Bryson et al., 2005] (version 2.4), PHD [Rost and Sander, 1993a] (version 0.1) and SSpro [Cheng et al., 2005a] (version 4.1) for our DR prediction. PSIPRED and SSpro need the PSSM information to support their prediction. In our experiments, PSIPRED has applied the PSSMs generated by aligning against the UniRef100 database. The SSpro predictor has applied the PSSMs generated by aligning against the nr database. PHD does not need PSSM for secondary structure prediction. Prediction results of each of the three secondary structure predictors are presented as three binary values corresponding to helices, sheets and coils. The solvent accessibility predictor ACCpro4.1 [Cheng et al., 2005a] is also used and the result of prediction for each residue is presented in terms of two binary values.

Groups 8, 9 and 10 are PSSM features. The alignment database we applied is UniRef100. We measure three feature groups derived from PSSMs. The first group (Group 8) uses PSSM scores directly and each residue corresponds to 20 PSSM scores along with two information values. The second feature group (Group 9) averages PSSM values over the input window. Given a window size of N, a window corresponds to a $N \times 22$ matrix. The average score is calculated by averaging over each row in the matrix, which results in a 1×22 matrix. These 22 values are features of the middle residue in the window. This approach has the advantage of the introduction of local information in the PSSMs while maintaining a small number of training features. The last PSSM feature group (Group 10) adopts a window concatenation approach. All values in the $N \times 22$ matrix are considered as features of the middle residue in the window. Therefore this feature group includes all local PSSM information for the target residue.

6.5.3 Results

In this section, we do a series of experiments based on our training dataset *SDB* and CASP7. As introduced before, DRs of three or less residues are considered unreliable and are ignored. Meanwhile, window based features will encounter difficulty in calculating the features for residues at the N and C termini. Therefore, we have ignored (window size-1)/2 resides at the N and C termini in both training and prediction procedures to assure a fair comparison of all feature groups.

Results comparing different feature groups are introduced first. The ten-fold cross-validation test result of our WT-based short DR predictor which considers terminal residues is then presented. Finally, a blind test is carried out on CASP7 targets and the results are compared with other best performing DR predictors.

Comparison with Different Feature Groups

In Table 6.8, we compare the AUC achieved by the predictor based on our selected WTs and predictors based on other feature groups introduced in Section 6.5.2. For each feature group, the bold font annotates the highest AUC values. Given a window size of 15, our short DR predictor IUPwavelet achieves an AUC result of 79.1% for the middle area of protein sequences, which is 5.1% higher than that of AAC. This improvement over the AAC is greater than that achieved by VSL2 of Peng et al. [2006]. They measured the performance of predictors by AAC and a combination of AAC, PSSM and PHD (results of the secondary structure prediction). With the addition of PSSM and PHD, the improvement in accuracy over AAC is 3%. Table 6.8 clarifies several interesting points.

- Small windows generally are more suitable than large windows for short DR prediction. Nine out of ten groups achieve their best performance when the window size is either 7 or 15. The only exception is feature Group 4, whose best result is achieved with a window of 31 residues.

- The WT is the most suitable feature in all non-alignment based feature groups in short

Table 6.8: Comparison of the other feature groups. If the size is 7, then 3 residues at the N/C terminal are not measured.

Window	Feature group									
size	1	2	3	4	5	6 (IUPwavelet)	7	8	9	10
7	71.4%	66.9%	71.5%	65.7%	69.1%	**79.6%**	**69.0%**	**72.7%**	80.7%	80.3%
15	**74.0%**	**70.8%**	**74.3%**	70.8%	**72.7%**	79.1%	63.4%	71.0%	**81.0%**	**81.4%**
23	73.2%	70.2%	73.4%	73.4%	71.6%	79.0%	65.6%	69.9%	78.4%	80.8%
31	72.1%	68.4%	72.1%	**74.0%**	70.2%	78.8%	68.1%	69.3%	76.2%	80.3%
39	71.6%	66.7%	71.5%	73.7%	69.3%	78.9%	68.1%	68.7%	74.3%	79.9%

DR prediction. It also achieves higher accuracy than the predictor trained by the prediction results of secondary structure and solvent accessibility. The WT is only less accurate than the last two feature groups based on PSSM.

- The first PSSM feature group (feature Group 8), which does not apply any local information, performs significantly worse than the WT. When local information is considered, feature Groups 9 and 10 achieved the best AUC values of 81% and 81.4% when the window size was 15. These results are better than all other feature groups. For feature Group 9, when the window size was higher than 15, the accuracy of prediction dropped dramatically and the AUC value can be lower than that from IUPwavelet.

Figure 6.9 shows the ROC curves of ten-fold cross-validation in the training dataset with different feature groups with a window size of 7. The AUC values of the WT and two PSSM based feature groups are substantially larger than the other seven groups (P-value $< 10^{-8}$).

Considering the high percentage of disordered residues at N and C termini, we have taken them into consideration in the ten-fold cross-validation test with the results shown in Figure 6.10. The test reveals that IUPwavelet achieves an AUC value of 85.5%, around 2.3% lower than the predictor DISpro, the second most accurate prediction server in the CASP7 competition. Because we used the same training dataset, ten-fold cross-validation results are comparable. The corresponding P-value is less than 10^{-3}, which indicates that DISpro is statistically more accurate than our IUPwavelet in this dataset. We believe this could be due to the different machine learning models applied by DISpro. It may also be attributed to features of PSSMs and to the prediction results of the secondary structure and solvent accessibility used in DISpro.

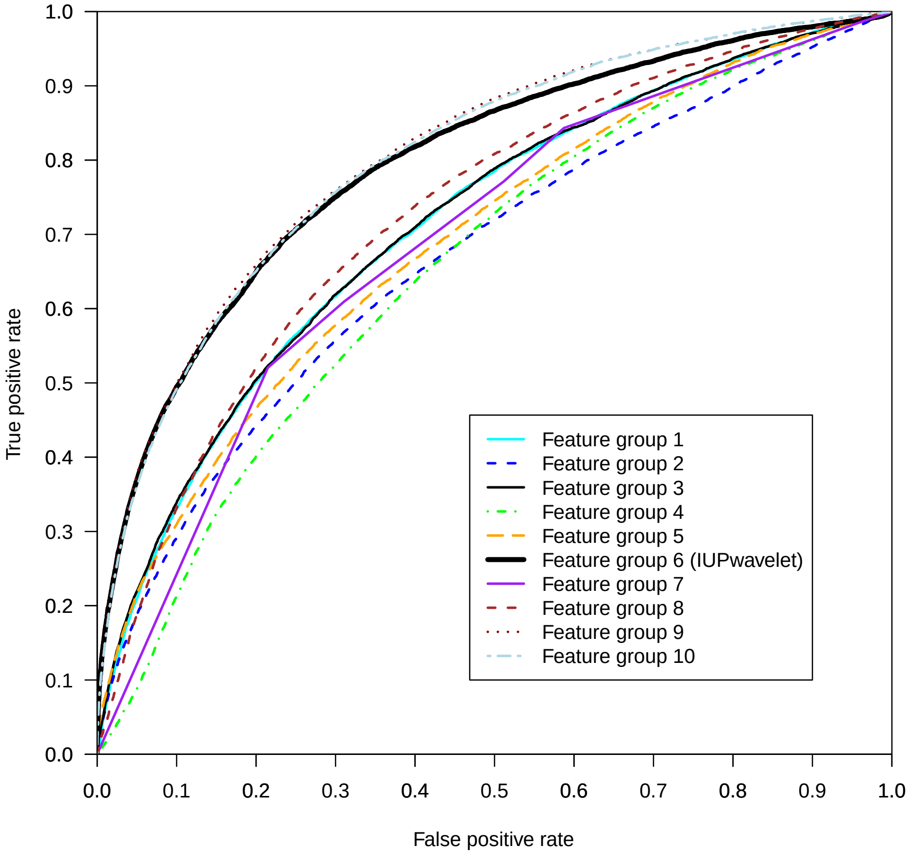


Figure 6.9: Ten-fold cross-validation test on the database SDB *of different feature groups under window size seven*

Blind Test on CASP7

Results of blind tests on CASP7 are shown in Table 6.9 (N and C terminal residues are measured). In this experiment, we have included the AAC feature group and terminal status into our feature space of the IUPwavelet. Terminal status is also called space status which is a simple float value annotating the percentage of space in the terminal window. We start to count the space when the number of residues in the current window is less than the defined window size. Given a window size of 15, seven residues at the N/C terminal have positive terminal status and the closer they are to the termini the larger this value is. Since both AAC and terminal status can be easily calculated from the query sequence itself, they have a negligible influence on our speed of prediction.

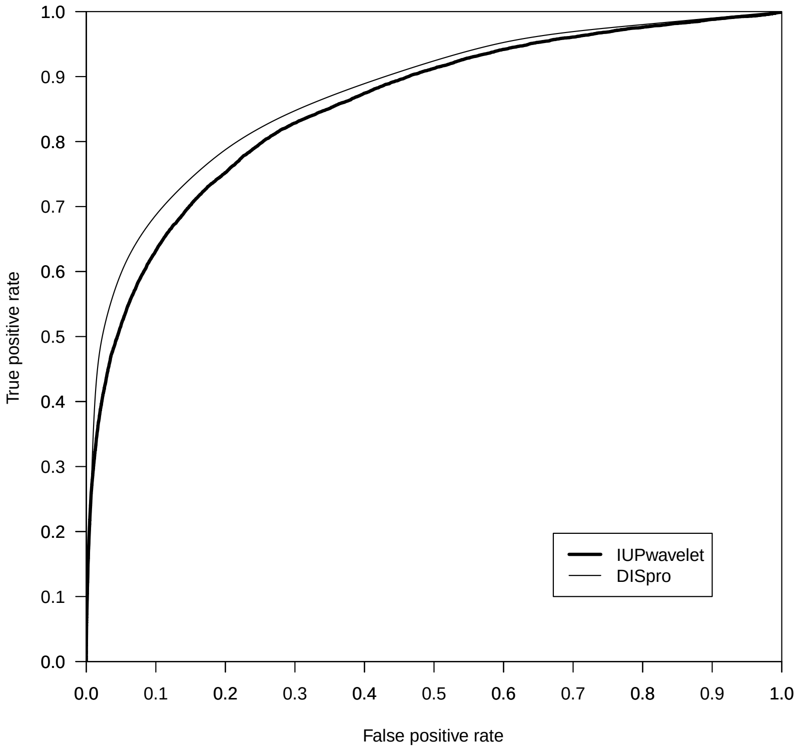


Figure 6.10: Comparison of ten-fold cross-validation test results with DISpro on the database SDB

Table 6.9 illustrates that our predictor IUPwavelet is ranked six among all 29 predictors in the CASP7 competition. However, all top three predictors ISTZORAN [Peng et al., 2006], CBRC-DR [Shimizu et al., 2005; Hirose et al., 2007; Shimizu et al., 2007b] and Fais [Ishida et al., 2006] are registered as human experts. Most of the top 10 predictors have applied alignment based features[2] [Bordoli et al., 2007; Peng et al., 2006; Cheng et al., 2005b; Ward et al., 2004]. In comparison, our approach is much simpler and faster.

