Protein structure prediction

Zinc-binding sites, one-dimensional structure and remote homology

Nanjiang Shu

Doctoral Thesis

Department of Physical, Inorganic, and Structural Chemistry
(Division of Structural Chemistry)
Stockholm University
Stockholm 2010

Doctoral Dissertation 2010

Structural Chemistry
Arrhenius Laboratory
Stockholm University
S-106 91 Stockholm
Sweden

Faculty opponent:

- Professor Geoff Barton, School of Life Sciences, University of Dundee, Dundee, UK

Evaluation committee:

- Professor Elisabeth Sauer-Eriksson, Umeå University, Sweden
- Associate Professor Erik Lindahl, Stockholm University, Sweden
- Professor Hans Ågren, KTH, Sweden
- Reserv: Professor Jozef Kowalewski, Stockholm University, Sweden

ISBN: 978-91-7155-984-5
Printed in Sweden by US-AB

The cover image is modified from Figure 1 and Figure 6 in the thesis

To my beloved family

Abstract

Predicting the three-dimensional (3D) structure of proteins is a central problem in biology. These computationally predicted 3D protein structures have been successfully applied in many fields of biomedicine, e.g. family assignments and drug discovery. The accurate detection of remotely homologous templates is critical for the successful prediction of the 3D structure of proteins. Also, the prediction of one-dimensional (1D) protein structures such as secondary structures and shape strings are useful for predicting the 3D structure of proteins and important for understanding the sequence-structure relationship. In addition, the prediction of the functional sites of proteins, such as metal-binding sites, can not only reveal the important function of proteins (even in the absence of the 3D structure) but also facilitate the prediction of the 3D structure.

Here, three novel methods in the field of protein structure prediction are presented: PREDZINC, a method for predicting zinc-binding sites in proteins; Frag1D, a method for predicting the 1D structure of proteins; and FragMatch, a method for detecting remotely homologous proteins. These methods compete satisfactorily with the best methods previously published and contribute to the task of protein structure prediction.

List of publications

I. Shu, N., Zhou, T. and Hovmöller, S. (2008) Prediction of zinc-binding sites in proteins from sequence, *Bioinformatics*, **24**, 775-782.

II. Zhou, T.*, Shu, N.* and Hovmöller, S. (2009) A Novel method for accurate one-dimensional protein structure prediction based on fragment matching, *Bioinformatics*, btp679. [*Equal contribution]

III. Shu, N., Zhou, T. and Hovmöller, S. (2010) Protein homology detection by profile based fragment matching, in manuscript.

IV. Shu, N., Hovmöller, S. and Zhou, T. (2008) Describing and comparing protein structures using shape strings, *Curr Protein Pept Sci*, **9**, 310-324.

Abbreviations

1D	One-dimensional
3D	Three-dimensional
CASP	Critical Assessment of protein Structure Prediction
CATH	A hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class (C), Architecture (A), Topology (T) and Homologous superfamily (H).
CH	Cysteine or histidine
CHDE	Cysteine, histidine, aspartate or glutamate
CSA	Catalytic Site Atlas
DSSP	Definition of the secondary structure of proteins
HMM	Hidden Markov Model
HSSP	Homology-derived structures of proteins
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PSSM	Position Specific Substitution Matrix
Q3	Overall per-residue accuracy for three-state secondary structure prediction
ROC	Receiver Operating Characteristic
S3	Overall per-residue accuracy for three-state shape string prediction
S8	Overall per-residue accuracy for eight-state shape string prediction
SCOP	Structural Classification of Proteins
SOV	Segment Overlap measure
SSE	Secondary Structure Element
SVM	Support Vector Machine

Contents

1 Introduction

Proteins are essential to life. In bodies, proteins fold into certain three-dimensional (3D) structures, called the native structures. The functions of proteins rely on their native structures. Determining the 3D structure of proteins has become a major task for modern biological research. Protein structures are determined experimentally by X-ray crystallography, NMR spectroscopy and cryo-electro microscopy (cryo-EM). Since the determination of the first protein structure, myoglobin, by Kendrew and his colleagues 50 years ago (Kendrew *et al.*, 1958), the number of experimentally solved protein structures deposited in the Protein Data Bank (PDB, www.pdb.org/pdb) (Berman *et al.*, 2000) has reached 56 951 (as of Nov. 18, 2009; there are also 4626 other biological macromolecular structures such as DNA in the PDB) and this number is still doubling about every three years. However, this exciting number can be disappointing for the biologists who need 3D models of proteins in their research. As of Nov. 2009, there are ~9.7 million protein sequences deposited in the UniProtKB/TrEMBL database (The-UniProt-Consortium, 2009). The chance of a protein sequence to have a solved structure has dropped to 0.6% (56951 / 9700000 * 100%) by Nov. 2009; while this number was 2.1% in Dec. 2004 and 1.6% in Dec. 2007. It has to be noticed that many entries deposited in the PDB are the same proteins but have been solved in different conditions (e.g. different concentrations and different temperatures) for various scientific purposes. For example, 1171 entries in the PDB are structures of lysozyme. When taking this into account, the chance of a protein sequence to have a solved structure is even lower. To narrow the gap between the number of solved sequences and the number of solved structures, efficient and accurate computational prediction methods are highly demanded.

Since Anfinsen beautifully demonstrated that bovine pancreatic ribonuclease could regain its native 3D structure after unfolding (Anfinsen, 1973), it has been believed that the 3D structure of proteins are determined by their amino acid sequences. Numerous methods have been developed for predicting protein structures from amino acid sequences in the past decades. Progress in predicting 3D structures of proteins from amino acids has been shown in the Critical Assessment of protein Structure Prediction (CASP) in recent years (Moult *et al.*, 2003; Moult *et al.*, 2005; Moult *et al.*, 2007; Moult *et al.*, 2009). CASP is a world-wide competition of 3D protein structure predictions held every two years. As revealed by CASP, the accuracy of the predicted structure model mainly relies on the successfulness of the detection of structurally similar templates. It is generally accepted that proteins with similar amino acid sequences are structurally similar. Although exceptions that two proteins (constructed by man) sharing 88% sequence

identity but with totally different fold have also been observed (Alexander *et al.*, 2007), for natural proteins, it is still safe to say that if two proteins share > 30% sequence identity, they are structurally similar. However, to detect the homology between proteins sharing less than 25% sequence identity is a challenge. Therefore, methods which can accurately detect remotely homologous templates become essential. The accuracy of the predicted structural models using *de novo* prediction methods, i.e. predicting 3D structures without any structural template, is still far away from practical requirements. Protein structures in reduced form, e.g. protein secondary structures (represented by H: helix, S: sheet and R: random coil) and shape strings (see Figure **1** for definition) can be predicted at high accuracy. Such predicted one-dimensional (1D) structures can assist the prediction of 3D structures.

Many proteins need to interact with other molecules or ions in order to function properly. Metals are among the most common molecules or ions that interact with proteins. Metal ions are present in about one third of the proteins deposited in the PDB and they play a variety of roles in many biological processes, from structure stabilization to enzyme catalysis. Zinc is the second (only after iron) most abundant transition metal found in eukaryotic organisms (Coleman, 1992). The function of zinc in proteins can generally be divided into two categories: structural and catalytic. An example for the former is that zinc-fingers which comprise the largest class of transcription factors in the human genome are structurally stable only in the presence of zinc (Tupler *et al.*, 2001). An example for the latter is that zinc ions serve as powerful electrophilic catalysts in many hydrolases and lyases (McCall *et al.*, 2000).

A protein might not be of great biological interest unless its function has been annotated. The identification and localization of zinc-binding sites (if they exist) is not only important for functional annotation of many proteins but also helpful to the prediction of 3D protein structures. The accurate prediction of zinc-binding residues in sequences can be used directly to screen zinc-binding proteins in genomes. The predicted zinc-binding proteins can also be used to complement the current metalloprotein or catalytic site database, e.g. MDB (Metalloprotein Database and Browser, metallo.scripps.edu/) and CSA (Catalytic Site Atlas, http://www.ebi.ac.uk/thornton-srv/databases/CSA/). In addition, the accurate prediction of zinc-binding proteins might be used to select protein enzymes capable of catalyzing inorganic reactions and mediating the formation of crystals, which is fundamental in material synthesis (Feldheim and Eaton, 2007). Advances in DNA/protein sequencing techniques and the much slower traditional function annotation methods have led to an increasing gap between the number of functionally uncharacterized protein sequences and

that of well-annotated protein structures. Fast and automatic annotation tools based on computational biology are required.

Here, three methods in the field of protein structure prediction are described: (1) Frag1D: a method for predicting 1D protein structures, including secondary structures and shape strings; (2) FragMatch: a method for detecting remote homologues; and (3) PREDZINC: a method for predicting zinc-binding sites of proteins from amino acid sequences, including the prediction of (i) whether a protein is zinc-binding or not and (ii) residues that bind to zinc. Aside from the above three methods, different methods for describing and comparing protein structures are reviewed, with the emphasis on those methods that represent protein structures as 1D geometrical strings, especially shape strings.

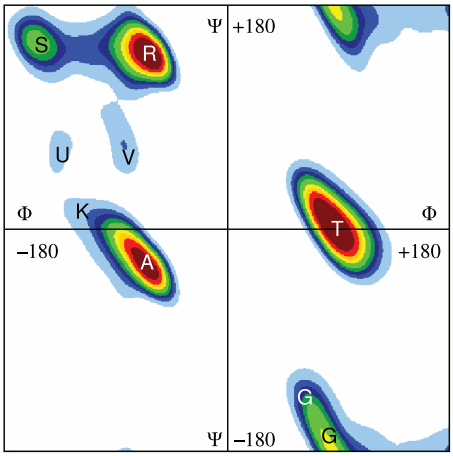


Figure 1: The definitions of eight-state (S, R, U, V, K, A, T and G) shape strings on the Ramachandran plot (Ison *et al.*, 2005). The typical shapes for α-helices and β-sheets are A and S respectively. Shape R represents the so-called polyproline type II structure. Shape K is often found at ends of helices or in 3_{10} helices. T denotes the turn region and G is special for glycine. Three-state shape strings are obtained by mapping S, R, U and V to S, K and A to H, T and G to T. The Ramachandran plot shown here is a montage from two plots; the left part shows the Ramachandran plot for all amino acids found in random coil, while the left half of the figure is that found for all glycine residues. Both are taken from Hovmöller *et al.* (2002). [From Fig. 1 in Paper **II**]

1.1 Background: 3D protein structure prediction

Predicting the 3D structure of proteins from their amino acid sequences has been a major interest for researchers in various disciplines for many years

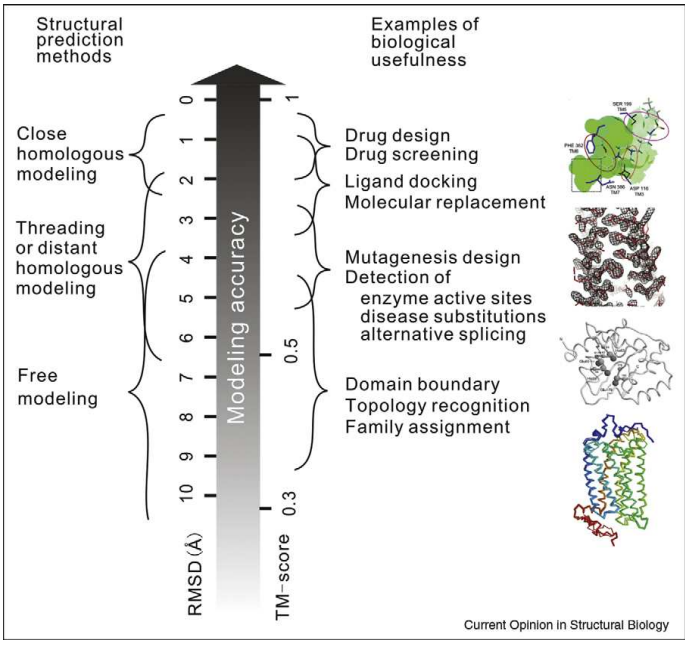
(Lewis and Scheraga, 1971). The purposes of protein structure prediction vary from the high intellectual challenge of elucidating the protein folding process, to diverse applications that might be possible once the accurate 3D structure of proteins can be predicted.

Protein structure prediction is usually divided into three categories: *ab initio* (or *de nono*) prediction, fold recognition (or threading) and homology modelling, based on to which extent the homology information in sequence and structure databases has been used to construct the structural model. *Ab initio* prediction refers to the prediction of protein structures where neither homologues nor fold templates can be found in the PDB. In its purest form, *ab initio* prediction predicts protein structures based entirely on physical and chemical laws, e.g. simulation of folding process using molecular dynamics. However, it often applies to the prediction using only local structure information or starting from a secondary structure prediction. In CASP (Moult *et al.*, 2009) it almost always refers to the latter. Fold recognition means the prediction of protein structures for which only templates that might have similar folds but without obvious homologous to the target can be detected in the PDB. Finally, when a close homologue to the target protein can be identified in the PDB, the prediction is often referred to as homology modelling.

To date the most accurate protein structure prediction methods are still based on homology modelling, although significant progress has been made in fold recognition and *ab initio* prediction, according to the results revealed in recent CASP experiments. The boundary between these three categories is becoming increasingly unclear (Zhang, 2008); nowadays even the prediction tasks classified as *de novo* prediction are usually based on available 3D fragmental structures. However, the accuracy of a predicted model is mainly determined by the availability of templates. For proteins with close homologous templates, the predicted 3D protein structures can be as close as 1-2Å root mean square deviation (RMSD) to their native structures. For proteins having only distantly related templates in the PDB, the predicted structures can be as close as 2-6Å to their native structures. The errors are mainly caused by incorrectly predicted loop regions (Jauch *et al.*, 2007). For proteins without any homologous templates, successful predictions have only been reported for small proteins, with less than 100 residues (Zhang, 2008). The best predicted models can be as close as 4-8Å to the native structures. The relationship among the algorithms, accuracy and biological usefulness of protein structure predictions are illustrated in Figure **2**.

Although there is only one final goal for protein structure prediction, that is, predicting 3D structures from amino acid sequences, sub-problems such as protein secondary structure prediction, protein backbone dihedral angle prediction, homology detection, binding-site prediction and protein-

protein interaction prediction are also of great interest. Some of these problems are inevitable steps in 3D protein structure prediction. Two examples are protein secondary structure prediction for *ab initio* 3D structure prediction (Bonneau and Baker, 2001) and homology detection for homology modelling. Others, e.g. binding-site prediction and protein-protein interaction prediction, are not directly used in 3D structure prediction, but will be of great help to the 3D structure prediction if they can be accurately predicted. In addition, the prediction of binding-sites and protein-protein interactions can immediately be applied in functional annotation and protein design, even without the knowledge of the full 3D structure (Laurie and Jackson, 2006).



TM-score: Template modelling (TM) score, a scoring function for assessing the quality of protein structure templates and predicted structural model (Zhang and Skolnick, 2004).

Figure 2: Approximate relationship among the algorithms, accuracy and biological usefulness of protein structure predictions. [Reproduced from Zhang, (2009) with permission]

1.2 Predicting 1D protein structures

Predicting the secondary structure of proteins has long been considered as an important stage for 3D structure prediction. Since the first protein structures were solved by X-ray crystallography, attempts have been made to predict

the secondary structure of proteins as α-helix, β-sheet and random coil from their amino acid sequences. Chou and Fasman (1974) pioneered the secondary structure prediction based on simple statistics of the probabilities of each individual amino acid appearing at each of the three states, namely H (helix), S (sheet) and R (random coil). Although Chou and Fasman claimed nearly 80% Q3 (overall three-state per-residue accuracy) in their original work when tested on 19 proteins available at that time, it has been proved that the Chou and Fasman method predicts protein secondary structures only at 50-60% accuracy (Kabsch and Sander, 1983b). The over-optimistic results reported in the original work of Chou and Fasman is caused by the very small and non-representative dataset they used (due to the limited number of protein structures that were solved at that time). Moreover, Chou and Fassman failed to separate the training set and the test set. Later on, by using the propensities for segments of 3-51 adjacent residues, the Q3 accuracy of the secondary structure prediction was improved steadily to above 60% (Deleage and Roux, 1987; Holley and Karplus, 1989; Kneller *et al.*, 1990; Muggleton *et al.*, 1992; Presnell *et al.*, 1992). The breakthrough of the third-generation secondary structure prediction was made by using the evolutionary information and advanced algorithms such as neural networks: Q3 was improved to over 70% (Rost and Sander, 1993). The evolutionary information was derived from the divergence of amino acids among homologous proteins to the protein to be predicted. With the emergence of new sequence database searching tools such as Hidden Markov Models (HMM) (Eddy, 1998) and PSI-BLAST (Altschul *et al.*, 1997), large-scale real-time database searching became feasible. Consequently, reliable profiles (see Figure **25** in Appendix 3 for an example) built from large sequence families became achievable. By using PSI-BLAST to build profiles, David Jones made a big step forward in secondary structure prediction: a Q3 of 76.5% was obtained when tested on 187 unique folds (Jones, 1999b). Recently developed methods (Wood and Hirst, 2004; Dor and Zhou, 2007; Homaeian *et al.*, 2007) are almost without exception based on sequence profiles generated by PSI-BLAST. The Q3 for those methods is approaching 80% and slightly better result may be obtained by combining several of these methods (Cheng *et al.*, 2007). Rost *et al.* proposed that the upper limit of the secondary structure prediction is 88% Q3 by analyzing the structural divergence among homologous proteins (Rost *et al.*, 1994). The accuracy of recently developed secondary structure prediction methods is approaching this proposed upper limit but there is still a long way to go, since every 1% step forward is becoming more difficult as it is approaching the upper limit.

The accurate prediction of secondary structure can improve the sensitivity of threading methods (Jones, 1999a) and is critical to many *de novo* structure prediction methods (Bradley *et al.*, 2003). However, for on average ~40% of all residues in random coils, the classical secondary

structure representation carries no structural information. On the other hand, the backbone protein structure is precisely described by a series of torsion angle pairs (φ, ψ), one pair for each residue, due to the planarity of the peptide bond. The torsion angle pairs of native protein structures are actually clustered into distinct regions. Therefore the backbone protein structure can be rather accurately described by a 1D string of symbols representing the clustered regions of φ/ψ torsion angle pairs, called shape strings (Ison *et al.*, 2005) (see Figure **1** for definition). Shape strings describe the conformations of residues in regular secondary structure elements (SSE), e.g. shape A corresponds to the regular α-helix (centered at φ = –61°, ψ = –41° on the Ramachandran plot) and shape S corresponds to the regular β-sheet (centered at φ = –116°, ψ = 128° on the Ramachandran plot) (Hovmöller *et al.*, 2002). Shape strings also classify residues in random coils into several states, thus containing much richer conformation information. It has been shown that shape strings can be used for efficient searching for similar structures in a database (Paper **IV**) and the precise backbone structure can be reconstructed from shape strings (Gong *et al.*, 2005; Ison *et al.*, 2005). Only recently, attempts to predict also the conformation of the protein backbone in segments of random coil have been made. Bystroff *et al.* predicted 11-state shape strings with an overall MDA score (Bystroff *et al.*, 2000) of 58.8%, using a Hidden Markov Model. The MDA score is defined as the fraction of residues that are found in predicted eight-residue segments in which no predicted φ/ψ angles differ by more than 120° from the true structure. Kuang *et al.* predicted three-state shape strings with overall per-residue accuracy (S3) of 79.5% and for four-state shape strings, 78.4%, using Support Vector Machines (SVM) (Kuang *et al.*, 2004). Our method, Frag1D predicted the three-state shape strings at 81.7% S3, i.e. 2.2% better than that of Kuang's method (Paper **II**), using the same shape string definition as in Kuang's work. Note that slightly different definitions on how to discretize clustered regions of φ/ψ angle pairs on the Ramachandran plot have been used for these works (see the comparison of different definitions in Paper **IV,** Fig. 6). The baseline for the three-state shape string prediction is higher than that for the three-state secondary structure prediction. The average abundances of the three secondary structure states H, S and R are 38.1%, 21.7% and 40.3% respectively (see Table **5** in section 2.3.1). Therefore, a random guess of the secondary structure, given the condition that the proportions must be correct, will yield Q3 = $(0.381^2 + 0.217^2 + 0.403^2)$ = 35.5%. For three-state shape strings, the average compositions for H (A+K), S (S+R+U+V) and T (T+G) are 51.7%, 42.6% and 5.7% respectively (Table **5**), and thus the S3 of a random guess is $(0.517^2 + 0.426^2 + 0.057^2)$ = 45.2%. Nevertheless, even the best result reported for the three-state shape string prediction is 79.5% (Kuang *et al.*, 2004), at the same level as the secondary structure prediction. More accurate methods for predicting shape strings are required.