6.5.4 Time Efficiency

We have discussed previously that applying the WT for short DR prediction is more time efficient than that of multiple alignment based approaches. In this section, we compare the speed of the prediction between IUPwavelet and some top ranked predictors of the CASP7 competition (VSL2, DISOPRED2 and DISpro). Although ISTZORAN is registered as a

[2] `http://predictioncenter.org/casp7/meeting_docs/abstractsd.pdf`

Table 6.9: Comparison with other predictors on CASP7

Groups	Specificity	Sensitivity	S_w	ACC	ROC(AUC)
590	0.837	0.725	0.562	0.781	0.860
253	0.966	0.454	0.420	0.710	0.850
443	0.924	0.556	0.481	0.740	0.844
470	0.953	0.425	0.378	0.689	0.837
140	0.854	0.597	0.451	0.726	0.822
IUPwavelet	0.859	0.613	0.472	0.736	0.804
609	0.912	0.527	0.440	0.720	0.804
271	0.883	0.536	0.419	0.710	0.804
272	0.839	0.591	0.430	0.715	0.798
538	0.971	0.327	0.298	0.649	0.796
572	20.947	0.396	0.343	0.672	0.777
153	0.908	0.383	0.291	0.646	0.758
681	0.906	0.371	0.277	0.639	0.726
393	0.788	0.558	0.346	0.673	0.724
168	0.788	0.558	0.346	0.673	0.724
188	0.997	0.222	0.219	0.610	0.710
132	0.971	0.201	0.172	0.586	0.704
686	0.971	0.338	0.309	0.655	0.704
Naiv	0.830	0.488	0.319	0.659	0.696
594	0.993	0.066	0.058	0.529	0.671
284	0.937	0.280	0.218	0.609	0.609

Table 6.10: Comparison of prediction time (seconds) on SDB

Percentage	VSL2	DISOPRED2	DISpro	IUPwavelet
20%	52,802	25,584	14,574	335
40%	111,100	53,850	30,306	675
60%	167,100	81,974	45,677	1,005
80%	225,302	110,376	61,637	1,340
100%	281,126	133,432	76,992	1,670

human expert in CASP7, it applies VSL2 as its server. The prediction is based on our training database *SDB*. Table 6.10 illustrates the speed of prediction. The first column represents the percentage of sequences predicted in the training database *SDB* by selected predictors. All other numbers in the table are prediction time measured in seconds. Obviously, IUPwavelet uses significantly less time than VSL2, DISOPRED2 and DISpro in short DR prediction.

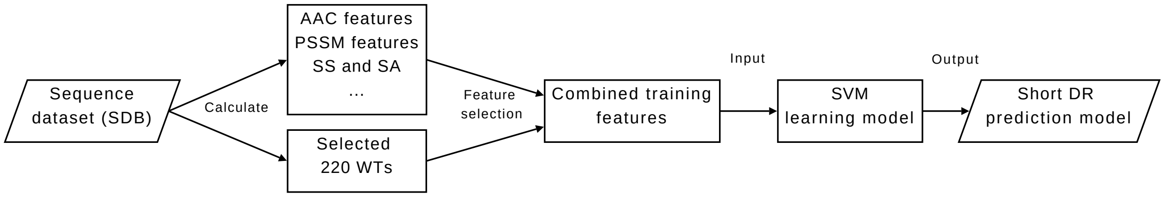


Figure 6.11: Feature combination learning model

Table 6.11: Four feature clusters

Cluster	Cluster name	#Feature
1	AAC based (Group 1 + 4 reduced-AAC)	20 + 4 = 24
2	PSSM from UniRef100 and nr (Group 10 + 300 PSSM features)	330 + 300 = 630
3	Prediction results of SS and SA (Group 7)	11 + 2 = 13
4	WT and space status (Group 6 + 1 space status)	220 + 1 = 221

6.6 Wavelet Transform Enhanced Accurate Short DR Prediction

In the previous section, we have shown that AAindex based WTs perform better than other features for short DR prediction except for the PSSM-based feature groups. The predictor IUPwavelet, based on WTs, achieves accurate prediction as well as high efficiency. To achieve a higher accuracy of prediction, in this section, we combine other feature groups including AAC, PSSM and the prediction results of secondary structure and solvent accessibility with WTs. The learning process is illustrated in Figure 6.11.

Given the ten existing groups proposed in Table 6.8, there are $2^{10} - 1 = 1,023$ combinations, which takes a long time to measure the accuracy level involved. Therefore, we categorise some selected feature groups into four feature clusters, which have only 15 combinations.

6.6.1 Feature Clusters

We organise features into four clusters. In Table 6.11, most clusters are based on feature groups proposed in Chapter 6.5.2.

In Cluster 1, four reduced-AAC are based on charge and hydrophobicity, as described in Table 3.2. Feature Cluster 2 includes PSSM features generated from databases UniRef100 and nr. Table 6.8 shows that PSSM achieves the highest accuracy of prediction when all PSSM values in the matrix are used for training. Given a window size of 15, the PSSM of the window generated from UniRef100 corresponds to $15 \times 22 = 330$ features; and the PSSM

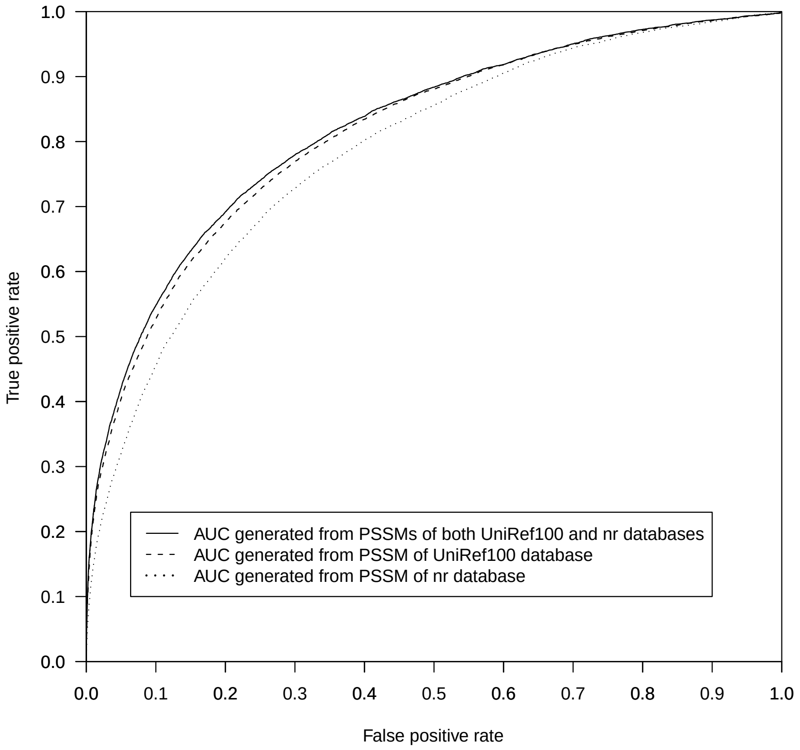


Figure 6.12: Accuracy of prediction in a ten-fold cross-validation test on the database SDB

of the window generated from nr corresponds to $15 \times 20 = 300$ features. To evaluate the performance of these PSSMs in short DR prediction, we have shown their ROC curves from the ten-fold cross-validation tests in Figure 6.12. It is clear that the PSSM combination has achieved a higher AUC value (82.1%) than the AUC value from nr (78.6%); it is also slightly higher than the AUC value from UniRef100 (81.4%). The corresponding P-value of a pairwise comparison between AUCs generated from the PSSM combination and nr is less than 10^{-10} and the corresponding P-value of a pairwise comparison between AUCs generated from the PSSM combination and UniRef100 is around 0.21. So feature Cluster 2 includes PSSMs calculated from both UniRef100 and nr databases. Feature Cluster 3 is the feature Group 7 as described in Section 6.5.2. Finally, feature Cluster 4 is a WT and one space status. 220 WTs selected in Section 6.5.1 are combined with space status, which is the percentage of space in sliding windows. Feature Cluster 4 includes 221 features. Predicted scores of confidence are always smoothed in a window of seven residues for all feature clusters.

Table 6.12: Contribution of WT when combined with other feature clusters in the ten-fold cross-validation tests

Cluster 1	Cluster 2	Cluster 3	Cluster 4 (AAindex WTs)	AUC(%)
			*	79.6
*				74.0
*			*	79.4
	*			82.1
	*		*	83.5
		*		63.4
		*	*	80.9
	*	*		82.6
	*	*	*	84.0
*	*	*	*	84.1

6.6.2 Combining Feature Clusters for Accurate Short DR Prediction

For a fair comparison of the four feature clusters, a window of 15 residues is applied and seven residues at the N and C termini are not measured. From all 15 combinations of four feature clusters, the predictor built from all four clusters achieves the highest ten-fold cross-validation test result with an AUC value of 84.1%. So overfitting is not a big issue in our short DR prediction, since the total number of features from four clusters is more than 800. The predictor built from Clusters 2, 3 and 4 is the second most accurate predictor with an AUC value of 84%.

To ascertain the contribution of the AAindex WTs, we show the AUC results for AAindex WTs combined with other feature clusters in Table 6.12. A "*" mark indicates if a feature cluster is included in the feature space of the predictor. AAindex WTs can achieve an AUC value of around 79.6% by itself, as shown in Table 6.12. However, it is not suitable to combine with feature Cluster 1 because the AUC value of combined clusters is lower than that of Cluster 4 alone. In contrast, the WT achieved higher AUC results when it was combined with feature Clusters 2 or 3. It improves the AUC value of Cluster 2, PSSM, by 1.4% and the AUC value of Cluster 3, prediction results of SS and SA, by 17.5%. When WT is combined with feature Clusters 2 and 3 at the same time, the AUC value is 84%, and close to the highest AUC value of 84.1% achieved by all four clusters.

Table 6.13: Performance of each feature cluster under different window sizes

Window	Cluster 1	Cluster 2	Cluster 4 (AAindex WTs)
7	70.3%	78.8%	78.2%
11	72.9%	80.6%	78.7%
15	**73.7%**	81.3%	78.9%
19	73.7%	**81.4%**	79.0%
23	73.2%	81.4%	**79.1%**

Window Size Tuning

As described in Table 6.8, the window size used in a feature group influences the accuracy of prediction. A predictor built with various feature clusters derived from the same window size is unlikely to achieve the best prediction accuracy when these clusters are combined. Different window sizes should be applied for different feature clusters. However, it is computationally expensive to consider all possible combinations of feature clusters and window sizes. A heuristic approach is adopted here to find appropriate window sizes. First, the accuracy of predictors based on a single feature cluster and different window sizes is measured. Table 6.13 illustrates the performance of these predictors (feature Cluster 3 is not shown as it is not window based). The first column is the training/prediction window for different feature clusters. Given that the largest window size measured is 23, for fair comparison, 11 residues at the N and C termini are excluded from the evaluation.

The most suitable windows of 15, 19 and 23 for feature Clusters 1, 2 and 4 are then used to build the predictor. All four feature clusters are included in the feature space and 11 N/C terminal residues are excluded from the evaluation. Finally, the size of the window is increased/decreased by two for each feature cluster at every step. This operation stops when the accuracy of the cross-validation test starts to decrease. Under window sizes 27, 13 and 23 for feature Clusters 1, 2 and 4, we achieve the highest AUC result of 84.1% for internal residues. It is 1.1% higher than the AUC result from the predictor with the training feature extracted from a universal window size of 15 residues.

6.6.3 Results

In our final ten-fold cross-validation test shown in Figure 6.13, N and C terminal residues are taken into consideration. Feature Cluster 3 and AAindex WTs are used to predict the terminal residues as they do not involve windowing. The ROC curve of the predictor built

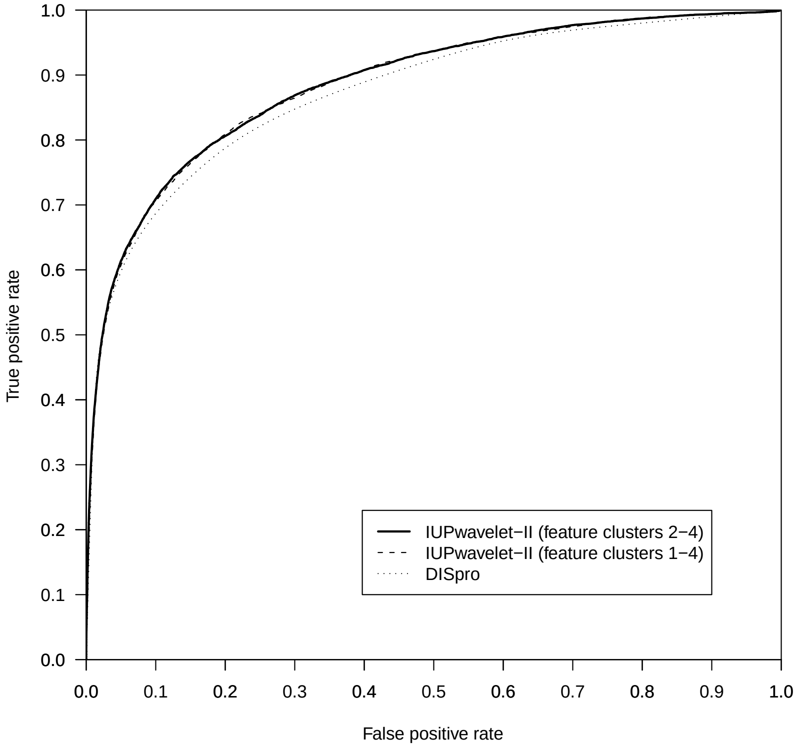


Figure 6.13: Results of the ten-fold cross-validation test on the database SDB

from all four clusters is shown as the dashed line in Figure 6.13. Previously selected window sizes of 27, 13 and 23 are applied for Clusters 1, 2 and 4. According to Table 6.12, the predictor based on the combination of feature Clusters 2, 3 and 4 also achieves a high AUC value. The ROC curve of the predictor trained by the combination of these three feature clusters is shown as the solid line of Figure 6.13.

Two curves for predictors trained by feature Clusters 1 to 4 and 2 to 4 almost overlap and their corresponding AUC values are 88.6% and 88.7%. These AUC values are higher than the cross-validation result of DISpro (87.8%), shown as the dotted line. Although two predictors called IUPwavelet-II have similar prediction accuracy in Figure 6.13, the predictor trained by feature Clusters 2 to 4 uses less features and is used in following experiments. In the significance test of Figure 6.13, the P-value of the pairwise comparison between AUCs generated from IUPwavelet-II and DISPro is 0.13.

Results of the cross-validation test reveal that IUPwavelet-II achieved a high accuracy for short DR prediction. IUPwavelet-II is then used for the blind test on the CASP7 targets.

The complete training dataset *SDB* is used to build the predictor. For the CASP7 targets only DRs of more than three residues are considered. The ROC of the blind test is shown as the bold line in Figure 6.14. Figure 6.14 (a) shows the performance of prediction on complete CASP7 sequences and the AUC value of IUPwavelet-II reaches 84.6%. This value is superior to all predictors registered as servers in the CASP7 competition and is equivalent to the human expert Fais [Ishida et al., 2006]. In Figure 6.14 (a), the dashed and dotted lines are from DISOPRED[3] and DISpro, the winner and runner-up servers in the CASP7 competition. Both curves have smaller AUC values compared to IUPwavelet-II. The corresponding P-values of pairwise comparison between the AUC generated from IUPwavelet-II and DISOPRED is 0.056; and the P-value between the AUC generated from IUPwavelet-II and DISPro is less than 10^{-3}. Recall our initial attempt at a short DR predictor in Section 6.2, where the AUC value of DRaai-S on CASP7 was only 72.2%, which is much lower than that of IUPwavelet-II.

From the report of CASP7 [Bordoli et al., 2007], DRs in the internal area of the protein sequences are more difficult to predict than DRs at the termini. When ten residues at the termini are excluded from our experiments, the ROC for internal residues is shown in Figure 6.14 (b). The AUC value of our predictor is 80.7%, which is again better than all server predictors including DISOPRED and DISpro. This result is slightly higher than the runner-up human expert CBRC-DR in the CASP7 competition.

We also validated IUPwavelet-II on datasets PDB and DisProt. The PDB dataset updates every day and thousands of new structures are added each year. We used those new updated sequences to carry out a blind test. The selection procedure was the same as that introduced in Section 3.1.2. New sequences with a pairwise sequence identity of more than 25% for any sequence in *SDB* were excluded from the test dataset. The final dataset, called *SDB2*, consisted of 508 sequences and 151,037 residues. 7.2% of the residues were annotated as disordered. We also selected sequences from DisProt 3.6 for our blind test. IUPwavelet-II is a short DR predictor, so sequences in DisProt 3.6 with only long DRs (>30) were excluded. The selected dataset is called *SDisProt*, which includes 41 sequences and 12,337 residues.

The predictor IUPwavelet-II was then applied to predict datasets *SDB2* and *SDisProt*. Results of the prediction are shown in Table 6.14. When the dataset of the blind test was *SDB2*, IUPwavelet-II achieved an AUC value of 86.8%. Under the default cutoff value,

[3] The DISOPRED score comes from its website http://bioinf.cs.ucl.ac.uk/disopred/disopred.html because more recent DISOPRED2.2 package does not provide the recall value when the false positive rate is more than 10%.

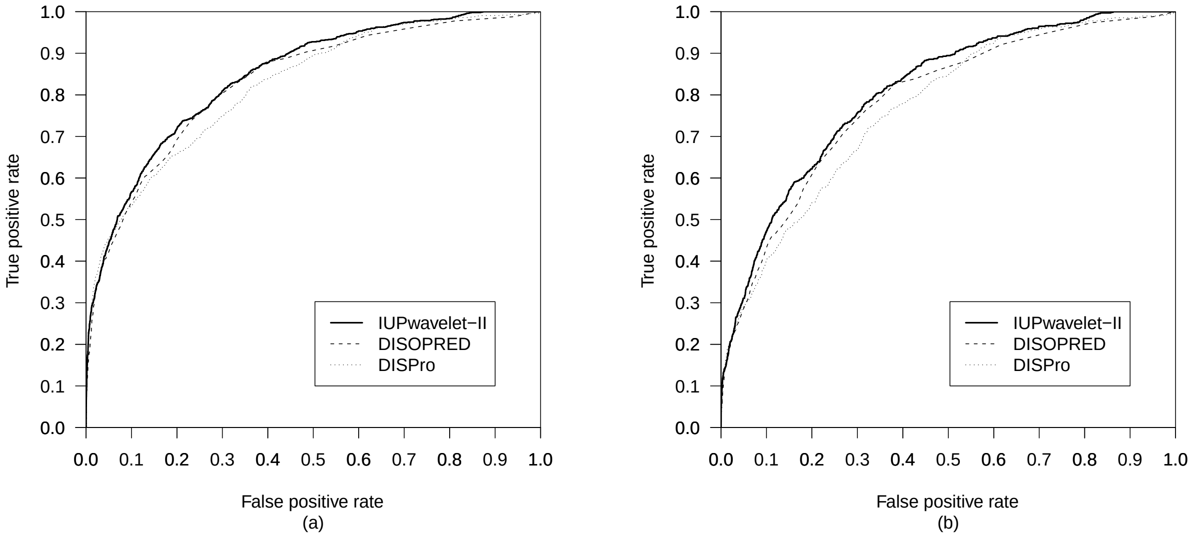


Figure 6.14: The blind test results on CASP7

Table 6.14: Blind test results on SDB2 *and* SDisProt

Datasets	AUC	Disordered residues		Ordered residues	
		Recall	Precision	Recall	Precision
SDB2	87.5%	86.0%	87.0%	85.2%	88.9%
SDisProt	82.6%	84.5%	85.5%	81.2%	87.5%

recall and precision over disordered residues were 86.0% and 87.0%; recall and precision over ordered residues were 85.2% and 88.9%, respectively. This AUC value is similar to the results of the ten-fold cross-validation test of IUPwavelet-II (88.7%). It confirms that the performance of IUPwavelet-II is consistent over multiple datasets including *SDB*, *SDB2* and CASP7 targets. In comparison, the results of prediction over dataset *SDisProt* was less successful, with an AUC value of 82.6%. Recall and precision over disordered and ordered regions under a default cutoff value were also lower than that of *SDB2*. It may be attributed to the limited number of sequences (41 sequences) and disordered residues (521 residues) in *SDisProt*.

6.7 Related Work

In this section, we first describe the application of WT for other structure prediction tasks, and then we discuss recent developments in short DR prediction.

Single Wavelet Transform for Structure Prediction

In recent years, there has been a growing interest in using wavelets for the analysis of sequence and functional genomics data [Liò, 2003]. Liò [2003] discussed the general application of wavelets in bioinformatics. He summarised the state of the art of wavelet research in different areas such as molecular biology, genome sequence, protein structure and microarray data analysis. Liò and Vannucci [2000] predict transmembrane helix locations using WT. They first generate a propensity profile by substituting the sequence with a propensity scale for amino acids in the membrane environment. Then the profile is converted into wavelet coefficients using the Daubechies' basis of WT scale eight. At the third step, wavelet coefficients that detect abrupt changes in the profile are selected. The corresponding regions of the selected wavelet coefficients are potential targets, transmembrane helices. Finally, wavelet shrinkage [Donoho and Johnstone, 1998] is applied to remove noise. This approach has achieved better or comparable accuracy of prediction with other predictors. A similar approach is applied to predict protein secondary structure [Chen et al., 2006a]. Hydrophobic values of 20 amino acids are selected to carry out the substitution of protein sequences and generate profiles. With a given basis, WT scales ranging from 1 to 64 are examined. The best scale (around 16) is applied to detect secondary structure. These applications of the WT are based on a single WT for structure prediction and are similar to our single WT based short DR predictor proposed in Section 6.4.