1.3 Detecting remote homologues

The function and structures of unknown sequences can often be accurately inferred if one can map the uncharacterized sequence to a well-annotated protein or protein family. This mapping procedure requires the detection of evolutionary relationship, or homology, between proteins. As mentioned at the beginning of the Introduction, presently the chance of a protein sequence to have a solved structure is only 0.6%, due to the big gap between the number of solved sequences and solved structures. However, recent analyses show that the coverage of existing protein folds represented by the solved 3D protein structures in the PDB is close to completion (Zhang *et al.*, 2006; Qi *et al.*, 2007). This means that for any new protein, it is likely to have a homologue with a solved 3D structure already in the PDB, and the structure of this new protein is similar to the solved one. This presents a challenge to computational biologists, that is, to find a method which can detect homologues for a given protein, if it exists. Around 1990, homology detection methods such as FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990) were developed using pairwise comparison of protein sequences with sequence and position independent substitution matrices, e.g. PAM (Dayhoff *et al.*, 1978) and BLOSUM (Henikoff and Henikoff, 1992). Brenner *et al.* (1998) once showed that sequence-sequence methods such as BLAST can detect most homologues with > 30% sequence identity to a target sequence. However, structural classifications of proteins as done in SCOP and CATH (see section 1.5 for more details) show that also proteins sharing very low sequence identities (10-20%) may still be homologues. For example, two forms of the protein triosephosphate isomerase in the PDB, 1HG3_A (from *Pyrococcus woesei*) and 1TRE_A (from *Escherichia coli*), share only 18% sequence identity but both belong to the SCOP family triosephosphate isomerase (see Fig. 1 in Paper **IV**). To detect the homology between those proteins sharing low sequence identity, i.e. to detect distantly related homologues, is still a challenge.

By comparing protein sequences to position specific substitution matrices (PSSM, also termed as profiles), methods such as PSI-BLAST (Altschul *et al.*, 1997), HMMer (Eddy, 1998) and SAM (Karplus *et al.*, 1998) are able to detect more remotely homologous proteins. Bussiere *et al*. (1998) showed that profile-sequence methods could detect three times as many homologues as the traditional sequence-sequence methods when the sequence identity was below 30%. Profiles are built by multiple sequence alignments among homologous proteins to the target (see Figure **25** in Appendix 3 for an example). PSI-BLAST automated the profile building together with the large-scale sequence database searching in a very efficient way and is thus widely used by biologists. Even more sensitive methods, e.g. PROF_SIM (Yona and Levitt, 2002), PRC (Madera and Gough, 2002),

COMPASS (Sadreyev and Grishin, 2003) and HHsearch (Soding, 2005), were developed by comparing profiles to profiles, which means that the position specific evolutionary information is used for both the query and the target sequence. Some of these methods employed HMM to model the sequence profile. Consequently, the sequence-profile and profile-profile methods became sequence-HMM methods (e.g. HMMer and SAM) and HMM-HMM methods (e.g. RPC and HHsearch). These HMM models are similar to normal sequence profiles but they contain the position-specific probabilities for insertions and deletions along the alignment, in addition to the amino acid frequencies in the columns of the multiple sequence alignment (Eddy, 1998). Some recent studies show also that improved sensitivity can be achieved by incorporating predicted secondary structures into profiles or HMMs (Soding, 2005; Wang *et al.*, 2009). At the same time, methods using supervised machine learning algorithms such as SVM were developed for remote homology detection and protein family classification. Such methods include SVM-pairwise (Liao and Noble, 2003), the Fisher-kernel (Jaakkola *et al.*, 2000), the mismatch kernel (Leslie *et al.*, 2004) and SW-PSSM (Rangwala and Karypis, 2005). By taking the advantages of SVM in binary classification, these methods are very accurate in distinguishing positive examples (homologues) and negative examples (non-homologues) when trained on a large dataset containing both positive examples and negative examples. Therefore, these machine learning based methods are extremely suitable for protein family classifications, albeit at the cost of computational time.

1.4 Predicting metal-binding sites in proteins

Due to the abundance and importance of metal-binding sites in proteins, many researchers have endeavoured in developing methods for predicting these metal-binding sites based on structures or amino acid sequences (Gregory *et al.*, 1993; Nakata, 1995; Andreini *et al.*, 2004; Sodhi *et al.*, 2004; Lin *et al.*, 2005; Schymkowitz *et al.*, 2005; Menchetti *et al.*, 2006; Passerini *et al.*, 2007; Ebert and Altman, 2008). However, even given the 3D structure, the detection of binding sites solely from geometric criteria in proteins without bound metal (e.g. apoproteins) is difficult, since the residues that bind to a metal often undergo conformational changes upon binding (Babor *et al.*, 2005). Therefore, structure based metal-binding prediction methods often employ sequence profiles (Sodhi *et al.*, 2004; Ebert and Altman, 2008) derived from multiple sequence alignments, due to the fact that metal-binding sites are often highly conserved (Ouzounis *et al.*, 1998).

It has been noticed that special sequence patterns exist among functional metal-binding sites, for example, the C2H2 zinc-finger motif (one of the

most ubiquitous zinc-binding motifs). Such motifs are now deposited in databases such as PROSITE (Hulo *et al.*, 2004), either as sequence patterns or as matrices. An example is a sequence pattern for the zinc finger C2H2 type domain signature, represented as C-*x*(2,4)-C-*x*(3)-[LIVMFYWC]-*x*(8)-H-*x*(3,5)-H. This pattern means a Cys (cysteine) followed by 2 to 4 residues of any type, followed by a Cys and then by exactly 3 residues of any type, followed by one residue of either lysine, isoleucine, valine, methionine, phenylalanine, tyrosine, tryptophan or cysteine, followed by exactly eight residues of any type and then a His (histidine) followed by 3 to 5 residues of any type and finally ended by a histidine. Alternatively, such information may also be represented as matrices. These patterns and matrices in PROSITE are generated from multiple sequence alignments of homologous motifs and are very sensitive for identifying metal-binding sites. However, the coverage of the PROSITE patterns is low. Take zinc-binding proteins for example: when searching in a non-redundant set of PDB chains containing 2727 chains, only ~29% of all zinc-binding chains (210 chains binding to biologically important zinc, see Table **3** in section 2.2.1) can be detected using the current PROSITE database (version 20.27, Feb. 26, 2008) by the program 'ps_scan' (version 1.57) (Gattiker *et al.*, 2002), although the precision is as high as 90%. Note that all these 2727 protein chains have been used to build this version of the PROSITE database and thus an over-fitting might have occurred. The actual accuracy of the PROSITE motif searching for zinc-binding proteins might be lower than reported here.

Methods for predicting zinc- and other metal-binding sites from sequence alone have received attention recently thanks to the large and increasing number of high-resolution protein structures in the PDB, advances in the machine learning methods such as neural networks (Lawrence, 1994) and SVM (Vapnik, 2000), as well as the availability of PSI-BLAST (Altschul *et al.*, 1997) which enables the creation of reliable sequence profiles. The two amino acids most frequently binding to zinc are Cys and His (see also Table **3).** Menchetti *et al.* (2006) and Passerini *et al.* (2007) predicted zinc-binding Cys and His by a local predictor and a gated predictor based on SVM. They observed that residues that bind to a zinc atom tend to be close in sequence. Based on this observation, they selected preliminary zinc-binding residue candidates with a semi-pattern [CH]*x*(0–7)[CH] (C is cysteine, H is histidine and CH stands for cysteine or histidine, x(0–7) stands for a consecutive substring of any amino acid with a length from 0 to 7). These selected residue pairs were encoded into feature vectors by PSSMs and SVM were then applied to distinguish zinc-binding residues from non-zinc-binding residues. Their method predicted zinc-binding Cys and His with 60% precision at 60% recall (see section 3.2 for the description of precision and recall) based on a five-fold cross-validation (Menchetti *et al.*, 2006) (see section 3.1 for the description of cross-validation). For the

less common zinc-binding residues Asp and Glu, the results were less satisfactory. Passerini *et al*. (2006) described a method for predicting metal-binding Cys and His in a generic way, based on a two-stage machine learning approach. The first step was similar to the method used in Menchetti *et al*. (2006), i.e. using SVM to classify feature vectors which encode preliminary selected zinc-biding residue candidates. After that, a three layer bi-directional recurrent neural network (BRNN) was used to further distinguish metal-binding and non metal-binding Cys and His. For zinc-binding Cys and His, SVM-BRNN predicted with 60% precision at 60% recall. Note that in the works of both Menchetti *et al.* (2006) and Passerini *et al*. (2006), positive examples are proteins containing zinc-binding sites and negative examples are non-metalloproteins. The exclusion of non-zinc metalloproteins from the negative examples tends to simplify the zinc-binding prediction, which might yield over-optimistic prediction results as reported in Menchetti *et al.* (2006) and Passerini *et al*. (2006). Although these methods are reasonably successful in locating zinc-binding sites in proteins, higher prediction accuracy is required for accurate functional annotations of vast amounts of uncharacterized protein sequence data.

1.5 Describing and comparing protein structures

Most protein molecules contain thousands or even tens of thousands of atoms. Their structures are so complex that perhaps the only way to comprehensively describe them is by listing the xyz coordinates of all atoms, as is done in the PDB. Close to half of the atoms in proteins are hydrogens (Andersson and Hovmöller, 2000), but for most structures they are not listed in the PDB, because they are very hard to detect by X-ray crystallography. Luckily, most hydrogen atom positions can easily be deduced from geometry. Using such verbose description of protein structures by listing xyz coordinates of all atoms in proteins using thousands of real numbers, although comprehensive, is not only difficult for the human brain to grasp, but also not easy for a computer to carry out large scale comparisons.

The existence of ordered regular conformations in proteins, stabilized by hydrogen bonds, was predicted already in 1951 (Pauling *et al.*). These regular conformations were called secondary structures, namely α-helices and β-sheets. Today, the secondary structure is usually defined by the definition of the secondary structure of proteins (DSSP) (Kabsch and Sander, 1983a). Thus, a protein structure can be described simply as a set of α-helices, β-sheets and with the rest as random coils. The secondary structure description of a protein captures the most important features of the protein in a rather concise way, which allows our human brain to grasp the most essential information of the protein. However, for the remaining part of the

protein structure, which on average comprises ~40% of all amino acids in proteins, the *random coil* in the secondary structure description carries no structural information.

As already mentioned, experimental characterization of proteins is both time-consuming and expensive and thus it is not feasible to study all proteins in all genomes experimentally. As a consequence, the function of an uncharacterized protein is often inferred from a characterized protein by sequence/structure comparison methods. Functional inference based on sequences only, which often refers to homology detection, is fundamental in computational biology due to the massive uncharacterized sequence data as described above. However, for proteins with sequence identity below 25%, the relationship between them can hardly be inferred from pairwise sequence comparison. To improve the homology detection for distantly related proteins, large scale hierarchical structure classification databases, such as SCOP (Murzin *et al.*, 1995) and CATH (Class, Architecture, Topology and Homology superfamily) (Orengo *et al.*, 1997), have been built by comparing all solved protein structures in the PDB.

1.5.1 SCOP (Structural Classification of Proteins)

The SCOP database is a comprehensive classification of all protein structures in the PDB according to structural, functional and evolutionary relationships among proteins. The basic classification unit in SCOP is the *protein domain* and domains are classified hierarchically into *classes*, *folds*, *superfamilies*, *families*, *proteins and species*. Small proteins are usually comprised of a single domain while domains in large proteins are often classified individually. The classification *species* is used to distinguish the structures of the same protein from different organisms.

First, different proteins are grouped into classes, including (1) all alpha, (2) all beta, (3) alpha and beta (α/β, α-helices and β-strands are interspersed), (4) alpha plus beta (α+β, α-helices and β-strands are largely segregated), (5) multi-domain, (6) membrane and cell surface proteins and peptides and (7) small proteins. Other classes (e.g. designed proteins) are also defined in the current SCOP classification. Nevertheless, they are not true classes but merely temporary holders for PDB entries that are useful to keep together. Proteins in each class are further clustered into folds if their major secondary structures are of the same arrangement. Furthermore, proteins are defined to belong to the same superfamily if they do not have a high sequence identity but their structural and functional features indicate a probable common origin. Finally, proteins are defined to belong to the same family if they meet at least one of the following two criteria which imply that they share a common evolutionary origin: (i) their sequence identity is $\geq$ 30%, and (ii) their functions and structures are very similar, even if they do not have a

high sequence identity. Proteins that are within the same superfamily but not belonging to the same family are usually regarded as remote homologues. A topological illustration of the SCOP hierarchical classification is shown in Figure **3**. SCOP is curated manually with visual inspections and structure comparisons by human experts. It has become the gold standard for homology relationships, e.g. for evaluating remote homology detection methods. However, the accuracy brought from the manual verification of human experts has to be sacrificed by the relatively low updating speed. The latest update of SCOP is June 2009, containing 110 800 domains which are clustered in 1195 folds, 1962 superfamilies and 3902 families. This update contains 38 221 PDB entries from before Feb. 23, 2009.

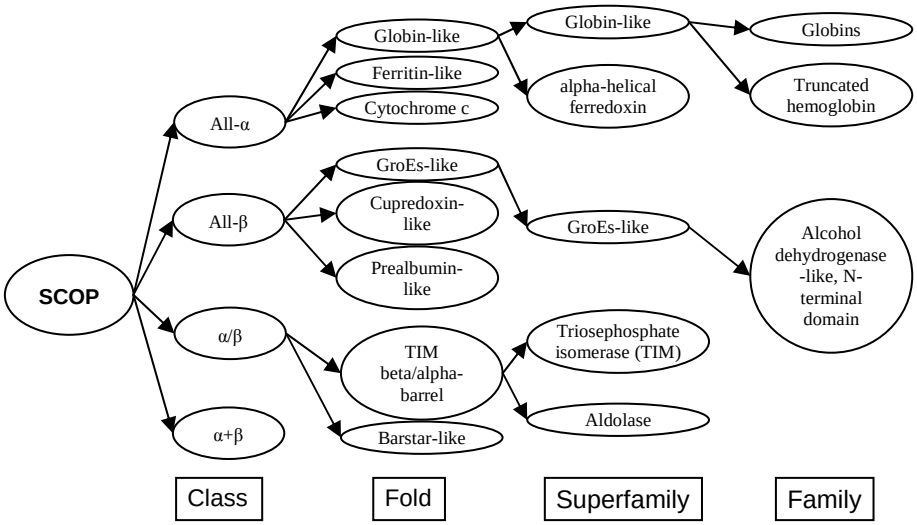


Figure 3: Illustration of part of the hierarchical classification of the SCOP database. There are in total 11 classes (of which 4 are shown here), 1195 folds, 1962 superfamilies and 3902 families in the current SCOP database (version 1.75, June 2009).

1.5.2 CATH (Class, Architecture, Topology and Homologous superfamily)

CATH is a semi-automated hierarchical classification of protein domain structures, in which, protein structures are clustered into four major levels, namely Class (C), Architecture (A), Topology (T) and Homologous superfamily (H), using automated computer programs and supervised by manual inspections. Proteins are automatically clustered into classes according to their secondary structure content. Architecture is currently

assigned manually according to the gross orientation of secondary structures. Furthermore, proteins are clustered into topologies (or folds) according to their topological connections and numbers of secondary structure elements. Finally, proteins with highly similar structures and functions are clustered into homologous superfamilies. The assignments of structures to fold groups and homologous superfamilies are made by sequence and structure comparisons. The latest update of the CATH database is June, 2009, version 3.3, containing 128 688 domains, clustered in 1233 topologies and 2386 homologous superfamilies. This update contains 53 132 PDB entries from before March 02, 2009.

1.5.3 Comparing protein structures

Structure classifications such as SCOP and CATH provide comprehensive descriptions of structural and evolutionary relationships between all proteins with known structures. Due to the fact that structures are more conserved than sequences (Rost, 1997), very distant evolutionary relationships can be revealed by structure comparison methods. Apart from the database building, structure comparison is often required when searching a newly determined structure in the PDB for similar structures, so that more functional annotations can be found. However, comparing protein structures by superposing all atoms of one protein onto the other as rigid bodies is very computationally expensive (even if simplified by superposing Cα atoms only). Still, it is widely used when subtle structural changes need to be detected, for example when a protein loads a ligand. An all-against-all comparison of all proteins in the PDB takes months of computational time. Moreover, rigid body superposition methods often fail to detect the global similarity between proteins with large motions such as hinge-bending (see Fig. 9 in Paper **IV** for an example). Efficient and yet accurate structure comparison methods are required, as more and more structures become available in the PDB. Secondary structure based comparison methods (Orengo *et al.*, 1992; Madej *et al.*, 1995; Kawabata and Nishikawa, 2000; Lu, 2000; Yang and Honig, 2000; Harrison *et al.*, 2003; Krissinel and Henrick, 2004; Vesterstrom and Taylor, 2006) have been introduced to facilitate structure comparison. These methods compare SSEs of protein structures first and then carry out a more careful Cα alignment between pairs of protein molecules [for reviews see Gibrat *et al.* (1996), Carugo and Pongor (2002) and Carugo (2006; 2007)]. In structure database searching, the first step is vital in rapidly eliminating non-similar structures and identifying the structurally similar parts between proteins, since it is this step that enables the efficiency of SSE based structure comparison methods. However, these methods are limited by the inherent drawbacks of the secondary structure description, that is, on average ~40% of amino acids in protein structures are simply classified as random coils (or loop regions) which carry no structural information.

Recent observations of rich and regular structural conformations in loop regions (Oliva *et al.*, 1997) inspired researchers to develop structure comparison methods by representing the backbone structures of proteins as 1D strings of backbone path in the 3D space (Zhi *et al.*, 2006) or shape strings (Ison *et al.*, 2005). A shape string is a 1D geometrical string with each symbol representing a clustered region of φ/ψ torsion angle pairs in the Ramachandran plot (Ramachandran and Sasisekharan, 1968). Although these methods have not been fully developed yet, their advantages in representing the protein structures in loop regions and the rapid database searching for similar structures have already been shown. It is worth to review these methods to arouse attentions from more researchers. It is also necessary to discuss the advantages and disadvantages of these methods compared to rigid body superposition methods and SSE based methods, and to point out the future search directions regarding shape strings. This has been done in Paper **IV** and also in the present summary.

2 Methods and materials

2.1 Properties of shape strings

2.1.1 Definition of shape strings

A shape string is a one dimensional string composed of eight symbols (i.e. A, K, S, R, U, V, T and G, see Figure **1**) which correspond to eight clustered regions of backbone dihedral angles (i.e. φ/ψ angles) in the Ramachandran plot (Ramachandran and Sasisekharan, 1968; Ison *et al.*, 2005).

The planarity of the peptide bond in proteins was noted already in 1951 by Pauling *et al*. As a result of this planarity, the backbone conformation in a polypeptide chain can be described by a pair of torsion angles, φ and ψ, per residue. Thus, the most compact, yet complete, description of the backbone structure needs just two numbers (strictly speaking there is also the ω angle, but it is almost always 180 degrees) per amino acid. In 1963, Ramachandran *et al*. noted that only a few combinations of these torsion angles are possible in proteins. They predicted three commonly allowed regions: α_R, α_L and β, for φ/ψ-angle pairs in the Ramachandran plot, based on the analysis of steric hindrances of short peptides (Figure **4a** and **4b**). Recent studies on the Ramachandran plot by using high-resolution X-ray crystallography protein structures in the PDB, showed that the allowed regions of φ/ψ-angle pairs in the observed plot differ from the original Ramachandran plot (Kleywegt and Jones, 1996; Chakrabarti and Pal, 2001; Hovmöller *et al.*, 2002; Lovell *et al.*, 2003). The first main difference is that α_R, α_L and β-sheet regions are diagonal in the observed Ramachandran plot (Figure **4c** and **4d**) while in the original Ramachandran plot the edges of these regions are mostly parallel to one or both of the φ or ψ axes (Figure **4a** and **4b**). The second is that the β-region is split into two diagonal lobes: the β-sheet region (left) and the polyproline II region (right) (Kleywegt and Jones, 1996; Hovmöller *et al.*, 2002) (Figure **4c**). The third is that the two most populated regions for glycine (Figure **4d**) are in regions predicted to be only permissible in the standard Ramachandran plot. These discrepancies were explained partly in terms of local electrostatic interaction by Ho *et al*. (2003).