Recent Developments in Short Disordered Region Prediction

Hecker et al. [2008] have updated DISpro by training the model with a much larger dataset of 799,153 residues. The new model achieves an AUC value 0.2% higher than DISOPRED on the CASP7 dataset. However, this result is still lower than ours. Two more recent predictors metaPrDOS [Ishida and Kinoshita, 2008] and DISOclust [McGuffin, 2008] have achieved an AUC value higher than ours on the CASP7 dataset. However, both focus on how to combine existing predictors to achieve a higher accuracy of prediction. The major contribution of their work is the meta approach. In contrast, we are looking for feature groups suitable for large scale short DR prediction that are also able to enhance the results of predictions when combined with other feature groups.

6.8 Summary

In this chapter we have applied WTs to AAindices for short DR prediction. Specifically, the single WT based predictor can achieve an AUC value slightly higher than that of the predictor based on AAC and the SVM. The predictor IUPwavelet, which applies multiple WTs and SVM, is then proposed to enhance the accuracy of prediction. Greedy selection is used to select WTs, which are adopted to train SVM based predictors. The greedy selection has reduced the original large feature space significantly. By investigating some commonly used feature groups in short DR prediction, IUPwavelet is the second best performing feature group following PSSM. Ten-fold cross-validation reveals that IUPwavelet achieves an AUC value of 85.5%. The AUC value of IUPwavelet on CASP7 targets is 80.4%, more accurate than most server predictors in the competition. The speed of prediction reveals that IUPwavelet is much faster than some best performing predictors including VSL2, DISOPRED2 and DISpro.

We have investigated WTs in combination with commonly used features in short DR prediction. The best performing combination of feature clusters includes the PSSM, the prediction result of secondary structure and solvent accessibility and AAindex based WTs. The performance of WT shows that it is not only a good indicator of short DRs by itself, but also a good candidate when combined with other features. The combined feature groups achieve a higher accuracy of prediction. Our AUC value following a ten-fold cross-validation test is 88.7%, 0.9% higher than that of DISpro. The blind test on the CASP7 dataset reveals the AUC value can reach 84.6%, which is more accurate than all registered server predictors and comparable to the best performing human expert predictors.

Chapter 7

Prediction of Global Disorder by Machine Learning Models

Another goal of IUP study is to predict whether a sequence is largely disordered, rather than pinpointing locations of DRs in protein sequences. We call this study global disorder prediction. Machine learning approaches are adopted by Weathers et al. [2004] and Shimizu et al. [2007b] to predict globally disordered sequences. Some other predictors [Prilusky et al., 2005; Galzitskaya et al., 2006a] make the prediction by calculating the physicochemical properties of query sequences without building learning models. Predictors by Dunker et al. [2000] and Oldfield et al. [2005a] are based on the cumulative distribution function (CDF) and prediction results from existing DR predictors.

In Chapter 4 we have shown that reduced-AAC can achieve a high accuracy in long DR prediction. In this chapter, predictors based on the decision tree, random forest and SVM are developed. Training features of these predictors include AAC, reduced-AAC, or combinations of them from protein sequences.

A set of rules can be derived from the decision tree built from our training dataset (details in Section 7.1) that describe complex conditions of AACs for the disordered or ordered status of proteins. These rules confirm that IUPs have low overall hydrophobicity, high net charge and low sequence complexity [Uversky, 2002a]. More importantly, they present complex AAC information that has been previously unknown. In comparison, predictors based on random forest and SVM are harder to interpret, but they achieve more accurate results of prediction than the predictor based on the decision tree.

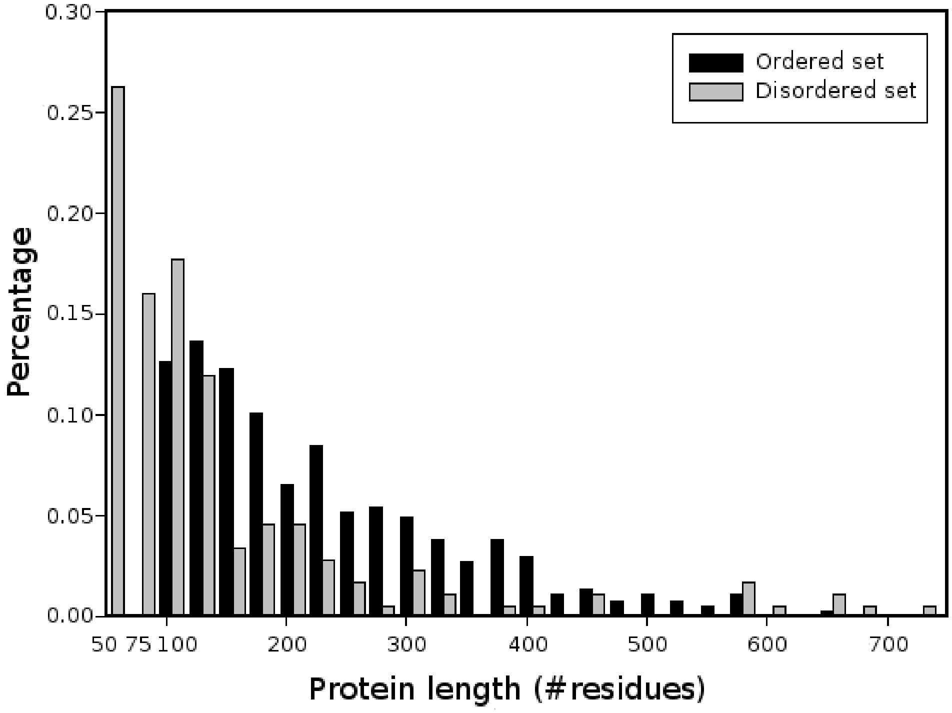


Figure 7.1: The distribution of ordered vs disordered segments

7.1 Material and Methods

We first show our order and disorder training databases, and then describe the construction of the AAC and reduced-AAC training features. We finally demonstrate how the decision tree, random forest and SVM are applied to the prediction.

The disordered training set called *GDDB* in our study is extracted from DisProt (release 2.1). There are 176 completely disordered sequences in this database. Our ordered training set *ODB* is extracted from PDB-Select-25. There are 366 completely structured sequences in *ODB*. The distribution of disordered and ordered segments of different lengths is plotted in Figure 7.1. We can see that there are more short segments in *GDDB* than in *ODB*. Specifically, more than 40% of disordered segments contain less than 100 residues. Ordered segments usually contain less than 700 residues. In contrast, disordered segments can contain more than one thousand residues (not shown in Figure 7.1).

Predicting IUPs from primary sequences is a binary classification problem. The unstructured sequences are tagged with the label *P* and the structured sequences are tagged with the label *N*. Given feature AAC, each protein sequence in the training dataset is represented by its AAC. By a single scan of the given protein sequence database, the composition database is constructed. With our training data, the IUP database is converted into a 176×20 matrix and the ordered database is converted into a 366×20 matrix. The AAC for each sequence of the IUP and ordered database are still tagged with the labels *P* and *N*.

To measure the performance of reduced-AACs in global disorder prediction, each AAC is reduced into four groups according to hydrophobicity and charge properties, as shown in Table 3.2. The corresponding matrices for disordered and structured databases are 176×4 and 366×4. Further, both AACs and reduced-AACs are used as feature spaces to build predictors.

Three machine learning models including decision tree, random forest and SVM are applied to build our predictors. With each of the three machine learning models, a predictor is built from the order and disorder training matrices. Predictors predict the possibility of a query protein sequence being an IUP or otherwise.

7.2 Results

Recall and precision on both classes, and ROC curves calculated from ten-fold cross-validation, are used to evaluate the accuracy of our predictors. In our discussions, IUPs are our main focus, and so recall, precision and AUC refer to the disordered class, unless otherwise specified. Accuracies based on the feature AAC are shown in Table 7.1. Generally, the recall and precision of ordered proteins are higher than for disordered proteins. This is because the number of ordered sequences is around twice that of disordered ones. Machine learning models generally focus on achieving the highest overall accuracy, which may sacrifice the accuracy of minority classes. In Table 7.1, among three predictors built from feature AACs, SVM based predictor achieves the highest accuracy of prediction. Under the default cutoff threshold, recall and precision values are 89.2% and 84%. The AUC value of the SVM based predictor is 95.8%, which is slightly higher than that of the predictor based on random forest (95.1%). Predictors based on random forest and SVM perform clearly better than the predictor based on the decision tree, whose AUC value is only 80.1%.

In Table 7.2, the accuracy of prediction based on reduced-AACs is shown. Compared with Table 7.1, AUC values change to 84.4%, 89.7% and 92.3%, respectively for predictors based on the decision tree, random forest and SVM. It seems that the reduced-AAC does

Table 7.1: Ten-fold cross-validation test based on AAC

	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
Decision tree	80.1%	72.2%	79.4%	91.0%	87.2%
Random forest	95.1%	79.0%	92.7%	97.0%	90.6%
SVM	95.8%	89.2%	84.0%	91.8%	94.6%

Table 7.2: Ten-fold cross-validation test based on reduced-AAC

	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
Decision tree	84.4%	76.1%	80.7%	91.3%	88.8%
Random forest	89.7%	74.4%	83.4%	92.9%	88.3%
SVM	92.3%	76.7%	88.8%	95.4%	89.5%

not provide adequate information for these predictors.

ROC curves of predictors based on AACs and reduced-AACs are shown in Figures 7.2 and 7.3. In Figure 7.2, both random forest and SVM predictors perform substantially better than the decision tree predictor (P-value $< 10^{-8}$). The AUC of SVM is slightly larger than that of the random forest (P-value $= 0.41$). The decision tree achieves a true positive rate of 80% when the false positive rate is around 44%. Random forest and SVM achieve the same true positive rate with a false positive rate 4.4% and 1.9%. Therefore, predictors based on random forest and SVM achieve a good global IUP recall rate without introducing too many errors.

The ROC curve of the decision tree is not smooth compared with that of SVM and random forest. In Figure 7.2, when the false positive rate is between 15% and 25%, the true positive rate is almost the same. This is due to the limited number of rules for prediction. There are not enough rules corresponding to each cutoff value along the false positive rate axis. A similar problem is present in Figure 7.3, where the number of training features is less and only a few rules are generated.

Curves in Figure 7.3 are closer to each other, and the AUC values of both random forest and SVM have dropped due to insufficient information provided from the four training features. These two ROC curves have a P value $= 0.1$. However, the decision tree has a much larger AUC value when the number of features is reduced to four.

Finally, we build predictors based on the combination of AAC and reduced-AAC features, and results are shown in Figure 7.4. All three predictors perform better than their

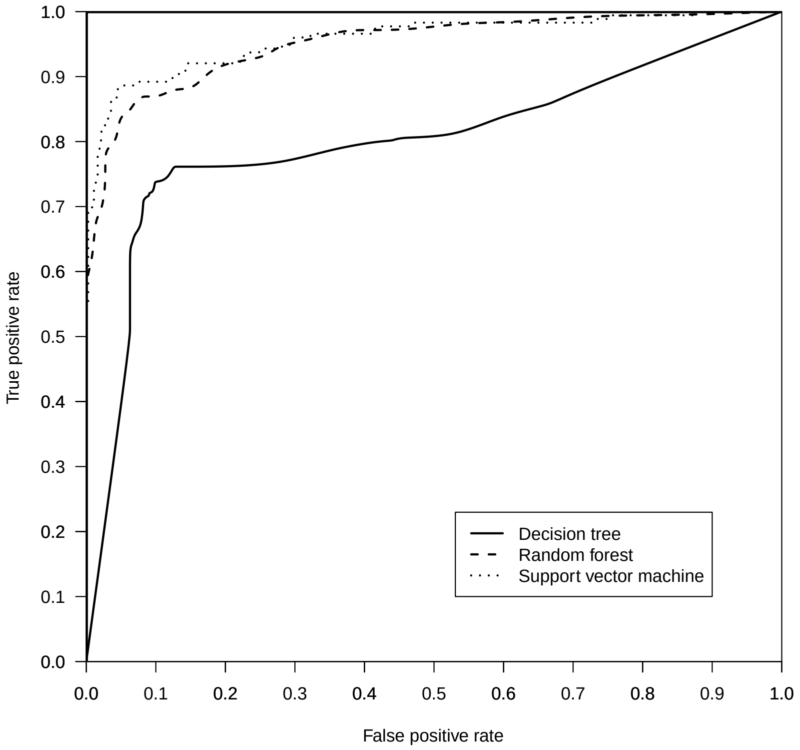


Figure 7.2: The ROC curve of AAC-based global disorder predictors

corresponding predictors based on AACs or reduced-AACs only. In Figure 7.4, AUC results of predictors based on the decision tree, random forest and SVM are 85.5%, 96.1% and 96%, respectively. The ROC curve difference between random forest and SVM is trivial, but the curve of the decision tree is significantly lower than the other two (P-value $< 10^{-7}$)

7.3 Learning Amino Acid Composition Rules for IUPs

Using the default settings for parameters of the decision tree model, only 12 AAC rules are learnt for *GDDB*, and 4 AAC rules for *ODB*. These rules are listed in Tables 7.3 and 7.4, respectively. Rules are listed in decreasing order of estimated predictive accuracy (the last column of Tables 7.3 and 7.4). All rules are very concise and understandable. The rules that describe the disordered state are much simpler than those describing the ordered state. This is due to the biased composition and lower sequence complexity of the sequences in the *GDDB* dataset. Conversely, the AAC in the sequences of the *ODB* dataset are more uniform

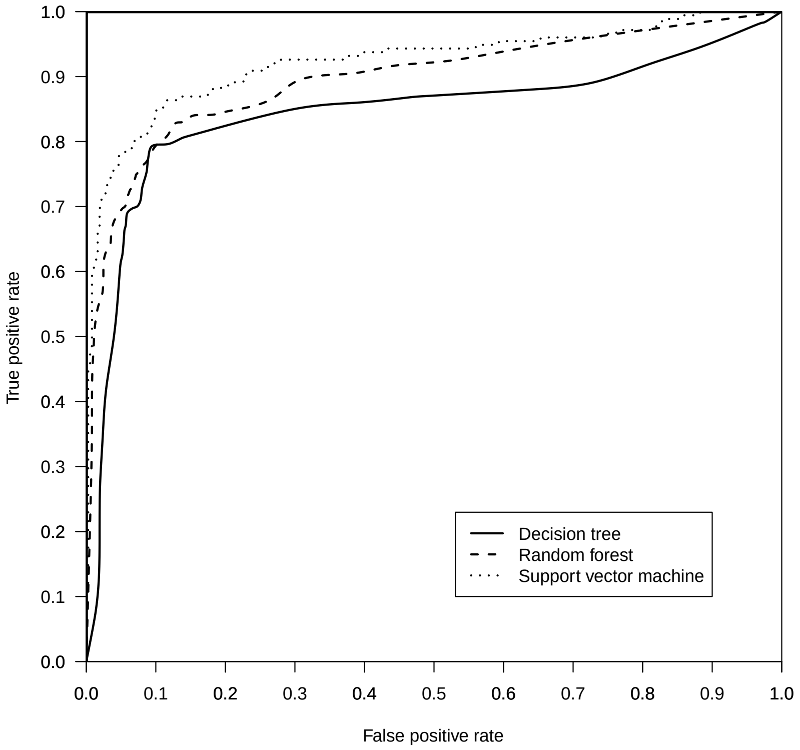


Figure 7.3: The ROC curve of reduced-AAC based global disorder predictors

Table 7.3: Disordered AAC-rules from GDDB

	Rule	Confidence		Rule	Confidence
1	F≤0.013; L≤0.118; Y≤ 0.025	97.6%	7	E>0.124	94.7%
2	F≤0.007	97.4%	8	S>0.090; V≤0.034	94.6%
3	H ≤0.069; K>0.122; Y≤0.045	96.6%	9	G>0.093; I≤0.022	90.5%
4	P>0.103	95.9%	10	D>0.101; I≤0.035	89.9%
5	I≤0.022; K≤0.122; W≤0.003	95.5%	11	D>0.101; S>0.081; V>0.034	89.1%
6	C≤0.004; R≤0.011; Y≤0.039	95.3%	12	C>0.027; H>0.042; K≤0.122	79.4%

and sequence complexities are much higher. As a result, there are fewer rules and they tend to be more complicated.

Some disordered AAC rules in Table 7.3 are extremely explicit from the point of view of physicochemical properties of amino acids, such as rules 2, 4 and 7. They indicate that sequences extremely depleted in Phe ($F \leq 0.70\%$) or extremely enriched in Pro ($P > 10.3\%$) or Glu ($E > 12.4\%$) are very likely to be in a disordered state. Rule 1 shows that if a sequence

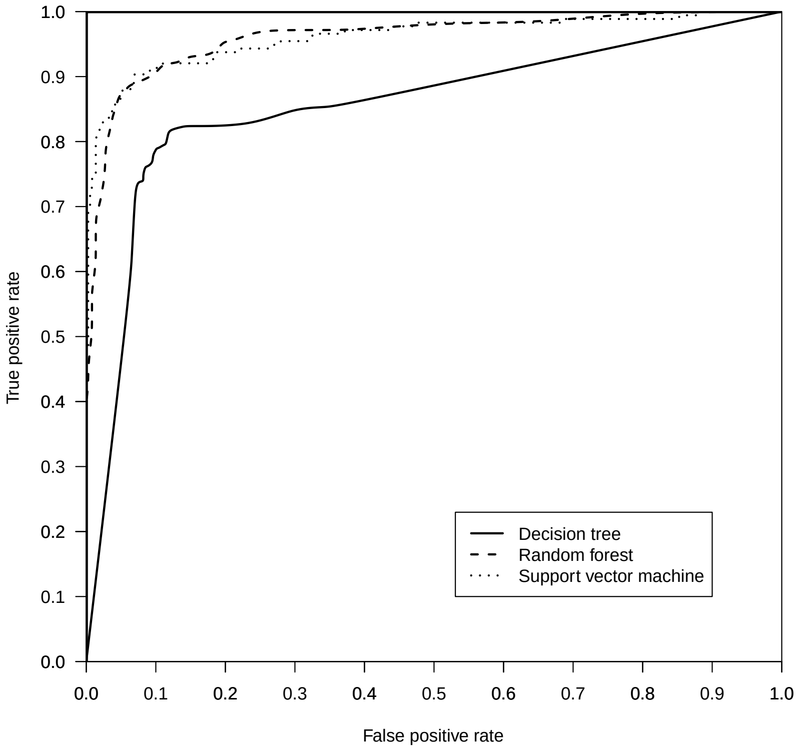


Figure 7.4: The ROC curve of the AAC and reduced-AAC based predictors

Table 7.4: Ordered AAC-rules from ODB

	Rule	Confidence
1	D≤0.101; E≤0.124; F>0.013; H≤0.042; I>0.022; K≤0.122; P≤0.103; V>0.033	95.5%
2	E≤0.124; F>0.013; G≤0.104; K≤0.122; P≤0.103; S≤0.090; W>0.003	94.5%
3	E≤0.124; F>0.013; P≤0.103; Y>0.045	94.4%
4	F>0.013; H>0.069	85.7%

lacks Phe (F), Leu (L) and Tyr (Y) at the same time, it most likely is in a disordered state. Most of the other rules listed in Table 7.3 are the combination of abundance in polar or hydrophilic residues and dearth of hydrophobic residues. Interestingly, positively charged residues His (H), Lys (K), and the sulphur-containing residue Cys (C) are environment-dependent in their state. For example, the sequence could be in a disordered state if the content of Lys (K) is greater than 12.2%, but that of His (H) less than 6.9%, and that of Tyr (Y) less than 4.5% at the same time (rule 3). In contrast, the sequence could also be in the disordered state if the content of Lys (K) is less than 12.2% but the content of Ile (I) is less than 2.2% and that of Trp (W) is less than 0.3% (rule 5), or the content of Cys (C) is larger than 2.7% and that of His (H) is larger than 4.2% (rule 12).

The disordered and ordered rules not only confirm that residues Phe, Tyr, Trp, Ile, Leu and Val are ordered promoters and Pro, Glu, Gln, Ser and Gly are disordered promoters as indicated by others [Dunker et al., 2002a; Uversky et al., 2000; Uversky, 2002a], but that they also describe the detailed and complicated impact from the combinations of different amino acids.

7.4 Feature Extraction

Feature extraction refers to the process where a new reduced feature space is constructed from the original feature space through some transformation. Grouping 20 AACs into four groups of reduced-AACs is essentially a kind of feature extraction, which extracts new features from the original feature set through functional mapping. From the perspective of physicochemical properties, amino acids are grouped into four categories, which reduces the original feature space. Alternatively, feature extraction can be achieved through statistics or linear algebra supported algorithms such as PCA (Principal Components Analysis) and subset selection. PCA is an unsupervised space reduction technique, which seeks to map or embed data points from a high dimensional space to a low dimensional space while keeping the relevant linear structure intact [Boutsidis et al., 2008]. In contrast, subset selection tries to remove irrelevant and redundant features from the data and improve the performance of predictors. It will not generate new features.

We have applied both PCA and subset selection algorithms to 20 AACs. Only four of the most important features are extracted/selected by PCA and subset selection. The performance of predictors built from these features generated from PCA and subset selection are shown in Table 7.5.