Knowing that the allowed combinations of φ/ψ angles in the Ramachandran plot are highly clustered, we can assign a symbol to each cluster in the Ramachandran plot as defined by Figure **1**. The backbone structure of a protein can then be expressed as a 1D string of such symbols (one symbol for each amino acid), i.e. a shape string. Each shape symbol in

the shape string corresponds to a certain region of backbone dihedral angles in the Ramachandran plot. The shape string of an entire protein carries a description of the entire 3D backbone structure. In contrast, the common secondary structure description with only 3 symbols, H (helix), S (sheet) and R (random coil), can describe helices and sheets accurately, but carries no information about the structure of the other 40% of all residues that are in loop regions.

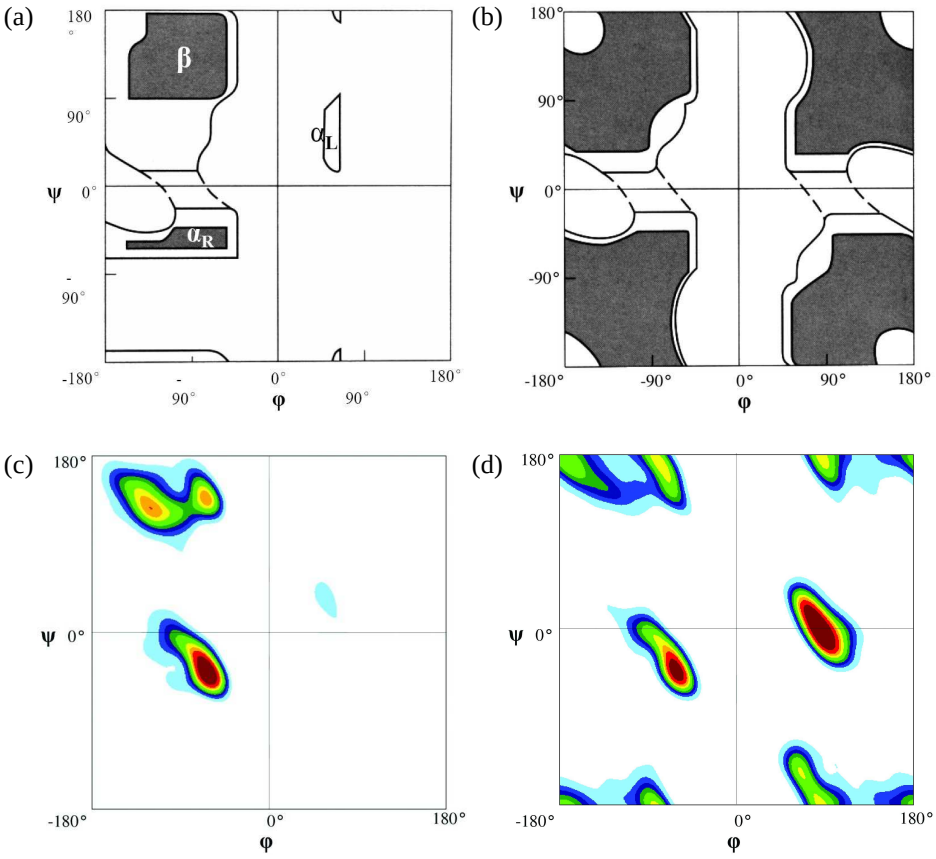


Figure 4: (a) and (b) are the classical Ramachandran plots predicted by Ramachandran and Saisekharan (1968), and (c) and (d) are obtained from high resolution X-ray protein structures from the PDB by Hovmöller *et al.* (2002). (a) is actually modeled for alanine, but often taken as typical for all non-glycines except proline, while (b) is for glycine. (c) is for all 19 non-glycines amino acids and (d) is for glycine. [Reproduced from Hovmöller *et al.*, (2002) with permission]

2.1.2 Statistics on shape strings

Among the eight shapes (Figure **1**), the A shape is the most abundant with ~45% of all residues (Table **1**). This is because almost all residues in α-helices are of the A shape and also a significant part of residues in random coil have the A shape. The second most abundant shape symbol is S which accounts for nearly 25% of all residues, since most residues in β-sheets and some in random coils are of the S shape. The R shape accounts for 16.4% of all residues. It corresponds to the so-called polyproline II region, but it is found also in many slightly distorted β-strands. The name polyproline II has historical roots (Adzhubei and Sternberg, 1993) but does not mean that all or even most of the residues in this region are proline. In fact only 15.8% of the residues in the polyproline II region are prolines (see http://www.fos.su.se/~pdbdna/). The K shape (6.5% of all residues) is typically found as a terminating residue of α-helices. The T shape is the left-handed alpha-helical region α_L and is the most common conformation for glycine (Figure **1**) but is rare for most other amino acids so that in total only 4.5% of all residues have T shape. The shapes U, V and G are less abundant, with 1.2-1.4% each, but they are also very important since they contain extra information in the loop regions which is lacking in the standard secondary structure description. The distribution of shape symbols changes dramatically given the shape symbol of the preceding amino acid (Table **1** and Figure **5**). For example, while in total, the A shape accounts for nearly 45% of all residues, the probability for a residue with A shape following a residue with the A shape is 80% but after an amino acid with the S shape, the probability for being the A shape is only 11%.

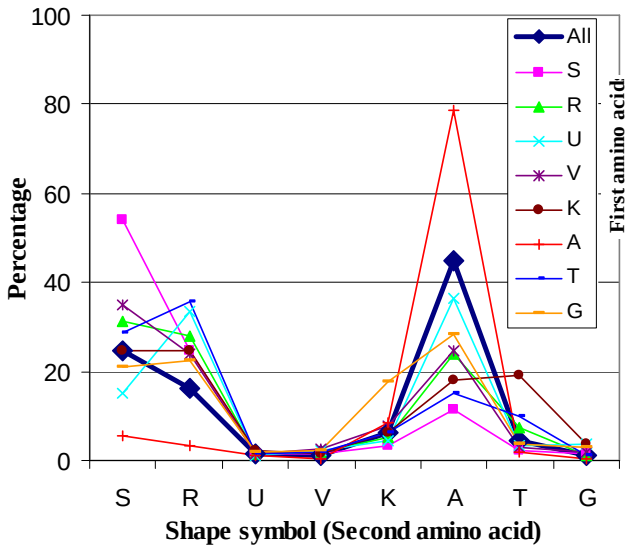


Figure 5: Distribution of eight shape symbols for all residues (represented by All in the legend, which gives out the background composition of shape symbols) and those following a residue with each of the eight shape symbols (S, R, U, V, K, A, T and G in the legend). See also Table **1** for detailed percentage values. [From Paper **IV**, Fig. 4]

Table 1: Percentages for the eight shape symbols for all residues and for those following a residue with each of the eight shape symbols. Complementary to Figure 5. [From Paper **IV**, Table 1]

Second a.a. / First a.a.	S (%)	R (%)	U (%)	V (%)	K (%)	A (%)	T (%)	G (%)
All	24.7	16.2	1.3	1.1	6.4	44.7	4.5	1.2
S	54.2	24.7	1.4	1.4	3.2	11.2	2.3	1.6
R	31.2	27.9	1.5	1.8	5.1	24.0	7.4	1.0
U	15.2	33.6	1.3	2.1	4.4	36.5	3.3	3.7
V	35.0	23.7	1.6	2.5	7.9	24.6	2.8	1.9
K	24.8	24.5	2.3	1.4	6.2	18.2	19.2	3.5
A	5.7	3.1	1.0	0.5	8.4	78.7	2.0	0.6
T	28.9	35.6	1.6	1.8	6.1	15.0	10.1	1.0
G	20.8	22.5	1.8	2.2	17.7	28.4	3.6	3.1

One of the most prominent advantages of shape strings over secondary structures is their ability to describe the detailed conformation in loop regions. Table **2** shows the distribution of short turns connecting two helices,

a helix and a strand, a strand and a helix and two strands. Some shape string fragments, e.g. RAKTR, appear very often, which indicates the existence of characteristic conformations also in loop regions.

Table 2: Frequencies of the shape string fragments of short turns or loops (2 to 5 amino acids long), connecting two helices (H*H), a helix and a strand (H*S), a strand and a helix (S*H), and two strands (S*S), respectively. For each case, the five most frequent shape string fragments are listed. As loops getting longer, there are of course more possible shape string fragments, making each individual shape string fragment less abundant, as seen by low percentages. Note, however, that the shape string fragment RAKTR is very common between two strands. See also Fig. 7 in Paper **IV** for the structural alignment of 313 protein segments each of which contains two strands connected by the five-long turn with the shape string RAKTR. The three-state secondary structure HSR (helix, sheet and random coil) is defined by mapping the eight-state DSSP (Kabsch and Sander, 1983a) definition to HSR with the scheme: H, I and P to H, E to S and the rest to R. Shape strings are defined according to Figure **1**. The existence of the A shape following a helix, e.g. the AS shape string fragment between two helices, are caused by differences in definitions of DSSP and shape strings. The statistics is based on a non-redundant set of PDB containing 4274 protein chains. [Modified from Paper **IV**, Table 2]

	H*H			H*S			S*H			S*S		
Size[a]	Shapes	Count	%	Shapes	Count	%	Shapes	Count	%	Shapes	Count	%
2	RR	551	18.6	RA	316	14.9	AS	438	17.5	TT	1600	38.2
	SR	260	8.8	TR	289	13.7	SR	339	13.6	GK	697	16.7
	KR	221	7.4	SA	215	10.2	RR	289	11.6	AK	251	6.0
	AS	170	5.7	TS	209	9.9	KS	226	9.1	GA	184	4.4
	RS	162	5.5	KT	132	6.2	SS	135	5.4	RT	161	3.9
3	TSR	248	8.4	KTR	410	13.3	SAS	132	6.8	SAK	115	5.6
	KRR	165	5.6	TRA	277	9.0	SKS	67	3.4	RRR	102	4.9
	ARR	156	5.3	KTS	264	8.5	RRR	66	3.4	AKG	91	4.4
	TRR	146	4.9	TSR	190	6.1	RAS	56	2.9	SAA	79	3.8
	KSR	112	3.8	ATR	148	4.8	ASR	54	2.8	ASR	74	3.6
4	KTRR	198	8.4	KTRA	203	7.2	AKRR	37	2.2	AAKT	286	9.1
	KTSR	118	5.0	ATRA	120	4.3	RRSR	34	2.0	AAAT	282	9.0
	ATRR	95	4.0	KTSR	110	3.9	SRRR	22	1.3	ASAK	144	4.6
	KTSS	56	2.4	KTRK	93	3.3	SAAR	21	1.3	RRTR	137	4.4
	KTRS	49	2.1	KTRR	88	3.1	RRRR	20	1.2	AKTR	131	4.2
5	KTASR	60	3.5	RRRTR	38	1.9	RAKRR	33	2.1	RAKTR	492	18.3
	ATASR	33	1.9	KTASA	38	1.9	RAARR	26	1.6	RAATR	153	5.7
	TSRRR	28	1.6	ATASA	32	1.6	RRTRR	23	1.4	RAKTS	97	3.6
	UAARR	18	1.1	RSRTR	30	1.5	SSAAS	15	0.9	SAKTR	76	2.8
	KTKSR	18	1.1	KTRAS	19	1.0	RRTRS	11	0.7	AAATR	67	2.5

[a]Size of the turn in between two secondary structure elements

2.2 Predicting zinc-binding sites in proteins

2.2.1 Statistics and properties of zinc-binding sites in proteins

Most Zn atoms in proteins (78%, see Table **3**) bind to 3 or 4 amino acid residues (called Zn3 and Zn4; Znm refers to Zn atoms coordinated by m amino acid residues), that is, 90% of all zinc-binding Cys (cysteine), His (histidine), Asp (apartate) and Glu (glutamate) are Zn3 or Zn4 binding. Zinc atoms that bind to 4 residues and have no bound water molecules are mostly structural, while those binding to 3 residues are generally catalytic (Auld, 2001). Figure **6** shows an example of a protein, alcohol dehydrogenase, with the PDB code 2OHX (Al-Karadaghi *et al.*, 1994), which contains both a Zn3 and a Zn4 binding site. Many Zn3 and Zn4 atoms have other metal atoms nearby, bridged by a side-chain atom or a water molecule. These bridging metal atoms work together to ensure the protein function. Such Zn atoms are called co-catalytic zinc, according to Auld (2001). Zn atoms that bind to only one or two residues are generally located on the surfaces of proteins. They are most probably bound to proteins during crystallization (McPherson, 1999) but have no biological function. We focused here on predicting biologically important zinc-binding sites, i.e. structural (Zn4), catalytic (Zn3) and co-catalytic zinc-binding sites. Inter-chain Zn atoms, e.g. Zn atoms that bind to two residues in one chain and one residue in another chain, and one Zn5 atom were also included. There were in total 295 biologically bound Zn atoms, binding to 531 Cys, 325 His, 92 Asp and 51 Glu (Table **3**).

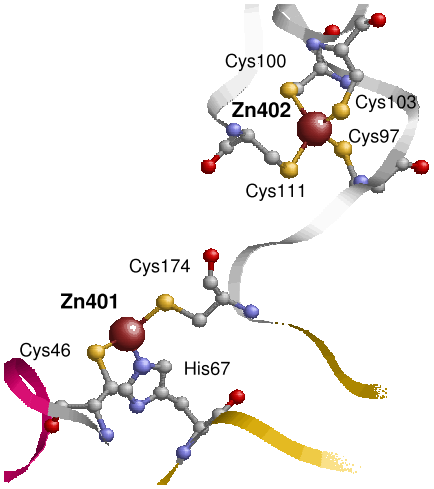


Figure 6: An example of a zinc-binding protein: liver alcohol dehydrogenase [PDB code 2OHX in (Al-Karadaghi, Cedergren and Hovmöller, 1994)]. Zn401 binds to three amino acid residues and is catalytic, whereas Zn402 is fully coordinated by four cysteines and plays a structural role. [From the supplementary data of Paper **I**]

Table 3: Number of residues bound to each type of Zn atom. The statistics are based on a non-redundant set of PDB retrieved by the UniqueProt (Mika and Rost, 2003) program with HSSP (homology derived secondary structure of proteins) distance (see Appendix 1 for definition of HSSP distance) set to zero. This dataset (containing 2727 chains with 564 444 residues) is the same as that used by Passerini *et al.* (2006) for testing the metal-binding site prediction. Among these 2727 chains, 1136 residues were identified binding to 375 zinc atoms. These 1136 residues were distributed in 235 chains (see Paper **I** for details about how zinc-binding residues were identified). The statistics for zinc-binding sites below are also based on this dataset. [Modified from Table 1 in Paper **I**]

	Cys	His	Asp	Glu	Others	Subtotal	No. of Zn atoms	No. of chains
Zn1[a]	1	10	9	10	3	33	34	19
Zn2[a]	3	32	15	26	7	83	45	37
Zn3[a]	25	134	54	30	7	250	89	73
Zn4[a]	499	190	41	24	15	769	205	148
Zn5[a]	7	1	0	0	2	10	2	2
Co-cat Zn[b]	46	59	38	22	10	175	67	35
Subtotal	535	366	116	85	24	1136	375	235
Subtotal[c]	531	325	92	51	24	1023	295	210

[a]Zn1, Zn2, Zn3, Zn4 and Zn5 are Zn atoms binding to 1, 2, 3, 4 and 5 amino acid residues, respectively. [b]Co-catalytic Zn: Zn atoms that bind to 3, 4 or 5 amino acids and are bridged to another metal atom(s) via side chain atoms or water molecules. [c]Subtotal for Zn3, Zn4, Zn5 and co-catalytic Zn.

2.2.1.1 Distribution of zinc-atoms per chain

Most zinc-binding protein chains (88%) contain only one or two Zn atoms (Figure **7**). This is the case for other metals as well, although a few metal-binding proteins are very metal-rich. For example, the protein cyanobacterial photosystem I with the PDB code 1JB0 contains 37 Mg atoms (embedded in ligands Alpha chlorophyll a) and two Fe atoms (embedded in ligands Iron/sulfur cluster) in a single polypeptide chain (the A chain) (Jordan *et al.*, 2001).

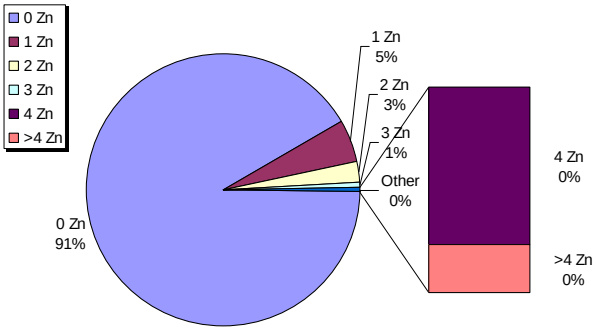


Figure 7: Percentages of chains having 0 Zn atoms, 1 Zn atom, 2 Zn atoms, 3 Zn atoms, 4 Zn atoms and over 4 Zn atoms, based on the dataset mentioned above which contains 2727 chains. [Unpublished results]

2.2.1.2 Distances between zinc-binding residues along the sequence

A protein chain is usually composed of hundreds of amino acids, all linearly connected by peptide bonds. The average length of zinc-binding chains is 219 amino acids (very close to the average length of all 2727 unique chains in our dataset, which is 206). Zinc-binding residues are usually rather closely located in sequence. For most zinc-binding sites, all the 3 or 4 zinc-binding residues are located within 100 residues in sequence (Figure **8**). About 50% of the zinc-binding residues are separated by less than 10 residues. For Zn4, the closest zinc-binding residues are most frequently separated by 2 amino acids and for Zn3, 1 or 3. The average distances of adjacent zinc-binding residues which bind to the same Zn atom are 32 for Zn3 and 22 for Zn4. For most zinc-binding residue groups (residues binding to the same Zn atom), there is at least one pair of residues closer than 10 amino acids, although other residues might be distantly separated. The average distance for the closest pair of each zinc-binding residue group is 11 for Zn3 and 4 for Zn4.