According to the AUC results of Table 7.5, predictors trained from four features extracted from PCA outperform that trained from subset selection. During the space reduction process, PCA projects data to lower dimensional space (four in this case) and tries to preserve all of the original relevant linear structure. Subset selection simply keeps the four most important features and ignores others. Four PCA features presumably contain more information from the original AAC than four subset selection features. However, AUC values from four PCA features in Table 7.5 are only comparable to that of the reduced-AAC in Table 7.2 for any

Table 7.5: The ten-fold cross-validation tests based on PCA and subset selection

PCA					
	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
Decision tree	82.7%	71.0%	85.6%	94.3%	87.1%
Random forest	89.5%	74.4%	89.7%	95.9%	88.6%
SVM	92.2%	78.4%	90.2%	95.9%	90.2%

Subset selection					
	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
Decision tree	79.6%	54.0%	84.8%	95.4%	81.2%
Random forest	86.7%	65.9%	77.9%	91.0%	84.7%
SVM	88.2%	59.7%	77.2%	91.5%	82.5%

learning model. These results suggest that domain knowledge based feature extraction can be an important alternative in global disorder prediction, especially given that it does not incur any computational cost.

7.5 Wavelet Transform Enhanced Global Disorder Prediction

We have discussed in Chapter 6 that the short DR predictor based on a WT has achieved comparable or even better prediction accuracy than existing DR predictors. In this section, we investigate whether the WT enhanced predictor can also perform well in global disorder prediction.

We have applied the short DR predictor IUPwavelet-II of Section 6.6 directly to the datasets *GDDB* and *ODB*. However, IUPwavelet-II is trained from sequences with mainly short DRs and patterns of corresponding WT can be completely different from those of *GDDB* and *ODB*. Many features used in IUPwavelet-II are alignment based, which are not appropriate for the completely disordered dataset *GDDB*. So we also built a new predictor using the machine learning model SVM. Instead of using all of the training features of IUPwavelet-II, we applied only 220 WTs as features of this predictor. Blind test results from predictor IUPwavelet-II and results of the ten-fold cross-validation test of the new predictor are illustrated in Tables 7.6 and 7.7.

According to Table 7.6, predictor IUPwavelet-II did not perform well with an AUC value of 80.5%. Both recall and precision on IUP and ordered sequences are similar to that of the decision tree trained from AACs. We believe it is due to the significant difference between the

Table 7.6: Blind test results from predictor IUPwavelet-II

Method	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
IUPwavelet-II	80.5%	74.0%	78.5%	91.2%	88.9%

Table 7.7: The ten-fold cross-validation tests based on wavelet transform

Method	AUC	IUP		Ordered	
		Recall	Precision	Recall	Precision
Wavelet transform	96.8%	85.5%	86.2%	93.6%	94.0%

training dataset of IUPwavelet-II and our global disordered sequences used for prediction. Table 7.7 shows that when we rebuild the predictor from WTs, the AUC value of the ten-fold cross-validation test is superior to all our previous results from various machine learning models trained by AACs or reduced-AACs. The improvement in the AUC value is at least 1%. Given that both random forest and SVM have achieved AUC values over 95% when the training feature is the AAC, the WT can provide complementary information in global disorder prediction.

7.6 Prediction for Plasmodium Falciparum and Yoelii Proteomes

According to results of the cross-validation test in Figure 7.4, the SVM based predictor achieves the true and false positive rates of 90.3% and 7.7% using default settings in global IUP prediction. We then apply this predictor to predict global IUPs in Plasmodium yoelii and Plasmodium falciparum proteomes. A systematic analysis of the global disorder of the Plasmodium falciparum and yoelii proteomes has not been undertaken previously. From the perspective of biology, many of the Plasmodium falciparum and yoelii proteins are known to contain extensive low complexity regions, which cause difficulties in identifying homologues via a BLAST search of databases [Feng et al., 2006]. Therefore, it has not been possible to describe the structure and functions of a large number of Plasmodium proteins from homologous sequences and we believe that many proteins in these proteomes are actually globally disordered.

7.6.1 Test Datasets

Both Plasmodium falciparum and Plasmodium yoelii are parasites. Plasmodium falciparum is well known for causing malaria in humans and is transmitted through the female anopheles mosquito. The majority (>90%) of human malarial infection and death is due to this parasite. Plasmodium yoelii is used in the laboratory to infect mice as a model of human malaria research. Both Plasmodium falciparum and yoelii proteomes are obtained from PlasmoDB (`http://plasmodb.org/plasmo/`)(March 2009). There are 5,283 sequences in the Plasmodium falciparum proteome and the average length is 771 amino acids. The Plasmodium yoelii proteome contains 7,758 proteins with an average length of 444 amino acids.

7.6.2 Results

At a false positive rate of 7.7%, 53.5% of sequences in the Plasmodium yoelii proteome and 51.5% of sequences in the Plasmodium falciparum proteome are predicted as globally disordered. Given that most sequences in both proteomes are much longer than the threshold of long DRs (>30 residues), these predicted DRs are therefore more likely to be real DRs compared to predicted short DRs in short DR prediction. The popularity of global disorder in the Plasmodium falciparum and yoelii proteomes suggest that IUPs can play particularly important roles in host-parasite interactions. We have compared our results to the results of Feng et al. [2006], in which the predictor DisEMBL-1.4 is used for prediction. Over 60% of sequences in Plasmodium falciparum and over 45% sequences in Plasmodium yoelii proteomes are predicted as IUPs containing long DRs (>30 residues) by DisEMBL-1.4 [Linding et al., 2003a]. Feng et al. [2006] indicate that proteins from the Plasmodium falciparum proteome (mammalian malaria parasites) have more content of disorder compared with those from the Plasmodium yoelii proteome (rodent malaria parasites). However, our results reveal that in terms of global disorder, both proteomes contain a similar percentage of globally disordered sequences.

In our analysis, several types of proteins have been identified as globally disordered. In many proteins, low complexity sequences characterised by stretches of Asn residues or other hydrophilic amino acids are found inserted within or between globular domains. A large number of Plasmodium proteins that have this kind of sequence inserted within the globular domains are predicted as globally disordered.

Inter-domain sequences are frequently disordered; therefore, the abundance of large multidomain sequences in Plasmodium species and in Plasmodium falciparum in particular contribute to the abundance of global disorder in these organisms. One sample disordered multi-domain protein is PfEMP1, which has flexible interdomain sequences to facilitate the protein-protein interactions between Duffy binding-like domains and host proteins that lead to cytoadherence [Smith et al., 2000].

Long extensions to highly ordered proteins can also lead to global disorder. Examples are merozoite surface protein 3 (MSP3) and related proteins. They have highly acidic disordered C-terminal regions that may be important for providing a negative charge to a cell surface that lacks sialoglycoproteins [McColl and Anders, 1997].

7.7 Summary

In this chapter we have proposed global IUP predictors based on the decision tree, random forest and SVM machine learning models. Both AAC and reduced-AAC features have been applied to our study. We show that predictors based on random forest and SVM built from AACs are very accurate with AUC results of over 95%. These two predictors are more accurate than predictors built from the decision tree. In contrast, a decision tree has the advantage of understandability, which derives meaningful rules for predicting global IUPs. The AAC rules derived are consistent with biological findings [Dyson and Wright, 2005; Uversky et al., 2000; Uversky, 2002a] and quantitatively specify the combined effect of AACs in global disorder prediction.

Twenty amino acids are reduced to four groups of reduced-AACs. Given the machine learning model decision tree, these reduced-AACs improve the AUC value by 4% over full AACs. The performance of the reduced-AAC is comparable or better than features generated from PCA or subset selection algorithms. Results of cross-validation demonstrate all three learning models achieve the best performance if both AACs and reduced-AACs are applied as training features. The AUC results of these predictors are 85.5%, 96.1% and 96%, respectively.

Chapter 8

Conclusions

Intrinsically unstructured or disordered proteins (IUPs/IDPs) are proteins lacking a fixed three dimensional structure or containing long DRs. IUPs play an important role in biology and disease. Identifying DRs in protein sequences can provide useful information on protein structure and function, and assist high-throughput protein structure determination. Machine learning techniques, including NNs and SVMs have been widely used in such predictions.

In this chapter, we summarise the work presented in this thesis, outline the contributions we have made and discuss directions for future work.

8.1 Summary

DRs of different lengths show different properties. We have developed machine learning based disorder predictors that effectively address long DRs, short DRs and global disorder predictions. We have shown that these predictors are very accurate at predicting corresponding DRs.

Given that most machine learning based DR predictors have applied NN or SVM models, the process of explaining these predictors and on what characteristics they made their decision is not easy. To improve understandability, we apply the predictors based on the decision tree in global disorder and long DR prediction. These predictors achieve a high accuracy of prediction and also clearly illustrate the learning and prediction processes. Our global disorder predictor illustrates that domain knowledge based feature extraction can be an important alternative in global disorder prediction. The accuracy of our global disorder predictors is close to optimal.

We have built two long DR predictors based on the decision tree and random forest. The decision tree based predictor shows that long DRs can be accurately predicted even after AACs are reduced based on amino acid physicochemical properties. The final voting results from three independent predictors alleviate the bias problem in training datasets. The predictor IUPforest-L based on random forest and built from MOREAU-BROTO autocorrelation achieves the highest accuracy of prediction among existing long DR predictors.

We also built several short DR predictors by transforming selected AAindices. The predictor IUPwavelet is built by selecting appropriate WTs first. Then these WTs are input to a SVM model to build the predictor. IUPwavelet achieves a comparable accuracy of prediction with existing predictors within much less time. Predictors built from WTs are more accurate than predictors built from other non-alignment based features. Selected WTs are then combined with features commonly used in short DR prediction and the IUPwavelet-II predictor is trained from the combined feature space. The accuracy of prediction from IUPwavelet-II is better than all server predictors in the CASP7 competition. It shows WTs can complement existing features in short DR prediction.

Novel features such as MOREAU-BROTO autocorrelation and WT are easy to calculate and suitable for large scale disorder prediction. These features can also be combined with other features to achieve a higher accuracy of prediction in disorder prediction.

Structural biologists can apply our predictors for fast analysis of the order/disorder status of proteins they are interested in.

8.2 Future Work

In Chapter 6, we show that WT is suitable for short DR prediction. The simple WT based on the wavelet rbio2.2 makes AAindex profiles much more correlated with the disorder/order status of residues. The AUC value is comparable with that of a SVM based predictor trained on AACs. However, the application of WT to long DR prediction has not been tested. The commonly used BLAST program ignores these low entropy regions in the alignment. Therefore, alignment based features including results of secondary structure prediction and PSSM can be less reliable. In contrast, WTs are generated from the sequence itself without alignment. They may provide valuable information to enhance the accuracy of prediction of long DRs. The application of WTs in long DR prediction is therefore interesting and deserves to be investigated. Given long and short DRs have different physicochemical properties, WTs suitable for short DR prediction may not perform well on long DR prediction. To build a long DR predictor trained from WTs, a WT selection process is needed.

The property that WT is derived from the sequence itself can be very useful for predicting orphan proteins. In protein study, orphans are proteins without (or very few) homologues, i.e. proteins without any known domains [Rost, 2002]. It has been reported in studies covering 60 microbial genomes that about 14% of the genes are orphans [Siew and Fischer, 2003; Ekman et al., 2005]. The existence of orphan proteins may be due to the spontaneous creation of new proteins [Rost, 2002] or proteins evolved too far away from their closest neighbours to be detected [Ramani and Marcotte, 2003]. Therefore, neighbours and ancestors of orphan proteins are difficult to find, which can lead to unreliable alignment and PSSM. Predictors trained from these PSSMs are less effective at predicting orphan proteins. In future work, we will build an orphan protein database for evaluating prediction accuracy between our predictor and other predictors trained on alignment information.

In this study, we have built length-specific predictors that achieve a high prediction accuracy on DRs of different lengths. According to results of the CASP7 competition, meta DR predictors achieve the highest prediction accuracy. In the future we will build a meta predictor which summarises the results of prediction from our predictors so as to predict comprehensive DRs. More specifically, a separate predictor will be built which can assign weights for prediction results of global disorder, long and short DR predictors. The final results of prediction are calculated as the weighted average of these component predictors. Hopefully this comprehensive DR predictor based on our length-specific DR predictors can achieve a higher accuracy of prediction than other predictors. As a further step, prediction results from our length-specific predictors can be combined with prediction results of other predictors to achieve even higher accuracy.

According to results of prediction of some existing disorder predictors, DRs have shown different “flavours” of disorder according to their length, position and various physicochemical properties [Vucetic et al., 2003; Shimizu et al., 2005; Peng et al., 2006]. To improve the accuracy of prediction further, we believe flavour-specific predictors should be built. However, it is still unknown what flavour or combination of flavours is the most suitable measure distinguishing one kind of disorder from others. One candidate is protein families, since proteins of the same family share the same ancestors and some other properties. More work should be done to investigate the accuracy of prediction based on different protein families. In particular, profile Hidden Markov Models (HMMs) can be used. A HMM is a statistical model in which the system being modelled is assumed to be a Markov process with unobserved states. HMM has been applied to present a family of DNA or protein sequences. In signal peptide prediction, HMM achieves a very high accuracy of prediction [Zhang and Wood, 2003]. Therefore, it may also perform well in DR prediction for sequences of the same family.

Bibliography

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

M. L. Anson and A. E. Mirsky. On some general properties of proteins. *Journal of General Physiology*, 9(2):169–179, 1925.

T. K. Attwood and D. J. Parry-Smith. *Introduction of bioinformatics*. Pearson Education, 2001.

A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.

W. C. Barker, J. S. Garavelli, P. B. McGarvey, C. R. Marzec, B. C. Orcutt, G. Y. Srinivasarao, L.-S. L. Yeh, R. S. Ledley, H.-W. Mewes, F. Pfeiffer, A. Tsugita, and C. Wu. The PIR-International Protein Sequence Database. *Nucleic Acids Research*, 27(1):39–43, 1998.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

M. J. Betts and R. B. Russell. *Bioinformatics for Geneticists*. Wiley, 2003.

M. Bhagwat and L. Aravind. *Comparative Genomics*. Humana Press, 2008.

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

J. Boberg, T. Salakoski, and M. Vihinen. Selection of a representative set of structures from brookhaven protein data bank. *Proteins: Structure, Function, and Bioinformatics*, 14(2): 265–276, 1992.

W. Bode, P. Schwager, and R. Huber. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. the refined crystal structures of the bovine 96 trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution. *Journal of Molecular Biology*, 118(1):99–112, 1978.

L. Bordoli, F. Kiefer, and T. Schwede, 2006. URL `http://predictioncenter.org/casp7/meeting/presentations/Presentations_assessors/CASP7_DR_Bordoli.pdf`.

L. Bordoli, F. Kiefer, and T. Schwede. Assessment of disorder predictions in CASP7. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):129–136, 2007.

H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, T. Oldfield, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, J. Swaminathan, M. Tagari, J. Tate, S. Tromm, S. Velankar, and W. Vranken. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Research*, 31 (1):458–462, 2003.

C. Boutsidis, M. W. Mahoney, and P. Drineas. Unsupervised feature selection for principal components analysis. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–69, New York, NY, USA, 2008. ACM.

L. Breiman. Random Forests. URL `http://oz.berkeley.edu/users/breiman/randomforest2001.pdf`. 2001a.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001b.

K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones. Protein structure prediction servers at University College London. *Nucleic Acids Research*, 33(Web Server Issue):36–38, 2005.

W. S. Bu, Z. P. Feng, Z. Zhang, and C. T. Zhang. Prediction of protein (domain) structural classes based on amino-acid index. *European Journal of Biochemistry*, 266(3):1043–1049, 1999.

P. D. Cary, T. Moss, and E. M. Bradbury. High-resolution proton-magnetic-resonance studies of chromatin core particles. *European Journal of Biochemistry*, 89(2):475–482, 1978.

J. M. Chandonia, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. Astral compendium enhancements. *Nucleic Acids Research*, 30(1):260–263, 2002.

D. T.-H. Chang, H.-Y. Huang, Y.-T. Syu, and C.-P. Wu. Real value prediction of protein solvent accessibility using enhanced PSSM features. *BMC Bioinformatics*, 9(Suppl 12), 2008.

H. Chen, F. Gu, and F. Liu. Predicting protein secondary structure using continuous wavelet transform and Chou-Fasman method. In *Proceedings of 27th annual international conference of the IEEE Engineering in Medicine and Biology Society*, pages 2603–2606, 2006a.

J. W. Chen, P. Romero, V. N. Uversky, and A. K. Dunker. Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. *Journal of Proteome Research*, 5(4):888–898, 2006b.

J. Cheng, A. Randall, M. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(Web Server Issue):72–76, 2005a.

J. Cheng, M. Sweredoski, and P. Baldi. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery*, 11(3):213–222, 2005b.

D. P. Clark. *Molecular Biology*. Academic Press, 2005.

K. Coeytaux and A. Poupon. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, 21(9):1891–1900, 2005.

T. U. Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 35 (Database Issue):193–197, 2007.

L. L. Conte, S. E. Brenneri, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 30(1):264–267, 2002.

G. M. Cooper and R. E. Hausman. *The Cell: A Molecular Approach*. Sinauer Associates, third edition, 2004.

G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker. Natively disordered proteins. In J. Buchner and T. Kiefhaber, editors, *Protein Folding Handbook*, chapter 8, pages 275–357. Wiley InterScience, 2005.

M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

R. Diaz-Uriarte and S. A. de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *Journal of Molecular Biology*, 347(4):827–839, 2005a.

Z. Dosztányi, V. Csizmók, P. Tompa, and I. Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005b.

Z. Dosztanyi, J. Chen, A. K. Dunker, I. S. I, and P. Tompa. Disorder and sequence repeats in hub proteins and their implications for network evolution. *Journal of Proteome Research*, 5(11):2985–2995, 2006.

A. K. Dunker and Z. Obradovic. The protein trinity–linking function and disorder. *Nature Biotechnology*, 19(9):805–806, 2001.

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

A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown. Intrinsic protein disorder in complete genomes. *Genome Informatics(Workshop on Genome Informatics)*, 11:161–171, 2000.

A. K. Dunker, J. D. Lawson, C. J. Brown, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. H. Kang, C. R. Kissinger,

R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, 19(1):26–59, 2001.

A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002a.

A. K. Dunker, C. J. Brown, and Z. Obradovic. Identification and functions of usefully disordered proteins. *Advances in Protein Chemistry*, 62:26–49, 2002b.

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

D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299(5881):371–374, 1982.

D. Ekman, A. K. Björklund, J. Frey-Skött, and A. Elofsson. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *Journal of Molecular Biology*, 348(1):231–243, 2005.

Z. P. Feng and C. T. Zhang. Prediction of membrane protein types based on the hydrophobic index of amino acids. *Journal of protein chemistry*, 19(4):269–275, 2000.

Z.-P. Feng, X. Zhang, P. Han, N. Arora, R. F. Anders, and R. S. Norton. Abundance of intrinsically unstructured proteins in p. falciparum and other apicomplexan parasite proteomes. *Molecular and Biochemical Parasitology*, 150(2):256–267, 2006.

F. Ferron, S. Longhi, B. Canard, and D. Karlin. A practical overview of protein disorder prediction methods. *Proteins*, 65(1):1–14, 2006.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of International Conference on Machine Learning*, pages 148–156, 1996.

O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics*, 22(23):2948–2949, 2006a.

O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Computational Biology*, 2(12):e177, 2006b.

O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov. Prediction of natively unfolded regions in protein chains. *Molecular Biology*, 40(2):341–348, 2006c.

S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya. To be folded or to be unfolded? *Protein Science*, 13(11):2871–2877, 2004.

J. P. Glusker, M. Lewis, and M. Rossi. *Crystal Structure Analysis for Chemists and Biologists*. John Wiley and Sons, 1994.

J. C. Goswami and A. K. Chan. *Fundamentals of Wavelets: Theory, Algorithms, and Applications*. Wiley-Interscience, 1999.

O. Grana, D. Baker, R. M. MacCallum, J. Meiler, M. Punta, B. Rost, M. L. Tress, and A. Valencia. CASP6 assessment of contact prediction. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):214–224, 2005.

B. Gutte and R. B. Merrifield. The total synthesis of an enzyme with ribonuclease a activity. *Journal of the American Chemical Society*, 91:501–502, 1969.

J. Han and M. Kamber. *Data Mining:Concepts and Techniques*. Morgan Kaufmann, first edition, 2000.

P. Han, X. Zhang, R. S. Norton, and Z. Feng. Predicting intrinsically unstructured proteins based on amino acid composition. In *Proceedings of the Fourth Australasian Data Mining Conference (AusDM05)*, pages 131–140, 2005.

P. Han, X. Zhang, R. S. Norton, and Z.-P. Feng. Predicting disordered regions in proteins based on decision trees of reduced amino acid composition. *Journal of Computational Biology*, 13(9):1579–1590, 2006.

P. Han, X. Zhang, R. S. Norton, and Z.-P. Feng. Reducing overfitting in predicting intrinsically unstructured proteins. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 515–522, 2007.

P. Han, X. Zhang, and Z.-P. Feng. Predicting disordered regions in proteins using the profiles of amino acid indices. *BMC Bioinformatics*, 10(Suppl 1), 2009a.

P. Han, X. Zhang, R. S. Norton, and Z.-P. Feng. Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinformatics*, 10(1), 2009b.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

C. Haynes and L. M. Iakoucheva. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic Acids Research*, 34(1):305–312, 2006.

J. Hecker, J. Y. Yang, and J. Cheng. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics*, 9(Suppl 1), 2008.

S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, 23(16): 2046–2053, 2007.

U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3(3):522–524, 1994.

L. Holm and C. Sander. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Research*, 24(1):206–209, 1996.

L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith, and E. J. Ackerman. Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Science*, 10(3):560–571, 2001.

T. Ishida and K. Kinoshita. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research*, 35(Web Server Issue):460–464, 2007.

T. Ishida and K. Kinoshita. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, 24(11):1344–1348, 2008.

T. Ishida, S. Nakamura, and K. Shimizu. Potential for assessing quality of protein structure based on contact number prediction. *Proteins: Structure, Function, and Bioinformatics*, 4(S8):940–947, 2006.

P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35:339–344, 2007.

T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of International Conference on Machine Learning*, pages 143–151, 2003.