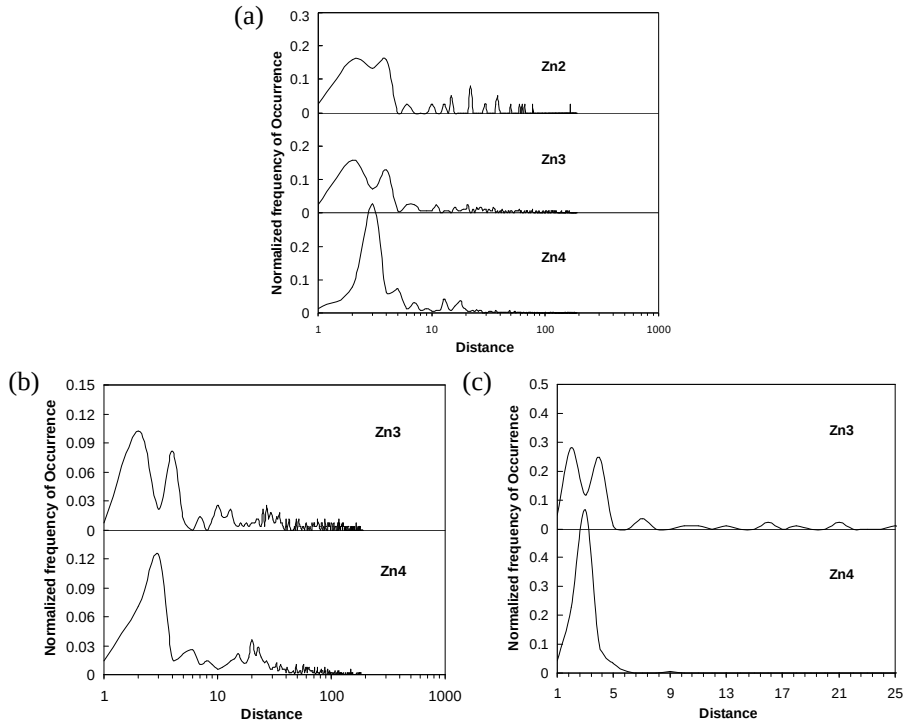


Figure 8: Distance between zinc-binding residues in sequence: (a) Distribution of the distance of adjacent zinc-binding residues binding to the same Zn atom. (b) Distribution of the distance in sequence from the first zinc-binding residue to the following ones that bind to the same Zn atom. (c) Distribution of the distance for the closest residue pair in each zinc-binding residue group (residues binding to the same Zn atom). [Unpublished results]

2.2.1.3 Amino acids at zinc-binding sites and their neighbouring sites

The four residues Cys, His, Asp and Glu constitute ~98% of all residues bound to zinc for Zn3 and Zn4 (Table **3**). His dominates for Zn3 (54% His, 22% Asp, 10% Cys and 12% Glu) and Cys for Zn4 (65% Cys, 25% His, 5.3% Asp and 3.1% Glu). Figure **9** shows clearly the dominance of CHDE (Cys, His, Asp or Glu) at zinc-binding sites. In stark contrast to this, the residues immediately adjacent to the zinc-binding sites show a frequency pattern quite close to the overall frequencies in proteins. It probably indicates that the type of amino acids adjacent to zinc-binding sites has no critical influence on the metal-binding domain. This observation might be important for protein engineering such as the study of zinc-binding site mutations (Windsor *et al.*, 1994). Note also in Figure **9** that His is predominant in Zn3 (catalytic zinc) binding sites, while Cys is preferred in Zn4 (structural zinc) binding sites. This dramatic difference in the preference of the ligand residues for Zn3 and Zn4 binding sites might be employed to distinguish them from each other in prediction.

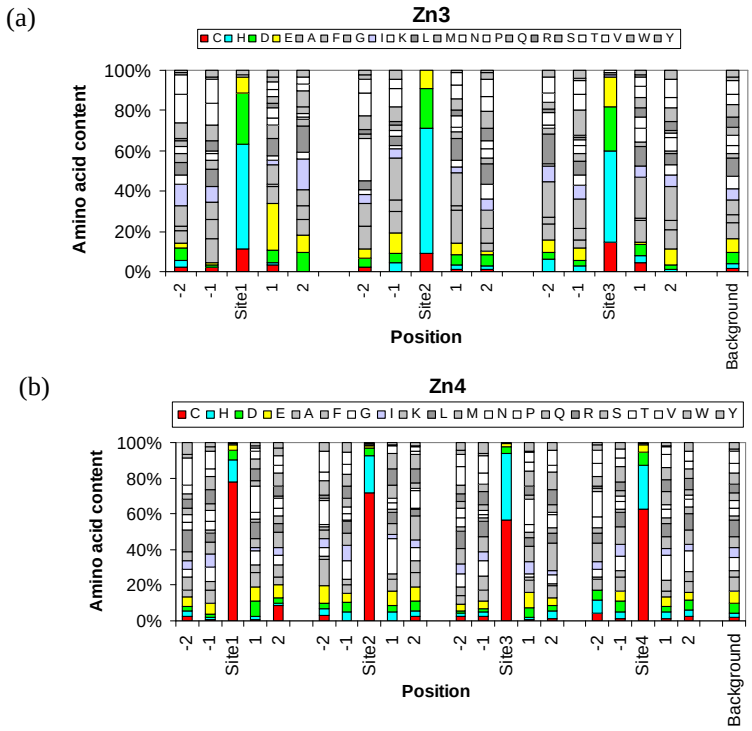


Figure 9: Amino acid content composition at zinc-binding sites and their four nearest adjacent residue positions for (a) Zn3 and (b) Zn4. The background amino acid content composition was estimated by averaging the amino acid compositions in all 2727 protein chains in the dataset mentioned above. If a residue within one of the four nearest adjacent residue positions to a zinc-binding site also binds to zinc, it was not included in calculating the amino acid content composition on that position. [Unpublished results]

2.2.1.4 Conservation level at zinc-binding sites

Ouzounis *et al.* (1998) showed that ligand binding residues are highly conserved (see Paper **I** for definition of the conservation level). This is certainly true for zinc-binding residues. For Zn3 and Zn4, the conservation level of zinc-binding residues is much higher than the background level (Figure **10**). However, for Zn1 and Zn2, their conservation levels are not significantly different from the background. This probably indicates that zinc atoms at these sites are not biologically essential. A more detailed analysis on the Zn3 and Zn4 binding sites and their adjacent residues shows that the conservation levels of residues at zinc-binding sites are dramatically higher than those of their adjacent residues (see the clear peaks in Figure **11)**.

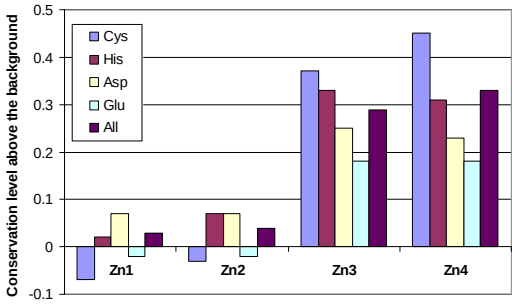


Figure 10: Conservation levels (ranging from 0 to 1) above the background level for residues bound to zinc according to the types of binding sites. The background conservation level for CHDE is 0.48-0.62. It was estimated by averaging the conservation levels of all amino acids of each type for all the 2727 unique chains described above. The conservation level is defined in Paper **I**. [Unpublished results]

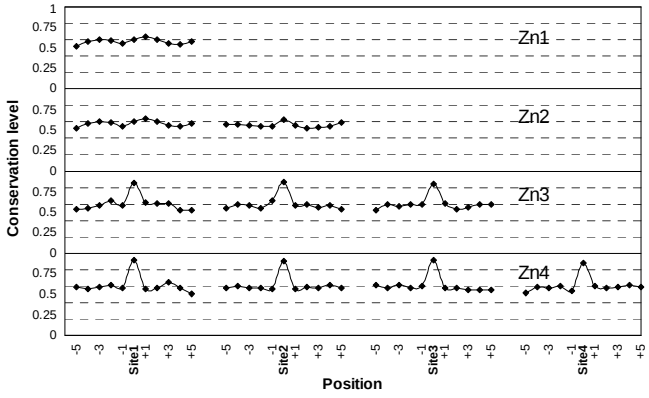


Figure 11: Average conservation levels at zinc-binding sites and their 10 nearest adjacent residue positions. If a residue within the 10 nearest adjacent residue positions to a zinc-binding site also binds to zinc, it is not included in calculating the average conservation level on that position. [Unpublished results]

2.2.2 Method description for PREDZINC

The zinc-binding prediction method consists of an SVM based predictor and a homology-based predictor. In this study, only four types of amino acids, i.e. Cys, His, Asp and Glu were predicted, since these four amino acids comprise ~98% of all residues binding to Zn3 and Zn4. For the SVM based predictor, CHDEs were selected in both the training set and the test set and were encoded into single-site vectors and pair-based vectors (see Appendix 2 for methods to encode single-site vectors and pair-based vectors) which represented a window of residues centered at each selected CHDE or a pair of selected CHDE respectively. The optimized model was learned by training the SVM on the training set and this model was then used by SVM to make the prediction on the test set. The publicly available Gist SVM package (version 2.1.1) (Pavlidis *et al.*, 2004), was used to implement SVM. The kernel was set as radial basis and all other parameters kept at their default values. SVM predictions on individually selected residues were obtained by combining the predictions using single-sites vectors and pair-based vectors with a gating network defined by

$$P(Y_g = 1 \mid f(x)) = P(Y_s = 1 \mid f(x)) + \{1 - P(Y_p = 1 \mid f(x))\} \cdot P(Y_s = 1 \mid f(x)) \qquad (1)$$

where x is the SVM input of each test instance, $f(x)$ denotes the margin of the test instance x, $P(Y_s=1 \mid f(x))$, $P(Y_p = 1 \mid f(x))$ and $P(Y_g = 1 \mid f(x))$ are the probabilities of zinc-binding predictions using single-site vectors, pair-based vectors and the gating network, respectively. For the homology-based predictor, each target chain in the test set was searched in the training set for remote homologues using a segment matching method (see the description of the homology detection method FragMatch). Homology-based predictions of zinc-binding residues were made by mapping the selected CHDE residue groups in the target chain to the binding sites in detected homologues. Finally, SVM predictions and homology-based predictions were combined to reach a consensus. The whole prediction procedure is illustrated in Figure **12**. Details about SVM based predictor and homology based predictor are described in Paper **I**.

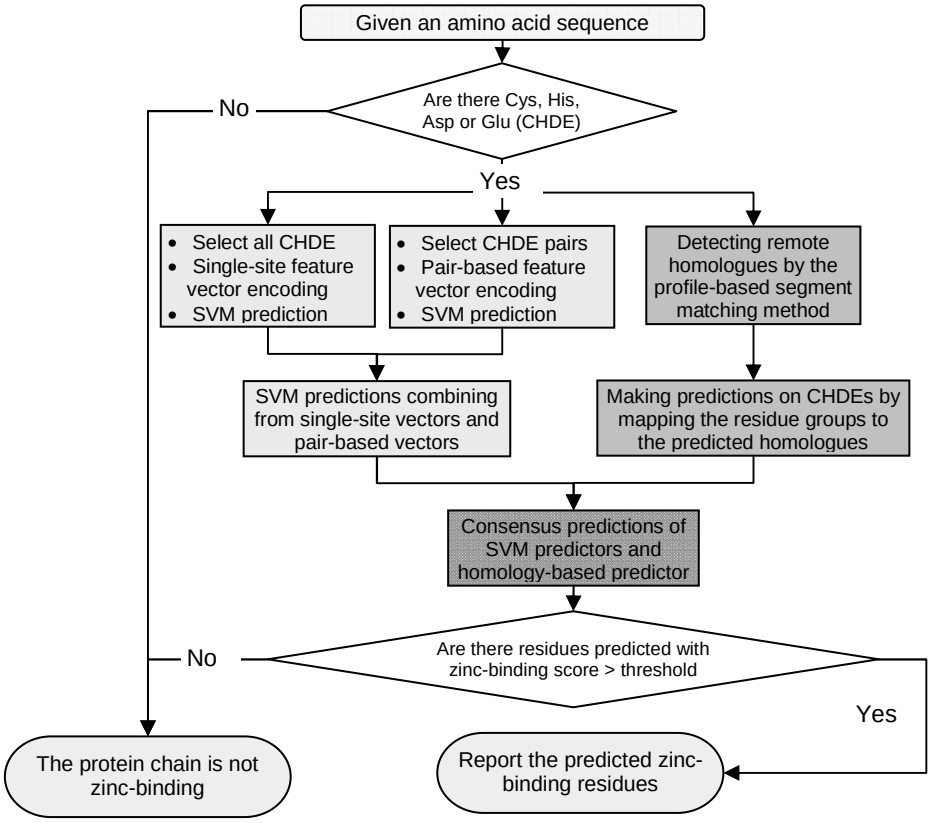


Figure 12: Flowchart for PREDZINC. SVM predictions and homology-based predictions are combined into the final consensus prediction. [Modified from Fig. 1 in Paper **I**]

2.2.2.1 Support Vector Machines (SVM)

SVM (Vapnik, 2000) is a supervised learning algorithm which is efficient in recognizing subtle patterns in large-scale and complex datasets. They have been widely used in different areas of computational biology (Byvatov and Schneider, 2003; Noble, 2004). SVM discriminates two different classes of feature vectors (n-dimensional vectors with numerical values which represent properties of the example) by first mapping the input vectors into a higher dimensional feature space using a kernel function and then doing a linear separation there. A simple case is a binary classification problem on a two-dimensional (2D) space where two sets of dots (square and round) need to be separated (Figure **13**). Square dots belong to class A, labelled as -1, while round dots belong to class B, labelled as +1. There are in principle numerous hyperplanes (lines in 2D space) to separate these two classes of

dots. As shown in Figure **13a**, hyperplane H1, H2 and H3 are successful classifications since they all separate two classes of dots correctly. However, H4 is an unsuccessful hyperplane since it mis-classifies some square dots. Two questions are: (i) which one of the three successful hyperplanes, H1, H2 and H3 is the best and (ii) does an optimal hyperplane exist and how can it be found? In mathematics, a hyperplane can be expressed as

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0 \tag{2}$$

where **w** is the vector normal to the hyperplane, **x** is a dot on the hyperplane and *b* is signed distance from the origin to the hyperplane (Figure **13b**). With this definition, all dots above the hyperplane have $f(\mathbf{x}) > 0$ and those below the hyperplane have $f(\mathbf{x}) < 0$. Therefore, in our example, a successful hyperplane should have negative $f(\mathbf{x})$ values for all square dots and positive values for the round dots.

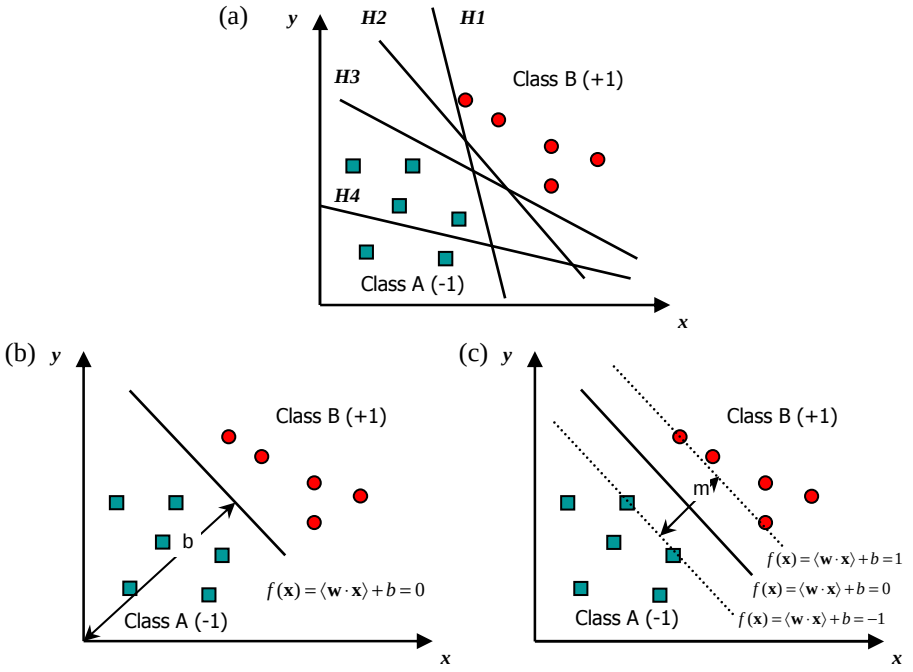


Figure 13: Illustration of binary classification on 2D space: (a) four hyperplanes that separate two classes of dots (square and round) on 2D space. H1, H2 and H3 separate these two classes of dots correctly (of which H2 is the best since it has the maximum margin of the three) while H4 does not, (b) the mathematical expression of a hyperplane, and (c) the separation hyperplane and its two parallel margin hyperplanes which hit the nearest dots on each class to the separation hyperplane.

For an ideal separation hyperplane, we would expect it not only to classify the visible dots (i.e. dots in the training set) but also the potentially added dots (i.e. dots in the test set) correctly. This requires the remaining space between the hyperplane and the nearest dot in class A and class B maximized so that as many dots as possible can be added without breaking the correct separation. In mathematics, this actually requires that the separation hyperplane should be as far away from the data of both classes as possible, or in another way, that the margin as shown in Figure **13c** should be maximized. The margin *m* can be calculated as

$$m = \frac{2}{\| \mathbf{w} \|} \tag{3}$$

where **w** is the vector normal to the hyperplane as described before. In reality, non-linear classification with different kernel functions has been used. Please refer to the work of Cristianini and Shawe-Taylor (2000) and Vapnik (2000) for more descriptions on SVM.

In this study, a feature vector represents the conservativity and physicochemical properties of selected amino acids which are either zinc-binding or not. The publicly available Gist SVM package [version 2.1.1, (Pavlidis *et al.*, 2004)] with the standard radial basis kernel of the form $exp[(-D(x,y)^2)/(2w^2)]$ was used to implement SVM.

2.3 Predicting the 1D structure of proteins

2.3.1 Data description

The dataset used in this study was a non-redundant set of protein chains in the PDB (as of June 2007) culled at 30% sequence identity by the PISCES server (Wang and Dunbrack, 2003), containing 5860 chains (1 480 756 amino acids). The three-state secondary structure (H: helix, S: sheet and R: random coil) of proteins was defined by converting the eight-state DSSP (Kabsch and Sander, 1983a) definition with the classical scheme: H, G and I to H, B and E to S and the rest to R. The eight-state shape string was defined according to Figure **1**. The three-state shape string was transformed from eight-state shape string with the following scheme: S, R, U and V to S, K and A to H, T and G to T. The relation between shape strings and secondary structures from DSSP is shown in Table **4** and Table **5**.

Table 4: The relationship between the eight-state DSSP definition and eight-state shape string definition. All numbers are given in percentages. [From supplementary Table 1 in Paper **II**]

Shape \ DSSP	B	E	G	H	I	S	T	R	Sum
S	0.55	16.54	0.00	0.00	0.00	2.01	0.05	5.21	24.37
R	0.48	4.10	0.04	0.00	0.00	1.63	0.96	8.86	16.07
U	0.04	0.23	0.02	0.01	0.00	0.26	0.09	0.69	1.33
V	0.02	0.21	0.00	0.00	0.00	0.23	0.07	0.61	1.15
K	0.01	0.18	0.80	0.63	0.00	0.91	2.64	1.16	6.33
A	0.00	0.48	2.87	33.94	0.02	2.28	4.71	0.91	45.20
T	0.00	0.10	0.10	0.01	0.00	0.87	2.59	0.71	4.38
G	0.02	0.10	0.06	0.04	0.00	0.32	0.24	0.38	1.16
Sum	1.12	21.94	3.89	34.64	0.02	8.51	11.34	18.54	100

Table 5: The relationship between the three-state DSSP definition and the three-state shape string definition. All numbers are given in percentages. Almost all amino acids in helices or sheets according to the DSSP have the H or S shape, respectively, but the reverse is not true. As many as half of the amino acids with the S shape are actually found in stretches of random coils. [From Table 3 in Paper **II**]

Shape \ DSSP	Helix	Sheet	Random coil	Sum
Shape H (A+K)	37.8	0.7	13.2	51.7
Shape S (S+R+U+V)	0.1	20.8	21.7	42.6
Shape T (T+G)	0.2	0.2	5.3	5.7
Sum	38.1	21.7	40.3	100

2.3.2 Method description for Frag1D

Given a protein sequence to be predicted, a sliding window of N-residue (N varies from 7 to 15, typically 9) long fragment with their respective profiles (see Appendix 3 for how profiles are obtained) of this target sequence, was searched among all N-residue segments in the training set. At each position of a target sequence, the 100 segments with the highest profile-profile scores were kept, together with their accompanying PDB chain IDs and positions in the sequence. The profile-profile score between two compared N-residue segments was defined as

$$Score(\alpha,\beta)=\sum_{n=1}^{N}\left(\sum_{i=1}^{20}(\alpha_{ni}\log(\beta_{ni}/P_i)+\beta_{ni}\log(\alpha_{ni}/P_i))\right) \quad (4)$$

where α and β are profiles for the two compared N-residue segments respectively, N is the window size and P is the background frequency for the

20 standard amino acids. In this study, N was set to 9. This profile-profile score was derived from the PICASSO3 score (Mittelman *et al.*, 2003). After that, the above selected top 100 segments with the highest profile-profile scores were further sorted by the weighted profile-profile score and only the top 10 were kept after re-sorting. The weighted profile-profile score was defined as

$$Score2(\alpha,\beta) = \sum_{n=1}^{N}\left\{ Pinfo_n * \left(\sum_{i=1}^{20} (\alpha_{ni} \log(\beta_{ni} / P_i) + \beta_{ni} \log(\alpha_{ni} / P_i)) \right) \right\} \quad (5)$$

where P$info_{\mathrm{n}}$ is the information score which was defined as

$$Pinfo_n = (1 - \sum_{i=1}^{20} (X_{ni} * X_{ni}) * (1 - \sum_{i=1}^{20} X_{ni} * X_{ni}) \quad (6)$$

where $X_{ni} = (q_{ni} / p_i) / \sum_{i=1}^{20} (q_{\mathrm{ni}}/p_{\mathrm{i}})$, i = 1, 2, 3, …, 20, q_{ni} denotes the probability for amino acid *i* at position *j* in the profile, p_i is the background frequency for amino acid *i*. Equation (6) is empirical; the closer the profile is to the background composition, the larger the P_{info} score is. This score ranges from 0 to 0.90. Score2 [defined by Equation (5)] was assigned to each of these selected segments.