D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

D. T. Jones and J. J. Ward. Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Genetics*, 53(Suppl 6):573–578, 2003.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: Amino Acid Index Database. *Nucleic Acids Research*, 27(1):368–369, 1999.

J. Kyte and R. F. Dolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

J. W. Lee, J. B. Lee, M. Park, and S. H. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics Data Analysis*, 48(4):869–885, 2003.

A. M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, 2002.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic. Predicting protein disorder for N-,C-,and internal regions. *Genome Informatics*, 10:30–40, 1999.

K. U. Linderstrom-Lang and J. A. Schellman. Protein structure and enzyme activity. *Reviews of Modern Physics*, 1:443–510, 1959.

R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. Protein disorder prediction: implications for structural proteomics. *Structure*, 11(11):1453–1459, 2003a.

R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, 31(13):3701–3708, 2003b.

P. Liò. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.

P. Liò and M. Vannucci. Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, 16(4):376–382, 2000.

S. Lise and D. T. Jones. Sequence patterns associated with disordered regions in proteins. *Proteins: Structure, Function, and Bioinformatics*, 58(1):144–150, 2004.

J. Liu and B. Rost. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Research*, 31(13):3833–3835, 2003.

J. Liu, H. Tan, and B. Rost. Loopy proteins appear conserved in evolution. *Journal of Molecular Biology*, 332(1):53–64, 2002.

H. Lodish, A. Berk, P. T. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell. *Molecular Cell Biology*. Scientific American Library, fifth edition, 2004.

A. Lupas. Prediction and analysis of coiled-coil structures. *Methods in Enzymology*, 266: 513–525, 1996.

R. M. MacCallum. Order/disorder prediction with self organizing maps. URL `http://www.forcasp.org/paper2127.html`.

A. Maton, J. Hopkins, C. W. McLaughlin, S. Johnson, M. Q. Warner, D. LaHart, and J. D. Wright. *Human Biology and Health*. Prentice Hall, 1993.

D. J. McColl and R. F. Anders. Conservation of structural motifs and antigenic diversity in the plasmodium falciparum merozoite surface protein-3 (msp-3). *Molecular and Biochemical Parasitology*, 90(1):21–31, 1997.

L. J. McGuffin. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, 24(16):1798–1804, 2008.

W. Mendenhall, R. J. Beaver, and B. M. Beaver. *Introduction to Probability and Statistics*. Duxbury Press, 2003.

S. Mika and B. Rost. UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Research*, 31(13):3789–3791, 2003.

M. L. Miller, L. J. Jensen, F. Diella, C. Jorgensen, M. Tinti, L. Li, M. Hsiung, S. A. Parker, J. Bordeaux, T. Sicheritz-Ponten, M. Olhovsky, A. Pasculescu, J. Alexander, S. Knapp, N. Blom, P. Bork, S. Li, G. Cesareni, T. Pawson, B. E. Turk, M. B. Yaffe, S. Brunak, and R. Linding. Linear motif atlas for phosphorylation-dependent signaling. *Science signaling*, 1(35):ra2, 2008.

T. M. Mitchell. *Machine learning.* McGraw-Hill, 1997.

D. W. Mount. *Bioinformatics: Sequence and Genome Analysis.* Cold Spring Harbor Laboratory Press, first edition, 2001.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

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

Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins: Structure, Function, and Bioinformatics*, 61(Suppl 7):176–182, 2005.

M. Ohgushi and A. Wada. 'molten-globule state': a compact form of globular proteins with mobile side-chains. *FEBS Letters*, 164(1):21–24, 1983.

C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6): 1989–2000, 2005a.

C. J. Oldfield, E. L. Ulrich, Y. Cheng, A. K. Dunker, and J. L. Markley. Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins: Structure, Function, and Bioinformatics*, 59(3):444–453, 2005b.

N. Pattabiraman, K. Namboodiri, A. Lowrey, and B. P. Gaber. NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Sequences and Data Analysis*, 3(5):387–405, 1990.

K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic. Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of Bioinformatics and Computational Biology*, 3(1):35–60, 2005.

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

G. Pollastri, D. Przybylski, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47(2):228–235, 2002.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21(16):3435–3438, 2005.

J. Prilusky, O. Noivirt, and K. Levy, 2008. URL `http://predictioncenter.org/casp8/doc/presentations/CASP8_DR_Sussman.pdf`.

R. A. Pullen, J. A. Jenkins, I. J. Tickle, S. P. Wood, and T. L. Blundell. The relation of polypeptide hormone structure and flexibility to receptor binding: the relevance of X-ray studies on insulins, glucagon and human placental lactogen. *Molecular and Cellular Biochemistry*, 8(1):5–20, 1975.

M. Punta and B. Rost. PROFcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–2968, 2005.

Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple source. In *Pacific Symposium on Biocomputing*, volume 10, pages 531–542, 2005.

J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufman Publishers, 1993.

P. Radivojac, Z. Obradovic, C. J. Brown, and A. K. Dunker. Prediction of boundaries between intrinsically ordered and disordered protein regions. In *Pacific Symposium on Biocomputing*, pages 216–227, Lihue, Hawaii, U.S.A, January 2003.

P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker. Protein flexibility and intrinsic disorder. *Protein Science*, 13(1):71–80, 2004.

P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker. Intrinsic disorder and functional proteomics. *Biophysical Journal*, 92(5):1439–1456, 2007.

A. K. Ramani and E. M. Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, 327(1):273–284, 2003.

G. Rätsch, B. Schölkopf, S. Mika, and K. R. Müller. SVM and Boosting: One class. Technical report, GMD FIRST, 2000.

P. Romero, Z. Obradovic, and A. K. Dunker. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics*, 8:110–124, 1997a.

P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, and A. K. Dunker. Identifying disordered regions in proteins from amino acid sequences. In *IEEE International Conference on Neural Networks*, volume 1, pages 90–95, 1997b.

P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot, and A. K. Dunker. Thousands of proteins likely to have long disordered regions. In *Pacific Symposium on Biocomputing*, pages 437–448, 1998.

P. Romero, Z. Obradovic, and A. K. Dunker. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Letters*, 462(3):363–367, 1999.

P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Genetics*, 42(1): 38–48, 2001.

B. Rost. Did evolution leap to create the protein universe? *Current opinion in structural biology*, 12(3):409–416, 2002.

B. Rost. Review: protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134(2–3):204–218, 2001.

B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993a.

B. Rost and C. Sander. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences*, 90(16): 7558–7562, 1993b.

B. Rost, P. Fariselli, and R. Casadio. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science*, 5(8):1704–1718, 1996.

M. A. Rowski and J. M. Bujnicki. GeneSilico protein structure prediction meta-server. *Nucleic Acids Research*, 31(13):3305–3307, 2003.

R. B. Russell and T. J. Gibson. A careful disorderliness in the proteome: Sites for interaction and targets for future therapies. *FEBS Letters*, 582(8):1271–1275, 2008.

A. Savitzky and M. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

A. Schlessinger, J. Liu, and B. Rost. Natively unstructured loops differ from other loops. *PLoS Computational Biology*, 3(7), 2007a.

A. Schlessinger, M. Punta, and B. Rost. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, 23(18):2376–2384, 2007b.

A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost. Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE*, 4(2):e4433, 2009.

O. Schweers, E. Schönbrunn-Hanebeck, A. Marx, and E. Mandelkow. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *The Journal of Biological Chemistry*, 269(39):24290–24297, 1994.

C. E. Shannon. A mathematical theory of communication. *Bell Labs Technical Journal*, 27: 379–423, 1948.

K. Shimizu, Y. Muraoka, S. Hirose, and T. Noguchi. Feature selection based on physicochemical properties of redefined N-term region and C-term regions for predicting disorder. In *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–6, 2005.

K. Shimizu, S. Hirose, and T. Noguchi. 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(17):2337–2338, 2007a.

K. Shimizu, Y. Muraoka, S. Hirose, K. Tomii, and T. Noguchi. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, 8(78), 2007b.

M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker. DisProt: the database of disordered proteins. *Nucleic Acids Research*, 35(Database Issue):786–793, 2007.

N. Siew and D. Fischer. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Structure, Function, and Bioinformatics*, 53(2):241–251, 2003.

J. Sim, S. Y. Kim, and J. Lee. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*, 21(12):2844–2849, 2005.

J. D. Smith, G. Subramanian, B. Gamain, D. I. Baruch, and L. Miller. Classification of adhesive domains in the plasmodium falciparum erythrocyte membrane protein 1 family. *Molecular and Biochemical Parasitology*, 110(2):293–310, 2000.

P. Sprent. *Applied Nonparametric Statistical Methods*. Chapman and Hall, second edition, 1993.

H.-G. Stark. *Wavelets and Signal Processing: An Application-Based Introduction*. Springer, 2005.

R. Thomson and R. Esnouf. Prediction of natively disordered regions in proteins using a bio-basis function neural network. In *Intelligent Data Engineering and Automated Learning*, volume 3177/2004, pages 108–116, 2004.

R. Thomson, T. C. Hodgman, Z. R. Yang, and A. K. Doyle. Characterising proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics*, 19(14): 1741–1747, 2003.

P. Tompa. Prediction of protein disorder. 2008. URL `http://ist.inserm.fr/basisateliers/atel185/Tompa1prediction.pdf`.

P. Tompa. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters*, 579:3346–3354, 2005.

P. Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27(10):527–533, 2002.

V. N. Uversky. Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, 11(4):739–756, 2002a.

V. N. Uversky. What does it mean to be natively unfolded. *European Journal of Biochemistry*, 269(1):2–12, 2002b.

V. N. Uversky, J. R. Gillespie, and A. L. Fink. Why are natively unfolded proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics*, 41(3): 415–427, 2000.

V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18(5): 343–84, 2005.

M. Vihinen, E. Torkkila, and P. Riikonen. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, 19(2):141–149, 1994.

S. Vucetic, C. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. *Proteins: Structure, Function, and Genetics*, 52(4):573–584, 2003.

S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. GSikes, C. D. Newton, and A. K. Dunker. DisProt: A database of protein disorder. *Bioinformatics*, 21(1):137–140, 2005.

A. Vullo, O. Bortolami, G. Pollastri, and S. C. E. Tosatto. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Research*, 34(Web Server Issue):164–168, 2006.

J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337(3):635–645, 2004.

E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Letters*, 576(3): 348–352, 2004.

P. H. Weinreb, W. Zhen, A. W. Poon, K. A. Conway, and J. Peter T. Lansbury. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*, 35(43):13709–13715, 1996.

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

P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, 1999.

B. Wu, T. Abbott, D. Fishman, W. M. G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 19(13):1636–1643, 2003.

S. Wu and Y. Zhang. Lomets: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382, 2007.

Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, 21(16):3369–3376, 2005.

R. K. Young. *Wavelet Theory and its Applications*. Springer, first edition, 1992.

Z. Zhang and W. I. Wood. A profile hidden markov model for signal peptides generated by HMMER. *Bioinformatics*, 19(2):307–308, 2003.