Not all of these 10 selected N-residue segments were used to predict the local structure of the query segment, nor were they used with equal weights. Although the dataset was culled at ≤ 30% (or 25% or 20%) sequence identity, homologues to the target chain may still exist in the training set. These remotely homologous proteins can be accurately predicted by FragMatch (see description of the method FragMatch). The number of segments which were actually used for secondary structure and shape string prediction depended on whether presumed homologues are detected or not for the target chain. If a homologue to the target chain was predicted, only the top 5 segments were used for predicting the secondary structure, since the conformation of the selected segments was believed to be closer to the native conformation of the target protein to be predicted at that position. Otherwise the top 10 were used. Among these 5 or 10 segments actually used for local structure prediction, some may belong to the predicted homologues. Their scores [i.e. Score2 defined by Equation (5)] were multiplied by a factor between 1 and 3 based on the homology score which represented the confidence of the predicted homologues.

The probability for a residue of the target appearing at each state (H, S or R for three-state secondary structures and S, R, U, V, K, A, G or T for eight-state shape strings) was predicted as the sum of weighted scores of all matched segments with the state of the residue aligned at that position

equaling that state. As mentioned above, if there were homologues detected for the target chain, the top 5 candidate fragments for each position were used for prediction; otherwise the top 10 were used. Since a residue in an N-residue target segment may be aligned to at most 9 positions of a candidate segment, there were in total at most either 45 or 90 candidate segments aligned to a target segment depending on whether there were homologues predicted for this target chain or not (see Figure **14** for an example). The state with the highest probability was predicted as the secondary structure or shape string state for that residue. In case of equal probability, the secondary structure was predicted in descending order as R, S and H, and the shape string in the order of G, T, V, U, K, S, R and A. We have noted that S was often under-predicted. In order to remedy this, an empirical 3% probability score was added to the S state. The thus calculated probability for the residue to be predicted on each state was taken as the raw confidence of the prediction. However, the Q3, S3 and S8 were on average 5-10% better than this raw confidence. We thus normalized this raw confidence, such that for a prediction with a given confidence, one might on average expect the Q3, S3 and S8 accuracy to be the same as the confidence. The raw confidence was normalized by a linear function: $y = ax+b$, where x is the raw confidence and y is the normalized confidence. The parameters a and b were obtained by first plotting raw confidence against the real Q3, or S3 or S8, and then making a linear regression (see Figure **15**).

	Target	PFAQAYDS**V**AIRADVEM		PFAQAYDS**V**AIRADVEM	Chain ID
Pos[a]	No.	Candidate secondary structure	Wscore[b]	Candidate sequence	
1	1	HHHHHHHH**H**--------	2.2	PIMQGWDW**F**--------	1Y42X
1	2	HHHHHHHH**H**--------	2.2	PGLQALDE**E**--------	1N3LA
1	3	HHHHHHHH**R**--------	1.8	PAIQAAPS**F**--------	1R6UA
1	4	HHHHHRRS**S**--------	1.0	ELMAAADL**L**--------	2IW1A
1	5	HHHRRRRS**S**--------	1.0	ELLEEYDW**Y**--------	1S4NA
2	1	-HHHHHHH**H**H-------	2.2	-IMQGWDW**F**E-------	1Y42X
2	2	-HHHHHHH**H**H-------	2.2	-GLQALDE**E**Y-------	1N3LA
2	3	-HHHHHHH**R**H-------	1.8	-AIQAAPS**F**S-------	1R6UA
2	4	-HHHHHHH**H**H-------	1.0	-TLQAYDY**L**C-------	1P2XA
2	5	-HHHHHHH**H**H-------	1.0	-MLRAVDR**F**H-------	1YOVA
3	1	--HHHHHH**H**HH------	2.2	--MQGWDW**F**EL------	1Y42X
3	2	--HHHHHH**H**HR------	2.2	--LQALDE**E**YL------	1N3LA
3	3	--HHHHHH**R**HH------	1.8	--IQAAPS**F**SN------	1R6UA
3	4	--RRSSSS**S**SS------	1.0	--LDGARW**F**HF------	2AFBA
3	5	--HHRHHH**H**HH------	1.0	--LNVFEY**Y**SI------	1I5PA
4	1	---HHHHH**H**HHH-----	2.2	---QGWDW**F**ELF-----	1Y42X
4	2	---HHHHH**H**HRR-----	2.2	---QALDE**E**YLK-----	1N3LA
4	3	---HHHHH**R**HHH-----	1.8	---QAAPS**F**SNS-----	1R6UA
4	4	---SSSSS**S**SSS-----	1.0	---AGWDW**I**SAN-----	2ICHA
4	5	---HHHHH**H**HHH-----	1.0	---QAIDL**R**HLE-----	1W27A
5	1	----HHHH**H**HHHH----	2.2	----GWDW**F**ELFY----	1Y42X
5	2	----HHHH**H**HRRR----	2.2	----ALDE**E**YLKV----	1N3LA
5	3	----HHHH**H**RRRR----	2.0	----AADI**L**LYNT----	1I6LA
5	4	----HHHH**R**HHHR----	1.8	----AAPS**F**SNSF----	1R6UA
5	5	----RRRS**S**RRRR----	1.0	----AVDL**I**QIDA----	2HXTA
6	1	-----HHH**H**HRRRR---	2.2	-----LDE**E**YLKVD---	1N3LA
6	2	-----HHH**H**HHHH---	2.2	-----WDW**F**ELFYQ---	1Y42X
6	3	-----HHH**H**RRRR---	2.0	-----ADI**L**LYNTD---	1I6LA
6	4	-----SSSS**S**SSSS---	1.0	-----YDH**V**HVHTD---	2GAGA
6	5	-----SSSS**S**SRRR---	1.0	-----FDV**A**VVDAD---	2AVDA
7	1	------HH**H**HRRRRS--	2.2	------DE**E**YLKVDA--	1N3LA
7	2	------HH**H**HHHHH--	2.2	------DW**F**ELFYQQ--	1Y42X
7	3	------HH**H**RRRRS--	2.0	------DI**L**LYNTDI--	1I6LA
7	4	------RR**S**SSSSR--	1.0	------DT**V**LLQANV--	2FWHA
7	5	------HH**H**HHHHH--	1.0	------DI**I**WLQRDL--	2HBJA
8	1	-------H**H**HRRRRSS-	2.2	-------E**E**YLKVDAQ-	1N3LA
8	2	-------H**H**HRSSSS-	2.2	-------F**Y**QQGVQMQ-	1Y42X
8	3	-------H**H**RRRRSS-	2.0	-------I**L**LYNTDIV-	1I6LA
8	4	-------S**S**SSSSSS-	1.0	-------F**L**LLQMDFG-	2OBDA
8	5	-------S**S**SSSSSS-	1.0	-------F**L**LFGADVV-	1EWFA
9	1	--------**H**HRRRRSSS	2.2	--------**E**YLKVDAQF	1N3LA
9	2	--------**H**HRSSSSS	2.2	--------**Y**QQGVQMQI	1Y42X
9	3	--------**H**RRRRSSS	2.0	--------**L**LYNTDIVP	1I6LA
9	4	--------**H**HRRRSSS	1.0	--------**F**ALMFDQRL	2A7KA
9	5	--------**H**HHHRSSS	1.0	--------**L**ALACDIRV	1HZDA

[a]Pos: position of the fragment that contains the residue of interest [b]Wscore: weighted score of the candidate fragment

For the central residue **V** to be predicted (shown in bold in the target segments in this figure), we find H in 30 cases, S in10 cases and R in 5 cases.
Sum of weighted scores for H = 56.4 (sum of the values in the column 'Wscore' for which the residue in the center of the fragment is in the 'H' state) [Continues to the next page]

Sum of weighted scores for R = 9.0
Sum of weighted scores for S = 10.0 + (56.4+9.0+10.0)*0.03 = 12.3
Total weighted scores = 56.4 + 12.3 + 9.0 = 77.7
Probability of H = 56.6 / 77.7 * 100%= 72.6%
Probability of S = 12.3 / 77.7 * 100% = 15.8%
Probability of R = 9.0 / 77.7 * 100% = 11.6%
The secondary structure of the residue **V** is predicted as H with a raw confidence of 0.726. About 63% of all residues are predicted with higher confidence than this one and 37% with lower confidence. Notice how few (only three, shaded in green) of the used sequences have a Val at the corresponding position.

Figure 14: An example for predicting the secondary structure. The residue VAL in the target chain from PDB entry 1H3F:A (amino acids PRO 176 to MET 192) is to be predicted. More than one homologue was predicted for the target chain and thus only the top 5 segments at each position were used for predicting the secondary structure. [From supplementary Figure 1 in Paper **II**]

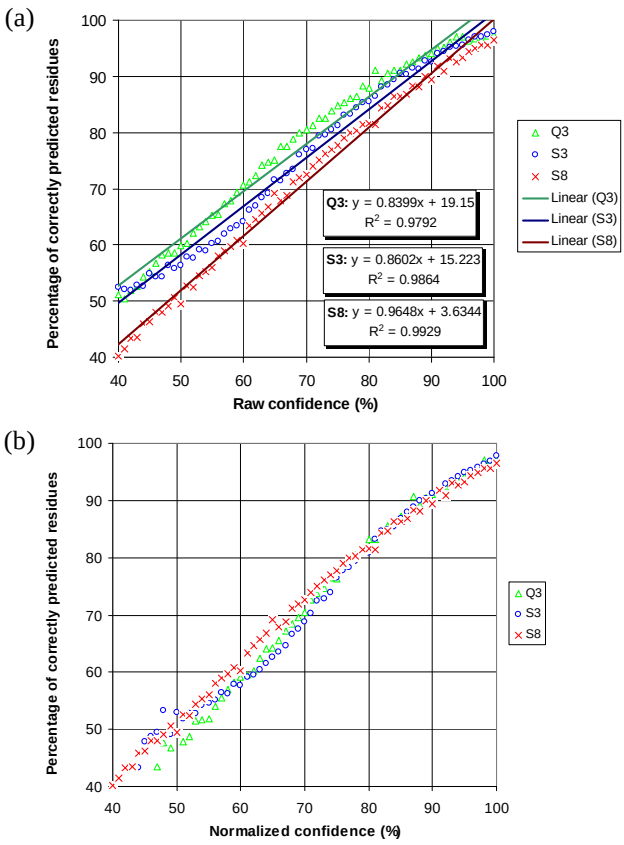


Figure 15: (a) Raw confidence versus the actual percentage of correctly predicted residues for three-state secondary structure, three-state shape strings and eight-state shape strings. (b) Normalized confidence versus the actual Q3, S3 and S8. The normalization functions as shown in the figure (a) were obtained by the linear regression as described in the text. [From supplementary Figure 3 in Paper **II**]

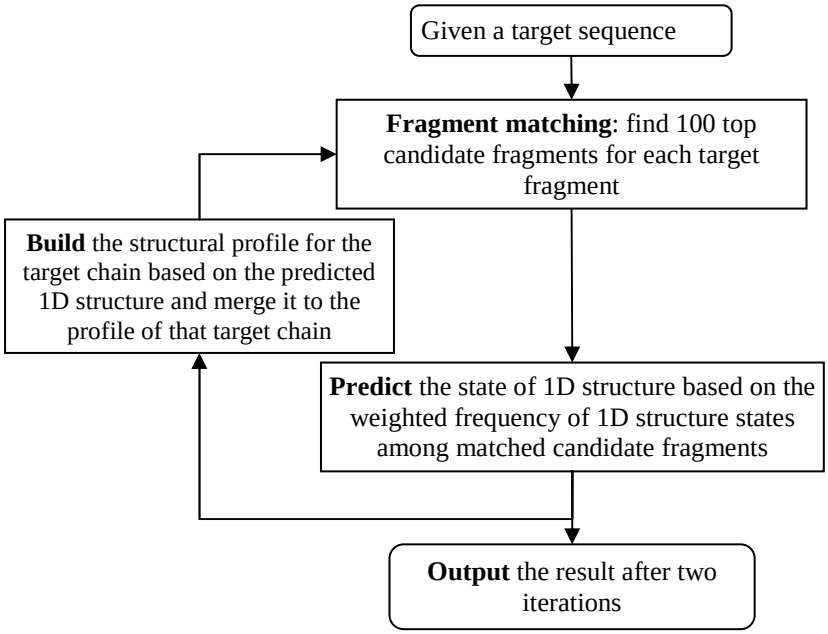


Figure 16: Outline of the 1D structure prediction procedure. [From Fig. 2, Paper **II**]

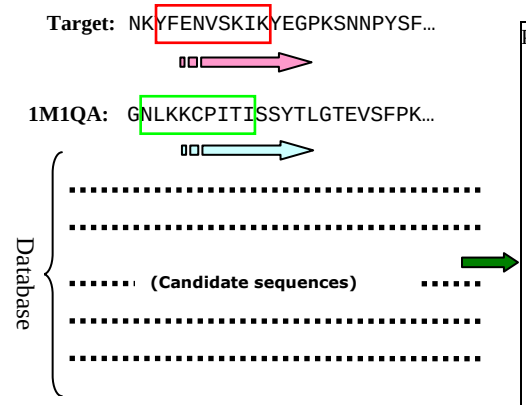
After the prediction was made, the 1D structure of the target sequence was available with an expected high accuracy. Therefore, structural profiles (see Appendix 3) for the target sequence could be built from the predicted 1D structure and then enrich profiles of the target sequence by structural profiles. A second round of prediction was thus carried out with the same setting as the first round, but with enriched profiles for also the target sequence. The whole prediction procedure is outlined in Figure **16**. In principle, this procedure can be iterated many rounds until it converges. However, we noted that Q3 dropped already at the third round. This is most probably because the inaccuracy of structural profiles embedded in the predicted 1D structure accumulates quickly as the iteration procedure progresses and thus the gain by using such structural profiles is soon counteracted by the loss caused by the accumulated inaccuracy. Therefore, the final results were obtained from the second round.

2.4 Detecting remote homologues

2.4.1 Method description for FragMatch

Given a target protein sequence, a sliding N-residue (N ranges from 5 to 17, typically 9 or 11) fragment of that sequence was searched against all such

fragments in the database by the profile-profile score defined by Equation (1) [see Appendix 3 for how profiles were obtained]. For each fragment in the target sequence, up to 150 candidate fragments from difference sequences in the database with highest scores to that target fragment were kept; the PDB chain identifier (CID) of the candidate sequence to which each selected high-scoring fragment belongs was recorded as well (see Figure **17** for a schematic diagram).



Profile example:

Num	AA	A	V	L	I	P	F	M	K	R	H	G	S	T	C	Y	N	E	W	D	Q
4	F	1	2	3	2	0	**73**	1	1	1	1	1	1	1	0	5	1	1	4	1	1
5	E	7	1	3	1	20	1	0	13	2	1	7	4	1	0	0	7	12	0	14	4
6	N	3	1	2	1	2	1	0	5	3	3	20	6	3	0	1	13	8	0	21	8
7	V	4	19	11	46	1	5	1	1	1	0	2	4	1	1	1	1	1	0	1	1
8	S	10	1	1	1	17	1	0	2	1	1	14	7	4	0	0	8	12	0	14	5
9	K	7	3	2	1	11	1	1	**40**	8	1	1	3	6	0	1	2	4	0	1	6
10	I	2	29	6	**50**	1	1	2	1	1	0	1	1	1	0	1	1	1	0	1	1
11	K	6	3	3	2	5	1	1	**21**	15	1	3	5	6	0	1	7	10	0	2	10
12	Y	1	2	3	2	1	22	1	1	1	1	1	1	1	0	**59**	1	1	1	1	1

Equation for calculating the fragmental profile-profile score:

$$Score(A,B) = \sum_{n=1}^{N}\left(\sum_{i=1}^{20}(A_{ni}\log(B_{ni}/P_i) + B_{ni}\log(A_{ni}/P_i))\right)$$

PosSegTar[1]	PosSegCan[2]	Chain ID	Score[3]
…	…	…	…
81	163	2HXWA	123
81	28	2GOMA	121
81	**106**	**2G0WA**	**120**
81	224	2O36A	120
…	…	…	…
82	228	1W5TA	209
82	27	1VQOX	208
82	**107**	**2G0WA**	**208**
82	295	2P4HX	203
…	…	…	…
126	41	2GLFA	708
126	317	2BG5A	705
126	297	1O5KA	703
126	**74**	**2G0WA**	**697**
126	130	2BKWA	697
…	…	…	…

[1]Position of the target segment. [2]Position of the candidate segment. [3]Fragmental profile-profile score calculated according to the equation on the left.

Figure 17: A schematic diagram of the segment matching method. Given a target sequence, a sliding N-residue fragment of that sequence is searched in the database for high scoring (defined by the profile-profile scoring equation) candidate fragments. For each target fragment, up to top 150 high scoring candidate fragments are kept, as shown in the table to the right. The bold lines in this table highlight a frequently appearing candidate chain: 2G0WA. Note that at some places along the sequence, but far from all, the most common amino acid found in related proteins is the same as the amino acid at this position (marked in red and bold in the profile example). [From Figure 1 in Paper **III**]

Some CID tends to appear frequently if they are evolutionary related to the target sequence. We made use of this to carry out the homology detection. For each candidate sequence, if there were candidate fragments from that sequence appearing in the candidate fragment list, a dot plot was drawn by plotting the positions of the fragments in the target sequence against the positions of the matched fragments in that candidate sequence (see Figure **18** for an example).

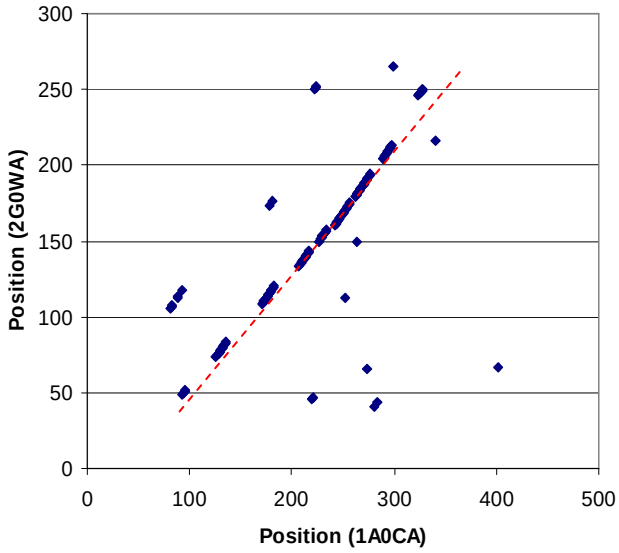


Figure 18: A typical dot plot between a target (1A0CA) and a candidate (2G0WA) sequence where dots form long and consecutive lines. This was obtained by using the protein sequence 1A0CA from the PDB as the target and searching in a non redundant set of PDB (with 5860 chains) cutting at ≤ 30% sequence identity. The dashed red line shows the location of the predicted homology region between 1A0CA and 2G0WA which actually belong to the xylose isomerase-like superfamily (c.1.15) according to SCOP version 1.73. Note that the sequence identity between 1A0CA and 2G0WA is only 16% (obtained by the program 'needle' from EMBOSS version 3.0.0). At this sequence identity level it is very difficult to detect a homologue based purely on comparing amino acid sequences. [From Figure 2 in Paper **III**]

More dots on a dot plot means more high-scoring fragments found between the target sequence and the candidate sequence and thus may indicate a probable homology between these two sequences. However, only when these dots form long and consecutive patterns, it is really a strong indication of the homology between these two sequences. This is similar to the classic dot plot (Maizel and Lenk, 1981) based purely on amino acids, but here we used the profile-profile score instead of the equivalence of

amino acids. A homology score was derived for each candidate sequence from the length and linearity of the pattern and then normalized by the sequence length. The algorithm for calculating the homology score from dot plots has been described in Paper **I** and is also summarized in Figure **19**. Generally speaking, a dot plot with more dots clustered as long, linear and consecutive lines results in a higher homology score.

Algorithm for calculating the homology score from a dot plot

Step 1: Initializing the score on the dot plot:
For each diagonal on the dot plot

1. Set the score of each dot as the number of consecutive dots
2. Record the position of the diagonal with the top 10 longest consecutive dot segments

Step 2: Calculating the score of each diagonal

For each diagonal with the top 10 longest consecutive dot segments

1. The diagonal score is calculated as the sum of the scores of all dots on the diagonal and the scores of all dots on the other diagonals divided by the distance to that diagonal
2. Divide the diagonal score by the average length of the target sequence and candidate sequence

Step 3: Calculating the homology score of the candidate

1. The homology score of the candidate is set as the highest diagonal score

Figure 19: Algorithm for calculating the homology score from dot plots. [From Figure 3 in Paper **III**]

If several related target sequences were available, e.g. when a protein family was used to classify all proteins in a genome, each individual target sequence was first searched in the database, producing a ranking in the same way as for a single target sequence. These rankings of different target sequences were then combined into a consensus ranking by

$$C_i = \left(\frac{\sum_{j=1}^{n} (\alpha_{j,i})^p}{n} \right)^{1/p} \qquad (7)$$

where C_i is the consensus homology score for candidate sequence i, $\alpha_{j,i}$ is the homology score for sequence i in ranking j, p is the power to raise the large homology scores in consensus and n is the number of rankings. p was set to 2 when combining rankings obtained by positive examples.

If a large training set with both positive and negative examples was available, more suitable parameters of the fragment matching method could be learned for the specific datasets. Liao and Nobel (2003) showed that additional accuracy could be obtained by modeling the difference between positive and negative examples. For the method FragMatch, when negative examples were available, these negative examples were also searched in the test set and a ranking of homology scores was obtained for each negative example. These rankings searched by negative examples were first combined by Equation (7) and then the negative consensus ranking was combined with the positive consensus ranking by

$$FC_i = \begin{cases} \dfrac{C_i}{N_i^{\,q}}, if\ C_i < 50\ \ and\ N_i \geq 60 \\ C_i, \qquad\qquad\qquad\qquad else \end{cases} \tag{8}$$

where FC_i is the final consensus homology score, C_i is the consensus score for rankings searched by positive examples, N_i is the consensus score for rankings searched by negative examples and q is the power for N_i. q was empirically set to 0.6.

2.4.2 Dataset for evaluating FragMatch

To evaluate the power of FragMatch for remote homology detection with a single target sequence, two datasets, one with 803 pairs of SCOP family level domain sequences and the other with 480 pairs of SCOP superfamily level domain sequences, were created. The former contains 1606 domain sequences each of which has another sequence within the same family while all others are not within the same superfamily. The latter contains 960 domain sequences each of which has another sequence within the same superfamily but not the same family while all others are not within the same superfamily. These two datasets were derived from 7890 SCOP domain sequences retrieved from the Astral Database (Chandonia *et al.*, 2004) (version 1.73, Nov. 2007) cutting at ≤ 30% sequence identity level. For each domain sequence in the dataset, FragMatch tries to identify its only homologue (either at superfamily level or family level) among the rest of sequences in the dataset.

To test the method for protein family classification, a well-benchmarked database retrieved from Astral Database (version 1.53) by an E-value threshold of 10E-25 was used. This dataset contains 4352 domain sequences,

grouped into 1938 families and 1001 superfamilies. For the reliability of the evaluation, only families containing at least 5 family members and 10 super family members outside of the family were selected. This resulted in 54 families (see Table **6**). For each family, the protein domains within the family were considered positive test examples, and the protein domains outside the family but within the same superfamily were taken as positive training examples. Negative examples were taken from outside of the positive sequences' fold, and were randomly split into training and test sets in the same ratio as the positive examples.

Table 6: List of 54 families for which each one contains at least 5 family members and 10 super family members outside of the family in 4352 domain sequences derived from SCOP version 1.53. [Reproduced from Liao and Noble, (2003) with permission]

	No. of sequences					No. of sequences			
	Positive Set		Negative Set			Positive Set		Negative Set	
SCOP ID	Training	Test	Training	Test	SCOP ID	Training	Test	Training	Test
1.27.1.1	12	6	2890	1444	2.9.1.4	21	10	2928	1393
1.27.1.2	10	8	2408	1926	3.1.8.1	19	8	3002	1263
1.36.1.2	29	7	3477	839	3.1.8.3	17	10	2686	1579
1.36.1.5	10	26	1199	3117	3.2.1.2	37	16	3002	1297
1.4.1.1	26	23	2256	1994	3.2.1.3	44	9	3569	730
1.4.1.2	41	8	3557	693	3.2.1.4	46	7	3732	567
1.4.1.3	40	9	3470	780	3.2.1.5	46	7	3732	567
1.41.1.2	36	6	3692	615	3.2.1.6	48	5	3894	405
1.41.1.5	17	25	1744	2563	3.2.1.7	48	5	3894	405
1.45.1.2	33	6	3650	663	3.3.1.2	22	7	3280	1043
2.1.1.1	90	31	3102	1068	3.3.1.5	13	16	1938	2385
2.1.1.2	99	22	3412	758	3.32.1.1	42	9	3542	759
2.1.1.3	113	8	3895	275	3.32.1.11	46	5	3880	421
2.1.1.4	88	33	3033	1137	3.32.1.13	43	8	3627	674
2.1.1.5	94	27	3240	930	3.32.1.8	40	11	3374	927
2.28.1.1	18	44	1246	3044	3.42.1.1	29	10	3208	1105
2.28.1.3	56	6	3875	415	3.42.1.5	26	13	2876	1437
2.38.4.1	30	5	3682	613	3.42.1.8	34	5	3761	552
2.38.4.3	24	11	2946	1349	7.3.10.1	11	95	423	3653
2.38.4.5	26	9	3191	1104	7.3.5.2	12	9	2330	1746
2.44.1.2	11	140	307	3894	7.3.6.1	33	9	3203	873
2.5.1.1	13	11	2345	1983	7.3.6.2	16	26	1553	2523
2.5.1.3	14	10	2525	1803	7.3.6.4	37	5	3591	485
2.52.1.2	12	5	3060	1275	7.39.1.2	20	7	3204	1121
2.56.1.2	11	8	2509	1824	7.39.1.3	13	14	2083	2242
2.9.1.2	17	14	2370	1951	7.41.5.1	10	9	2241	2016
2.9.1.3	26	5	3625	696	7.41.5.2	10	9	2241	2016

3 Performance measurement

3.1 Cross-validation

Cross-validation is an efficient and reliable approach to estimate the performance and generalizability of a program. In a *K*-fold cross-validation, the whole dataset is randomly split into *K* subsets (typically $K \in [5,10]$). The cross-validation process is repeated *K* times (the folds). In each repeat, one of the *K* subsets is retained as the test set, and the remaining $K-1$ subsets are used as training set. When *K* is equal to the number of examples in the dataset, *K*-fold cross-validation becomes leave-one-out cross-validation. The *K*-fold cross-validation allows efficient use of all examples in the dataset. Moreover, it minimizes the probability of getting an over-optimistic result by chance and thus allows the generalization of the overall results obtained from the cross-validation to real-world predictions on unknown sequence data.

3.2 Precision and recall

The precision is defined as TP/(TP+FP), where TP (true positives) refers to the number of correctly identified positive example, e.g. correctly predicted zinc-binding residues or proteins; FP (false positive) is the number of negative examples that are incorrectly predicted as positive, e.g. residues or proteins predicted to bind zinc, but are not zinc-binding according to the PDB. The recall is defined as TP/(TP+FN), where FN (false negative) is the number of positive examples that are incorrectly predicted as negative. In the study of zinc-binding site prediction, negative examples are far more abundant than positive examples. The negative to positive ratios are 26:1 and 93:1 for CH and CHDE respectively. For such an unbalanced dataset, receiver operating characteristic (ROC) curves (see below) can present an overly optimistic view of the performance of a method (Davis and Goadrich, 2006). The recall-precision curve, in which one plots the precision against the recall, has been proposed as an alternative to the ROC curve in dealing with datasets with great unbalance in the class distribution (Zhang *et al.*, 2004). The area under the recall-precision curve (AURPC) was used in our method for both model selection and performance measurement. AURPC was calculated by a method proposed by Davis and Goadrich (2006).

3.3 Receiver operating characteristic (ROC) curve

The ROC curve plots the sensitivity (i.e. recall) against 1–specificity (the specificity is equivalent to the precision) for a binary classifier system as its discrimination threshold is varied. It can also be represented by plotting the fraction of true positives (TPR = true positive rate) against the fraction of false positives (FPR = false positive rate). The area under the ROC curve (AUC, also termed as ROC score) is a simpler estimation of the performance of a binary classification. AUC is 1 for a perfect classification and a score of 0 means none of the examples are predicted as positive. The expected value of the AUC for a random classification is 0.5. In some cases, positive examples are much less than negative examples in the dataset to be classified. In such cases, the ROC score might not be ideal to distinguish different classifiers. For example, for a dataset containing 3 positive examples and 997 negative examples, the ROC score for a method that ranks the three positive examples at positions 1, 3 and 340 is 0.887, and the ROC score for another method that ranks the three positive examples at positions 39, 45 and 58 is 0.955. The latter is obviously better than the former just according to the ROC score. However, in practice it is hard to say which one is better. For structural biologists trying to solve new crystal structures, they often need to find homologues in the PDB. In that case, it does not matter very much to miss a few homologues as long as they can find some good homologous templates at the top of the list. Under such conditions, one might consider the former method superior to the latter. The ROC50 score, which measures the area under the ROC curve up to the first 50 false positives, is introduced to amend the limitations of the ROC score. The ROC50 score for the above two rankings are 0.990 and 0.190, respectively, indicating that the first ranking is more useful than the second. However, the ROC50 score is not always perfect and it can be misleading sometimes. Take the same illustrative dataset used above (with 3 positive examples and 997 negative examples) for example, the ROC50 score for a method that ranks the three positive examples at positions 1, 2 and 20 is 0.887, while another method that ranks these three positive examples at positions 1, 2 and 60 is 1.000, indicating a perfect ranking. Nevertheless, it is obvious that the first ranking is better than the second. In the study of remote homology detection, both the ROC score and the ROC50 score were used for evaluation.

4 Summary of scientific contributions

4.1 Prediction of zinc-binding sites in proteins (Paper I)

A method for predicting zinc-binding sites in proteins from sequences has been developed. When tested on a non-redundant dataset containing 2727 unique protein chains (see Table **3**), this method predicted zinc-binding Cys, His, Asp and Glu with 75% precision (86% for Cys and His only) at 50% recall according to a 5-fold cross-validation (Figure **20**). That is, when 50% of all zinc-binding Cys and His were picked out (by setting the prediction score to a certain threshold), 85% of the predicted zinc-binding Cys and His actually bind to zinc. When the protein level is considered, this method predicted protein chains containing zinc-binding Cys, His, Asp and Glu with 71% precision (75% precision for Cys and His only) at 50% recall. The chain level prediction accuracy was slightly lower than that of the residue level. This is mainly due to two factors. First, only chains with at least one zinc-binding residue predicted at the correct position were considered as correctly predicted. Chains with zinc-binding residues predicted but none of them at the correct position were not considered as correctly predicted. Second, most chains with two or more zinc-binding sites (thus with 6 or more zinc-binding residues) were better predicted than chains with only one zinc-binding site. When evaluated on the residue level, the former have higher weights. However, when evaluated on the chain level, they are equally weighted.

The zinc-binding predictions made by our method are very successful. First of all, there are 999 zinc-binding CHDE (856 for CH: Cys or His) but 93 630 CHDE (22 865 for CH) in total, thus the random prediction accuracy for CHDE is only 1.1% (3.7% for CH) on the residue level, and 7.7% on the protein level. The zinc-binding site prediction by our method has 71% precision (86% for CH) on the residue level and 70% on the chain level at 50% recall, i.e. substantially higher in accuracy than the random prediction. Secondly, when compared to a recently published paper (Passerini *et al.*, 2006), our method predicted zinc-binding Cys and His at ~10% higher precision at different recall levels (Figure **20**). The results on protein level are not given by Passerini *et al.* (2006), but the out-performance can also be expected, since the prediction accuracies on the residue level and protein level are highly correlated and most chains have only one or two zinc-binding sites (Figure 7).

We also showed that for ~46% of all target chains which have homologues predicted, the zinc-binding prediction accuracy for Cys and His was even higher; 90% precision at 70% recall (Fig. 4b in Paper **I**). This

means that the confidence for zinc-binding site prediction is much higher when homologues are detected. With more and more protein structures deposited in PDB, > 65% of the newly added proteins are estimated to have at least one homologue in the SCOP domain database (Ekman *et al.*, 2005). All such proteins can now be predicted at great accuracy for zinc-binding sites.

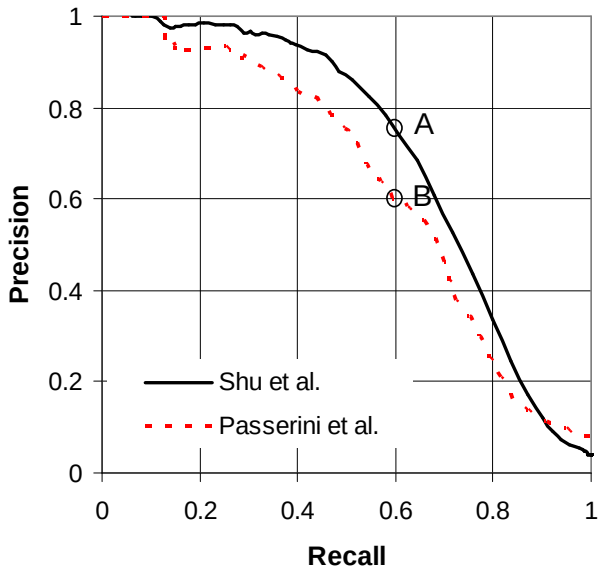


Figure 20: Comparison of the results predicted by our method and that of Passerini *et al.* (2006), for Cys and His on residue level, when tested on the same dataset. At the 60% recall level, our method predicted zinc-binding Cys and His with 76% precision (point A), whereas Passerini *et al.* predicted these two amino acids with 60% precision (point B). [Derived from Fig. 3a in Paper **I** and Fig. 4b in Passerini *et al.*, 2006]

Moreover, our zinc-binding prediction method is accurate enough even to detect potential 'errors' in the PDB. When analyzing the false positives predicted with high confidence, we found that many of them are actually zinc-binding according to the biochemical literature. The absence of zinc in the PDB files for those proteins might be caused by purification and crystallization in zinc-free conditions. Other proteins that were predicted to bind zinc, but had no evidence in the literature as zinc-binding, actually have several highly predicted zinc-binding residues close in 3D space (see Figure **21** for an example). It is highly likely that such a protein will bind zinc *in vivo*.

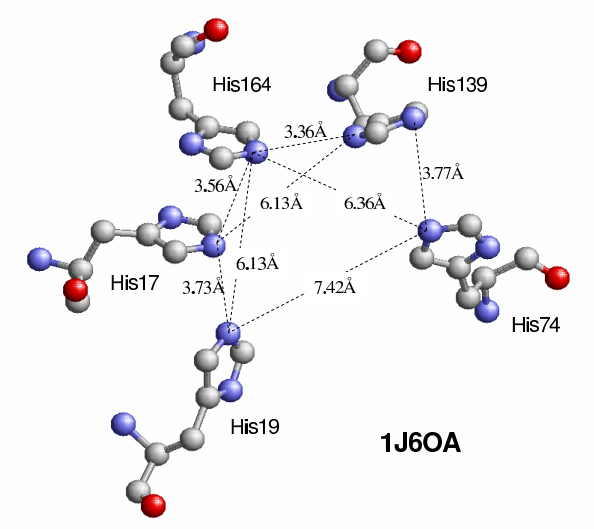


Figure 21: An example of a protein chain highly predicted as zinc-binding but with no bound zinc according to the PDB: chain A of the protein TatD-related deoxyribonuclease (PDB code 1J6O). Four residues His17, His19, His139 and His164 of 1J6OA were predicted at > 0.9 confidence score. His74 was predicted at 0.5 confidence score. These five histidines are closely located in 3D space. The residue sequence numbers in the 1J6O PDB file for His17, His19, His74, His139 and His164 are 4, 6, 61, 126 and 151 respectively. The sequence numbers in the PDB files do not always follow the index of residues in the chain. Therefore, the sequence numbers for these five histidines in the PDB file are different from the index of those residues in the sequence. [From the supplementary data of Paper **I**]

Our method is not only capable of predicting zinc-binding sites in proteins with rather high accuracy, but it can also be used for screening potential zinc-binding sites for protein design, since some apo proteins or proteins with 3 or 4 highly conserved CHDEs that might be close in 3D space can be predicted. In addition, it might also be a useful tool to complement the annotation of zinc-binding sites in PDB files for its ability in identifying occasional un-annotated zinc-binding sites in PDB files.

4.2 Prediction of 1D protein structures (Paper II)

A novel 1D structure prediction method, called Frag1D, was developed using a straightforward profile based fragment matching algorithm. The results show that this method predicted three sets of 1D structural alphabet, i.e. the classical three-state secondary structure, three-state shape strings and eight-state shape strings, successfully.

By exploiting the vast protein sequence and protein structure data available, we have brought the accuracy of the secondary structure prediction closer to the expected theoretical limit (88%, Rost *et al.*, 1994). The method was tested by a leave-one-out cross-validation on a non-redundant set of PDB cutting at ≤ 30% sequence identity (by the PISCES server, Wang and Dunbrack, 2003) containing 5860 protein chains (1.48 million amino acids). For the secondary structure prediction, the Q3 was 82.8%; and for the shape string prediction, the S3 and S8 were 85.0% and 71.5% respectively (see Table 1 and Table 4 in Paper **II**). For 80% of all amino acids predicted with the highest confidence, the Q3, S3 and S8 were as high as 88%, 92% and 79% respectively (Figure **22**). Note that these 80% residues were identified only from their predicted confidences. This means our program not only predicted the 1D protein structure with a good overall accuracy, but also identified quite well which sequences and which parts of the sequences that were better predicted.

We have also benchmarked Frag1D with the latest version of PSIPRED (version 2.2.17, 2008) (Jones, 1999b) for secondary structure prediction. PSIPRED is to date one of the best methods for secondary structure prediction. Frag1D predicted 0.3% better in Q3 when tested on 2241 chains with the same training set (see Appendix 4 for how the training set and the test set were created for this benchmark). The overall outperformance 0.3% Q3 might not be significant. However, for residues in helices and sheets, Frag1D predicted 2.3% better in Q3 compared to PSIPRED (see Table 2 in Paper **II**). In addition, the fact that Frag1D and PSIPRED predicted differently in helices, sheets and random coils, may benefit consensus methods such as JPred (Cole *et al.*, 2008) which combine the results of other original methods to take the merits of Frag1D and PSIPRED to obtain even better results.

For shape string predictions, we benchmarked Frag1D with a recently published method by Kuang *et al.* (2004). When tested on a non-redundant set of PDB chains cutting at ≤ 20% sequence identity, including 1296 chains, Kuang *et al*. predicted the three-state shape strings at 79.5% S3. With the same definition, Frag1D predicted at 81.7% S3, i.e. 2.2% better in accuracy. It has to be noticed that Kuang's definition is slightly different from that is defined in Figure **1**. For these 1296 chains (containing 304 585 amino acids), Kuang's definition and our definition agreed on 98.5% of all residues. Alternatively, with the definition according to Figure **1**, Frag1D predicted the three-state shape strings at 81.1% S3.

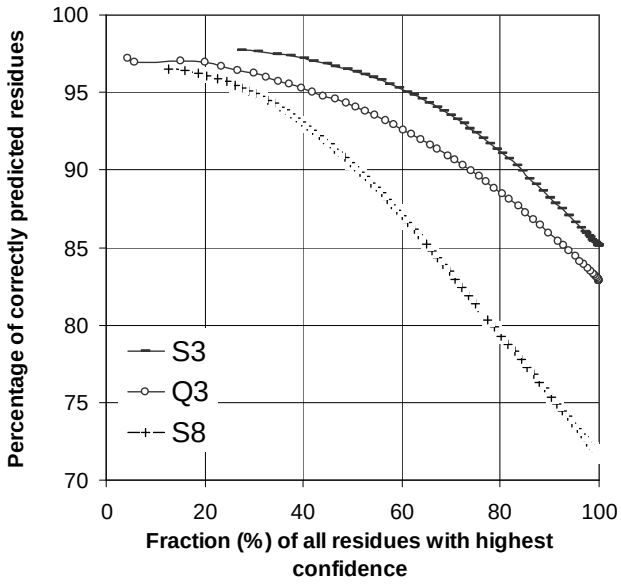


Figure 22: Correctly predicted secondary structure (Q3) and shape strings (S3 and S8) as a function of all residues above a certain confidence. For example, for the ~80% amino acids predicted with highest confidence, Q3, S3 and S8 are roughly 88%, 92% and 79% respectively. [From Fig. 3 in Paper **II**]

It has long been a topic of discussion that the accuracy of secondary structure prediction increases as the size of the database increases, even if the method has not been improved. We have investigated the effect of the size and sequence identity cutting level of the database on the prediction accuracy quantitatively. The results show that the Q3 increases by ~1% with every doubling of the database. Similar trends were observed for S3 and S8 as well (see Fig. 5 in Paper **II**).

4.3 Remote homology detection (Paper III)

A new method, called FragMatch, for detecting remote homologues was developed by using profile-based fragment matching and pattern generalisation based on high scoring candidate fragments on dot plots (see Figure **18** for an example). This method accepts either a single sequence query to search for homologues in a database, or a group of protein sequences with a number of positive examples and negative examples to classify an un-annotated sequence database to be positives or negatives. Therefore, FragMatch is suitable for various purposes in biology, such as finding homologous templates to solve protein crystal structures by

molecular replacement and protein family classification for a newly sequenced genome.

For the remote homology detection with a single query sequence, FragMatch was tested on two datasets; one with 480 superfamily domain pairs and the other with 803 family domain pairs (see section 2.4.2). On the family level, the best average ROC and ROC50 scores for FragMatch were 0.978 and 0.906 respectively (the window size was set to 11). This result was significantly better than that of HHsearch (version 1.5.1) (Soding, 2005), which obtained 0.944 and 0.867 for ROC and ROC50 respectively. Moreover, FragMatch was running ~3 times faster than HHsearch. The popular homology detection program PSI-BLAST obtained 0.845 and 0.694 for ROC and ROC50 scores respectively even supplied with PSI-BLAST checkpoint files built from the NCBI nr database with 6.5 million sequences. On the superfamily level, the best ROC and ROC50 scores were also obtained when setting the window size to 11. Although FragMatch predicted slightly worse ROC50 (0.707 versus 0.721) than HHsearch, the ROC was better (0.920 versus 0.913).

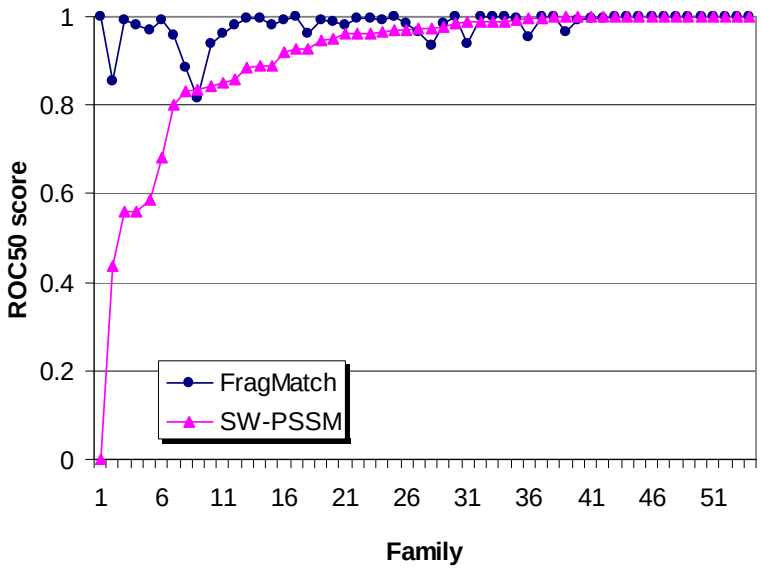


Figure 23: Comparison of ROC50 scores of FragMatch and SW-PSSM for all 54 family classifications. The X-axis is in the ascending order of the ROC50 scores by SW-PSSM. For most family classifications, SW-PSSM predicted homologues as good as FragMatch, with ROC50 scores very close to 1.0, as shown in the right part of the figure. For some family classifications, SW-PSSM did not predict homologues successfully, as shown in the left part of the figure. However, even for these family classifications, FragMatch obtained quite good ROC50 scores. [From Figure 5 in Paper **III**]

For the protein family classification with a training set including both positive examples and negative examples, FragMatch was tested on a well-benchmarked dataset with 4352 domain sequences (see section 2.4.2). The average ROC and ROC50 scores over 54 families for FragMatch were 0.981 and 0.924 respectively. The best result of previously published works tested on this dataset was reached by Rangwala and Karypis (2005). The average ROC and ROC50 scores for their method, SW-PSSM, were 0.981 and 0.904 respectively. Since in this classification task, there are far more negative examples than positive examples, the ROC50 score is a better measurement of the performance (see the discussion in section 3.3). Therefore, the out-performance of FragMatch over SW-PSSM is significant. The per family comparison (Figure **23**) shows clearly that FragMatch performed slightly worse on just 5 families but significantly better on many families.

4.4 Describing and comparing protein structures (Paper IV)

In this work, we reviewed various methods of describing and comparing tens of thousands of protein structures, with the emphasis on the recently developed methods that represent protein backbone structure as one-dimensional geometric strings. We showed that shape strings introduced by Ison *et al.* (2005) are as compact as secondary structures in describing protein backbone structure, and they capture more information for loop regions which comprise ~40% of all amino acids. Moreover, short protein backbone fragments with the same shape string are often highly similar in 3D space (see Fig. 7 in Paper **IV**), although each shape symbol represents a rather large area with a spread of torsion angle φ and ψ in the order of +/-20˚. It means that the 8-state conformation definition of Ison *et al.* (2005) is a good representation of constrains of the 3D path of backbone structures, which is also in accordance with the observation by Kolodny *et al.* (2002) that the conformation space of fragments of native structures is limited. With this observation, it becomes straightforward to construct the 3D backbone structure from a shape string, whereas it still remains a big problem when constructing the 3D structure from secondary structures. Nevertheless, the prediction of shape strings is still not sufficiently accurate and needs further study.

In addition, we showed that shape strings could be applied to improve fast structure database searching. We illustrated in two examples (Figs. 9 and 10 in Paper **IV**) that shape string comparison could reveal the global similarity between protein structures with a hinge bending, whereas rigid body superposition failed in spite of taking a longer computation time. Shape string comparisons can also reveal the difference in loop regions between

different protein structures with the same secondary structure elements, whereas SSE comparison failed in such cases for its lack of information in loop regions. In a large scale homology detection benchmark (Fig. 8 in Paper **IV**), the shape string alignment outperformed KL-string (Friedberg *et al.*, 2007) alignment and sequence based BLAST alignment. It fell behind FATCAT (Ye and Godzik, 2003) and CE (Shindyalov and Bourne, 1998), two Cα based structural alignment methods, but they are three orders of magnitude more time-consuming than shape string alignment.

5 Conclusions

Machine learning methods, data mining techniques and statistics based methods have been widely applied in biology for protein structure prediction, gene finding and other various areas. In this thesis, I have introduced three novel methods for predicting zinc-binding sites in proteins from amino acid sequences, predicting 1D protein structures and detecting remote homologues, respectively.

The zinc-binding site prediction method, PREDZINC, predicts whether a residue (or a protein chain) binds to zinc or not, taking advantage of recent advances in SVM and remote homology detection methods. This method predicted zinc-binding Cys, His, Asp and Glu with 75% precision (86% for Cys and His only) at 50% recall level, when tested on a non-redundant set of PDB containing 2727 unique protein chains. The predictions were so reliable that some occasional mis-annotated proteins regarding zinc-binding were found. This method should be useful for large scale screening for zinc-binding proteins in genomes and for checking poorly annotated proteins whether they are zinc-binding or not. However, the whole zinc-binding group, i.e. exactly which 3 or 4 residues that bind to the same zinc atom, can not be predicted by the method described in this thesis. The prediction of the whole zinc-binding group, and moreover, to distinguish catalytic zinc-binding sites from structural zinc-binding sites, is more challenging and will be of great help for metalloprotein design and 3D structure prediction since the freedom of the 3D structure of zinc-binding proteins will be restricted enormously if zinc-binding sites can be allocated.

The one-dimensional protein structure prediction method, Frag1D, predicted three sets of 1D protein structures with satisfactory results. When tested on a large (5860 chains including 1.48 million amino acids), non-redundant set of PDB chains cutting at $\leq$ 30% sequence identity, Frag1D predicted the protein secondary structure at 82.9% Q3 and shape strings at 85.1% S3 and 71.5% S8. Moreover, better predicted residues and sequences can also be identified by the predicted confidence. For 80% of residues predicted with the highest confidence, the Q3, S3 and S8 were 88%, 92% and 79% respectively.

The remote homology detection method, FragMatch, detected more than twice of the superfamily level homologues and missed less than half of the homologues at the family level compared to the most widely used homology detection method, PSI-BLAST. For protein family classifications, it also outperformed the best method previously published. One can expect more use of this method for structural biologists for structural template searching and genome classification and annotation.

In addition, various methods for describing and comparing protein structures have been reviewed. Some recently developed methods of representing protein structures as one-dimensional geometrical strings, especially shape strings, have been highlighted. Shape strings encode the backbone structures as 1D strings but carry rich structural information in loop regions. They are efficient in detecting the similarity and dissimilarity between protein structures and with them it is possible to construct the 3D structure. However, it should be noted that the current development on applications of shape strings is still very preliminary. More accurate alignments that make best use of the properties of shape strings as well as the analysis of the statistical significance of shape string alignment are required. Moreover, prediction of shape strings instead of the secondary structures of proteins might be an alternative way to start 3D structure prediction. Both the prediction of shape strings and the building of 3D structures from shape strings need further research. Shape strings facilitate fast database searching for similar structures, classification of loop regions and evaluation of model structures. We can expect more widely use of such methods in the near future.

Acknowledgements

First of all, I want to express my deepest gratitude to my supervisor, Prof. **Sven Hovmöller**, for introducing me to the field of protein structure prediction. With your enthusiasm, generosity, inspiration and broad knowledge in science, I enjoyed the fun journey of my PhD education during the past years and more importantly, I learned to do research independently.

I would also like to give special thanks to many other people who have helped a lot making this thesis possible. My sincere and wholehearted thanks go to:
Dr. Tuping Zhou, my close colleague, for being a fabulous friend and for offering numerous help both in scientific work and daily life. Dr. Roger Ison, for many valuable discussions on the shape strings. Prof. Aatto Laaksonen, for helping me with the computing resource. Prof. Arne Elofsson and Prof. Erik Lindahl at Center for Biomembrane Research, for allowing me to use the resources in your Center and for providing me the nice recommendations. Prof. Magnus Sandström, Prof. Margareta Sundberg and Prof. Arnold Maliniak, for establishing a friendly working environment at FOOS. Eva Pettersson, Ann-Britt Rönnell, Paula Jokela, Karin Häggbom Sandberg, Anna Su and Daniel Emanuelsson, for always being ready to help with administrative affairs. Emiliana Risberg, for your help in general chemistry teaching and for your kind encouragement in Swedish learning. Per-Erik Persson and Rolf Eriksson, for solving my computer related problems. My former officemates Daliang Zhang and Peter Oleynikov, and my current officemates Max Peskov and Mysore Srisharfor Santosh, for making our office a nice place to work in. Jonas Almqvist, for helping me out when I newly arrived in Stockholm.

Many thanks should also be given to all colleagues at FOOS for the wonderful atmosphere at work, especially Prof. Xiaodong Zou, Prof. Osamu Terasaki, Prof. Gunnar Svensson, Prof. Zhijian Shen, Lars Eriksson, Shuying Piao, Zhe Zhao, Joan Liu, Juanfang Ruan, Tiezhen Ren, Lei Shi, Junliang Sun, Yanbing Cai, Baolin Lee, Jovice BoonSing, Mikaela Gustafsson, Charlotte Bonneau, Kirsten Christensen, Yasuhiro Sakamoto, Yvonne Fors, Keiichi Miyasaka, Norihiro Muroyama, Wei Wan, Dong Zhang, Huijuan Yue, Daqing Cui, Changming Xu, Mirva Eriksson, Shiliang Huang, Changhong Xiao, Tom Willhammar, Miia Klingstedt, Samrand Shafeie, Guido Todde, Kuo Li.

Many thanks also to my friends and colleagues at Duke University, Prof. Bruce Donald, Prof. Kevin Xiao, Celeste Hodges, John MacMaster, Michael Zeng and Chittaranjan Tripathy, for your kind help during my stay in North Carolina; at Key Laboratory of Microgravity, Institute of Mechanics, Chinese Academy of Sciences, Prof. Mian Long, Prof. Zulai Tao, Drs.

Shujin Sun, Yuxin Gao and Zhiyi Ye, for the nice work we did together on fluid biomechanics.

My special thanks also go to all members in the Stockholm Chinese Choir for making my spare time so colourful and Dr. Jingxia Hao for being a close friend to me and my family.

My dearest parents and sister, thank you for always supporting and loving me; Fang Fang, my wife and best friend: with you my life has a meaning!

This thesis work is partially supported by Calidris, Sollentuna and Carl Tryggers Stiftelse.

6 Appendices

6.1 Appendix 1: HSSP distance

HSSP (homology derived secondary structure of proteins) distance is a measure of sequence similarity which takes both the pairwise sequence identity and alignment length into account (Rost, 1999). It is defined as

$$\text{HSSP distance} = \text{PIDE} - \text{HSSP\_PIDE}, \tag{9}$$

where PIDE is the percentage of pairwise sequence identity and HSSP_PIDE is defined as

$$\text{HSSP\_PIDE} = \begin{cases} 100, & \text{for } L \le 11 \\ 480 \cdot L^{-0.32 \cdot \left\{1+e^{-L/1000}\right\}}, & \text{for } L \le 450 \\ 19.5, & \text{for } L > 450 \end{cases} \tag{10}$$

where L is the length of the alignment. A common HSSP distance threshold is 0, which corresponds to 20% sequence identity for the alignment between two sequences with the length of 300 amino acids (size of a typical protein chain).

6.2 Appendix 2: vector encoding

6.2.1 Encoding of single-site vectors

A window of residues centered at a residue of interest is encoded into a vector of numerical values of size $(2k+1)\cdot p$, where k is the length of extension along the amino acid sequence on both sides of the centered residue and p is the number of numerical features used to describe each residue. In this study, the residue of interest is a selected amino acid residue C, H, D or E as described above. We used 39 numerical features to encode each residue. The first 20 items are the profile of multiple alignments derived from the position specific matrix generated by PSI-BLAST. The information content per position (denoted as score1) and the "relative weight of gapless real matches to pseudocounts" (score2) as in the last two columns of PSI-BLAST output, are encoded by 5 features each. The former is discretized into five bins (0 to 0.2), (0.2 to 0.5), (0.5 to 0.9), (0.9 to 2.0) and (2.0 to $+\infty$) and the latter into the bins (0 to 0.05), (0.05 to 0.8), (0.8 to 1.4), (1.4 to 2.0) and (2.0 to $+\infty$), so that except the first and last bins which hold a few very low or high scores, the numbers of residues within each bin are

essentially equal. Score1 represents the conservation level and score2 the number of aligned residues of that position. A modified one-hot encoding is applied for these bins; for example, score1 of 0.6 is encoded as (0 0 4 1 0). Some of the non-diagonal values of the matrix composed of vectors encoding these bins are set to 1 to represent the correlation between different bins. The diagonal values of the matrix are set to 4 so that the diagonal value of the matrix is equal to the average value of PSI-BLAST profiles for C, H, D and E on their corresponding column. The 31st item is the flag of the position of the residue, which is either 1 (within the sequence) or 0 (outside the sequence, normally at the start or end of the sequence). Five features are used to encode amino acid types which are classified as C, H, D, E and others (for example, C is encoded as 4 1 0 0 0 and H as 1 4 0 0 0, non diagonal values are set to 1 again to represent the fact that some zinc-binding Cys and His are interchangeable). Three features are used to encode hydrophobicity of each residue. Hydrophobicity (Black and Mould, 1991) of residues is classified as hydrophilic (R, D, E, H, N, Q and K), neutral (S, T, G, A) and hydrophobic (C, P, M, V, W, Y, I, L and F). A one-hot encoding is used for these bins, for example, the amino acid lysine (K) is encoded as (4 0 0), whereas proline (P) is (0 0 4). Finally, the feature vector of each residue position is multiplied by a weight defined as

$$W_j = 2 - \frac{|\, pos_j - pos_0 \,|}{k}, \qquad j = -k..k \tag{11}$$

where pos_j is the position of residue j in the sequence and pos_0 is the position of the centered residue of each window, and k is the length of extension on both sides of the centered residue described above. For example, if k = 10 and the position for the centring residue (pos_0) is 25, the weight for the residue at sequence position 30 is calculated as 1.5.

6.2.2 Encoding of pair-based vectors

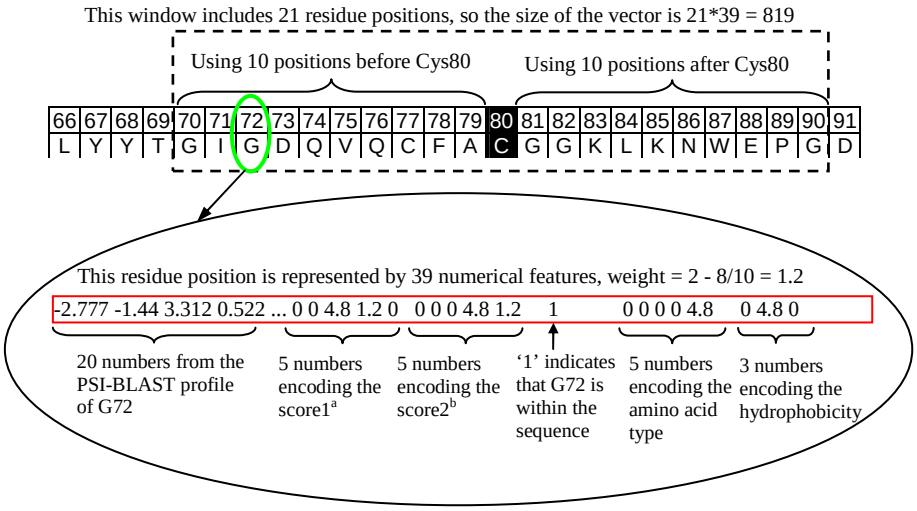
A pair-based vector encodes a window of residues centered by a pair of residues. It represents the correlation between a pair of residues. The encoding of each residue position is similar as for the single-site vectors. Each residue pair is represented by a vector of size $(2 \cdot k+2+2 \cdot w) \cdot p + 5$, where k and p are the same as described in the single-site vector, w is a constant as described below and 5 numerical numbers are used to encode the distance between the two residues of the pair. The number of residues separating the residue pair varies, while SVM requires the input to be a collection of fixed-length vectors (Noble, 2004). To solve this conflict, for residues between the pair, we took always w residues after the first residue in the pair and w residues before the second. Take w = 3 for example. If there are 8 residues (i.e. more than $2 \cdot w$ = 6) between the two zinc-binding residues p1 and p2 in

the pair, p1-x1-x2-x3-x4-x5-x6-x7-x8-p2 (x1-x8 represent the residues within the pair), residues x4 and x5 are not in the feature vector encoding. If there are 2 (less than w = 3) residues within the pair; p1-x1-x2-p2, the numerical features for the putative x3 are set to 0. An example is shown in Figure **24**. Finally, five numerical features are used to encode the distance between the residues of the pair. The distance is discretized into five bins [1 or 2], [3], [4 or 5], [6 to 20] and [> 21], such that the number of zinc-binding pairs within each bin is nearly equal. A one-hot encoding is used for these five bins, while the diagonal values are set to 4 (for example, the distance 3 is encoded as 0 4 0 0 0).

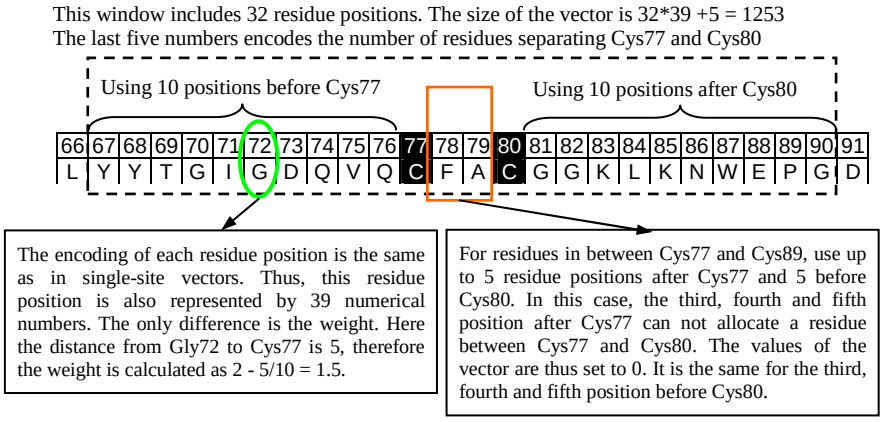
Amino acid sequence of 1C9QA (117 amino acids) from Lys66 to Asp91 (selected C, H, D and E are shaded in black)

66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91
L	Y	Y	T	G	I	G	D	Q	V	Q	C	F	A	C	G	G	K	L	K	N	W	E	P	G	D

(a) For a **single site** vector centred at Cys80:



(b) For a **pair-based** vector centred at Cys77 and Cys80:



[a]Score1: The information content of the PSSM profile generated by PSI-BLAST. [b]Score2: The relative weight of gapless real matches to pseudocounts of the PSSM profile generated by PSI-BLAST.

Figure 24: Examples for how a single-site vector and a pair-based vector are encoded, when k =10, w = 5 and p = 39, where k is the length of extension along the polypeptide chain on both sides of the entered residue, p is the number of numerical features used to describe each residue and w is the number of residues to take after the first residue in the pair and before the second residue in the pair.

6.3 Appendix 3: profiles

A profile is a table that lists the frequencies of the 20 amino acids at each residue position in the sequence from an evolutionary point of view. A log-odds score is defined as the logarithm of the ratio of the likelihood for two amino acids to be aligned to that of seeing these two amino acids matched by chance:

$$S_{ij} = \log(\frac{q_{ij}}{P_i P_j}) \quad (12)$$

where S_{ij} is the log-odds score between amino acid *i* and *j*, q_{ij} is the likelihood for amino acid *i* and *j* to be matched, P_i and P_j are background frequencies for amino acid *i* and *j* respectively and P_iP_j represents the probability of amino acids *i* and *j* being matched by chance. In this study, profiles were obtained by running PSI-BLAST (version 2.2.13) against the NCBI nr database (version April 2006) for three iterations with an E-value threshold of 0.001. The E-value is a statistical parameter that represents the number of hits one would expect to find by chance when searching a database of a particular size. For example, a hit with an E-value of 1 means that when searching a query sequence in a database of the current size, one would expect to see one match with a similar score simply by chance. The lower the E-value, or the closer it is to 0, the higher the significance of the match. Figure **25** shows a typical profile with log-odds values and weighted percentages that are generated by PSI-BLAST. Such profiles contain a summary of evolutionary information, since they show the average amino acid composition at each position along the protein sequence. This amino acid composition is calculated from many (often hundreds) of protein chains that are considered homologues. These putative homologues are picked out by PSI-BLAST from large sequence databases (e.g. the NCBI nr database which contains over 5 million sequences). The profiles generated by PSI-BLAST have been proven of tremendous value for protein structure prediction (Jones and Swindells, 2002).

(a)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	N	-4	-3	8	4	-5	-2	-2	-3	-2	-6	-6	-2	-5	-5	-4	-2	-2	-6	-5	-5
2	R	-1	2	2	-1	-5	-1	-4	5	-3	-5	-5	-3	-4	2	-5	0	-3	-4	-1	-5
3	N	-2	-2	1	-4	-4	-2	-2	-1	-4	-1	-4	-2	-3	-2	-4	-1	6	1	0	0
4	C	-3	-4	-5	-5	4	-1	-3	-5	0	-4	-4	-4	-4	2	-5	0	-4	1	9	-4
5	K	-1	2	-2	-4	-5	4	-1	-5	-1	-3	-4	4	-4	-1	1	-1	2	-5	1	-3
6	L	-4	-3	-6	-6	-4	-5	-6	-6	-6	7	2	-3	-2	2	-3	-5	-2	-5	-3	1
7	Q	-2	2	-4	-3	-4	2	-1	-4	2	0	-2	1	-3	-4	-4	0	-1	-5	-2	5
8	T	-2	-2	2	-3	-1	-2	-3	-3	-4	-4	-4	-3	-4	-5	3	4	5	-5	-4	-4
9	Q	2	-2	1	-2	-4	-1	-3	0	-4	2	-2	1	-3	-1	-2	2	-2	-5	-1	1
10	L	-1	2	-4	-4	-1	-2	-2	-4	0	2	2	3	1	-1	-3	-3	-3	-1	-3	2

[Continues to the next page]

(b)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	N	0	0	81	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	R	5	12	11	4	0	2	0	49	0	0	0	0	0	9	0	6	0	0	2	0
3	N	2	1	8	0	0	2	1	4	0	4	0	2	0	1	0	2	61	2	3	7
4	C	1	0	0	0	9	2	1	0	1	0	0	0	0	5	0	7	0	1	72	0
5	K	3	11	2	0	0	18	3	0	1	1	1	28	0	2	8	3	12	0	5	2
6	L	0	1	0	0	0	0	0	0	0	61	20	2	0	9	1	0	1	0	0	4
7	Q	3	10	0	1	0	9	3	0	4	3	3	9	0	0	0	8	3	0	1	41
8	T	1	2	11	0	1	1	0	1	0	0	1	0	0	0	13	29	38	0	0	0
9	Q	18	2	6	2	0	2	1	6	0	12	4	10	0	2	2	18	2	0	2	9
10	L	4	12	0	0	1	1	3	1	2	13	21	17	3	3	1	1	0	1	0	13

Figure 25: Part of a typical profile for a protein sequence generated by PSI-BLAST with (a) log-odds values and (b) weighted percentages. The first line lists the one letter code of the 20 amino acids found in proteins. The residue number in sequence and amino acid types are shown in the first two columns. The 20 values (which compose a profile at each residue position) in each row are scaled log-odds values in (a) and weighted percentages rounded to the nearest integer in (b) [see (Altschul *et al.*, 1997) for details] .

Profiles of the test protein sequences (proteins with only amino acid sequence information available) are represented by Q_{ij} (the estimated probability for residue *i* to be found on amino acid *j*, *j* represents 20 amino acids). Q_{ij} is calculated from weighted percentages by taking pseudo-counts into account, according to Altschul *et al.* (1997), which is defined as

$$Q_{ij} = \frac{\alpha f_{ij} + \beta P_j \sum f_{ik} e^{\lambda_u S_{kj}}}{\alpha + \beta} \tag{13}$$

where α and β are the relative weights given to observed and pseudocount residue frequencies, f_{ij} is a weighted percentage (see Figure **25b** for an example), that is, the observed frequency for residue *i* on amino acid *j*, P_j is the background frequency for amino acid *j* , S_{kj} is the substitution score from amino acid *k* to *j* as given in BLOSUM62 (Henikoff and Henikoff, 1992), and λ_u is a statistical parameter related to the database for PSI-BLAST.

For the training set, Q_{ij} profiles are enriched by structural profiles derived from blocks of similar shape string fragments. To get the blocks of shape strings, a sliding N-residue (here N is set to 9) fragment of shape strings of a given sequence in the training set is searched against all N-residue fragments of shape strings from all other sequences in the training set. First, up to top 200 N-residue shape string fragments are picked out by the similarity in shape strings. Then, the number of the initially selected fragments is further reduced to up to 100 by the similarity in amino acids and the water accessibility between the target fragment and candidate fragments. Once the blocks of shape strings are obtained, a position specific substitution matrix for each sliding N-residue fragment is derived from amino acid percentages at different columns of a L×N block (L is the

number of matched shape string fragments in the block) based on the same principles as in PSI-BLAST. The structural profile of the entire sequence is calculated as the average of substitution matrices of all sliding fragments for that sequence. Finally, the profile for the residue *i* in column *j* (F_{ij}) of that sequence is calculated as the linear combination of the Qij profile and structural profile according to,

$$F_{ij} = (1-\lambda) * Q_{ij} + \lambda * S_{ij} \quad (14)$$

where Q_{ij} is the same as mentioned before, S_{ij} is the profile for residue *i* on amino acid *j* (*j* is the index for 20 amino acids) derived from blocks of shape strings and λ (ranging from 0 to 1) is the parameter used to linearly combine the Qij matrix and Sij matrix. In practice, we find that a λ of 0.40 will often produce the best result for the remote homology detection and for the 1D structure prediction. Such linear combination method to merging the sequence profile and structural profile has also been used by Teodorescu *et al.* (2004) in protein threading and led to satisfactory results.

6.4 Appendix 4: dataset for benchmarking with PSIPRED

The training set was obtained from Dr. David Jones which has been used to build the weighting files for PSIPRED version 2.61. This training set contains 6598 protein chains with 1 563 587 amino acids. The average sequence length is 237. Note that many chains in this training set are of high sequence identity to each other. For example, the sequence identity of the chain 1JPTL and 1L7IL is as high as 90%. When cutting this training set down to ≤ 30% sequence identity by the PISCES server (Wang and Dunbrack, 2003), only 3644 chains remain.

The test set was constructed in the following ways. First, all PDB chains (as of June 10, 2009) cutting at ≤ 99% sequence identity were obtained with the following criteria: resolution < 2.5Å, R-value < 0.3 and only X-ray structures were used. In total 21 574 protein chains were retrieved. Then, those chains with the same chain IDs as examples in the training set were removed. This resulted in 15 256 protein chains. After that, PSI-BLAST (blastpgp) was run with an E-value threshold of 0.001 and three iterations for searching each of these 15 256 chains in the training set and chains with at least one significant hit in the training set were removed. To be a significant hit, the candidate should meet at least one of the following two criteria: (1) the sequence identity > 30%, the alignment length > 30 and the E-value < 0.1; (2) the sequence identity > 50%, the alignment length > 15 and the E-value < 5. This resulted in 3100 protein chains satisfying the above criteria.

Finally, these 3100 chains were cutting down to ≤ 30% sequence identity by the PISCES server, which resulted in 2421 unique chains for testing.

References

Adzhubei, A.A. and Sternberg, M.J. (1993) Left-handed polyproline II helices commonly occur in globular proteins, *J Mol Biol*, **229**, 472-493.

Al-Karadaghi, S., Cedergren-Zeppezauer, E.S. and Hovmöller, S. (1994) Refined crystal structure of liver alcohol dehydrogenase-NADH complex at 1.8 A resolution, *Acta Crystallogr D Biol Crystallogr*, **50**, 793-807.

Alexander, P.A., He, Y., Chen, Y., Orban, J. and Bryan, P.N. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function, *Proc Natl Acad Sci U S A*, **104**, 11963-11968.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

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

Andersson, K.M. and Hovmöller, S. (2000) The protein content in crystals and packing coefficients in different space groups, *Acta Crystallogr D Biol Crystallogr*, **56**, 789-790.

Andreini, C., Bertini, I. and Rosato, A. (2004) A hint to search for metalloproteins in gene banks, *Bioinformatics*, **20**, 1373-1380.

Anfinsen, C.B. (1973) Principles that govern the folding of protein chains, *Science*, **181**, 223-230.

Auld, D.S. (2001) Zinc coordination sphere in biochemical zinc sites, *Biometals*, **14**, 271-313.

Babor, M., Greenblatt, H.M., Edelman, M. and Sobolev, V. (2005) Flexibility of metal binding sites in proteins on a database scale, *Proteins*, **59**, 221-230.

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

Black, S.D. and Mould, D.R. (1991) Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications, *Anal Biochem*, **193**, 72-82.

Bonneau, R. and Baker, D. (2001) Ab initio protein structure prediction: progress and prospects, *Annu Rev Biophys Biomol Struct*, **30**, 173-189.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E. and Baker, D. (2003) Rosetta predictions in CASP5:

successes, failures, and prospects for complete automation, *Proteins*, **53 Suppl 6**, 457-468.

Brenner, S.E., Chothia, C. and Hubbard, T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc Natl Acad Sci U S A*, **95**, 6073-6078.

Bussiere, D.E., Pratt, S.D., Katz, L., Severin, J.M., Holzman, T. and Park, C.H. (1998) The structure of VanX reveals a novel amino-dipeptidase involved in mediating transposon-based vancomycin resistance, *Mol Cell*, **2**, 75-84.

Bystroff, C., Thorsson, V. and Baker, D. (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins, *J Mol Biol*, **301**, 173-190.

Byvatov, E. and Schneider, G. (2003) Support vector machine applications in bioinformatics, *Appl Bioinformatics*, **2**, 67-77.

Carugo, O. (2006) Rapid Methods for Comparing Protein Structures and Scanning Structure Databases, *Current Bioinformatics*, **1**, 75-83.

Carugo, O. (2007) Recent progress in measuring structural similarity between proteins, *Curr Protein Pept Sci*, **8**, 219-241.

Carugo, O. and Pongor, S. (2002) Recent progress in protein 3D structure comparison, *Curr Protein Pept Sci*, **3**, 441-449.

Chakrabarti, P. and Pal, D. (2001) The interrelationships of side-chain and main-chain conformations in proteins, *Prog Biophys Mol Biol*, **76**, 1-102.

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2004) The ASTRAL Compendium in 2004, *Nucleic Acids Res*, **32**, D189-192.

Cheng, H., Sen, T.Z., Jernigan, R.L. and Kloczkowski, A. (2007) Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: combining GOR V and Fragment Database Mining (FDM), *Bioinformatics*, **23**, 2628-2630.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation, *Biochemistry*, **13**, 222-245.

Cole, C., Barber, J.D. and Barton, G.J. (2008) The Jpred 3 secondary structure prediction server, *Nucleic Acids Res*, **36**, W197-201.

Coleman, J.E. (1992) Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins, *Annu Rev Biochem*, **61**, 897-946.

Cristianini, N. and Shawe-Taylor, J. (2000) *An introduction to support vector machines and other Kernel-based learning methods*. Cambridge University Press.

Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*. ACM Press, Pittsburgh, Pennsylvania.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins, *Atlas of protein sequence and structure*, **5**, 345-351.

Deleage, G. and Roux, B. (1987) An algorithm for protein secondary structure prediction based on class prediction, *Protein Eng*, **1**, 289-294.

Dor, O. and Zhou, Y. (2007) Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training, *Proteins*, **66**, 838-845.

Ebert, J.C. and Altman, R.B. (2008) Robust recognition of zinc binding sites in proteins, *Protein Sci*, **17**, 54-65.

Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**, 755-763.

Ekman, D., Bjorklund, A.K., Frey-Skott, J. and Elofsson, A. (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions, *J Mol Biol*, **348**, 231-243.

Feldheim, D.L. and Eaton, B.E. (2007) Selection of biomolecules capable of mediating the formation of nanocrystals, *Acs Nano*, **1**, 154-159.

Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z. and Godzik, A. (2007) Using an alignment of fragment strings for comparing protein structures, *Bioinformatics*, **23**, e219-224.

Gattiker, A., Gasteiger, E. and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool, *Appl Bioinformatics*, **1**, 107-108.

Gibrat, J.F., Madej, T. and Bryant, S.H. (1996) Surprising similarities in structure comparison, *Curr Opin Struct Biol*, **6**, 377-385.

Gong, H., Fleming, P.J. and Rose, G.D. (2005) Building native protein conformation from highly approximate backbone torsion angles, *Proc Natl Acad Sci U S A*, **102**, 16227-16232.

Gregory, D.S., Martin, A.C., Cheetham, J.C. and Rees, A.R. (1993) The prediction and characterization of metal binding sites in proteins, *Protein Eng*, **6**, 29-35.

Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J. and Orengo, C. (2003) Recognizing the fold of a protein structure, *Bioinformatics*, **19**, 1748-1759.

Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks, *Proc Natl Acad Sci U S A*, **89**, 10915-10919.

Ho, B.K., Thomas, A. and Brasseur, R. (2003) Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix, *Protein Sci*, **12**, 2508-2522.

Holley, L.H. and Karplus, M. (1989) Protein secondary structure prediction with a neural network, *Proc Natl Acad Sci U S A*, **86**, 152-156.

Homaeian, L., Kurgan, L.A., Ruan, J., Cios, K.J. and Chen, K. (2007) Prediction of protein secondary structure content for the twilight zone sequences, *Proteins*, **69**, 486-498.

Hovmöller, S., Zhou, T. and Ohlson, T. (2002) Conformations of amino acids in proteins, *Acta Crystallogr D Biol Crystallogr*, **58**, 768-776.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004) Recent improvements to the PROSITE database, *Nucleic Acids Res*, **32**, D134-137.

Ison, R.E., Hovmöller, S. and Kretsinger, R.H. (2005) Proteins and their shape strings. An exemplary computer representation of protein structure, *IEEE Eng Med Biol Mag*, **24**, 41-49.

Jaakkola, T., Diekhans, M. and Haussler, D. (2000) A discriminative framework for detecting remote protein homologies, *J Comput Biol*, **7**, 95-114.

Jauch, R., Yeo, H.C., Kolatkar, P.R. and Clarke, N.D. (2007) Assessment of CASP7 structure predictions for template free targets, *Proteins*, **69 Suppl 8**, 57-67.

Jones, D.T. (1999a) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences, *J Mol Biol*, **287**, 797-815.

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

Jones, D.T. and Swindells, M.B. (2002) Getting the most from PSI-BLAST, *Trends Biochem Sci*, **27**, 161-164.

Jordan, P., Fromme, P., Witt, H.T., Klukas, O., Saenger, W. and Krauss, N. (2001) Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution, *Nature*, **411**, 909-917.

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

Kabsch, W. and Sander, C. (1983b) How good are predictions of protein secondary structure?, *FEBS Lett*, **155**, 179-182.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, **14**, 846-856.

Kawabata, T. and Nishikawa, K. (2000) Protein structure comparison using the markov transition model of evolution, *Proteins*, **41**, 108-122.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis, *Nature*, **181**, 662-666.

Kleywegt, G.J. and Jones, T.A. (1996) Phi/psi-chology: Ramachandran revisited, *Structure*, **4**, 1395-1400.

Kneller, D.G., Cohen, F.E. and Langridge, R. (1990) Improvements in protein secondary structure prediction by an enhanced neural network, *J Mol Biol*, **214**, 171-182.

Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. (2002) Small libraries of protein fragments model native protein structures accurately, *J Mol Biol*, **323**, 297-307.

Krissinel, E. and Henrick, K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, *Acta Crystallogr D Biol Crystallogr*, **60**, 2256-2268.

Kuang, R., Leslie, C.S. and Yang, A.S. (2004) Protein backbone angle prediction with machine learning approaches, *Bioinformatics*, **20**, 1612-1621.

Laurie, A.T.R. and Jackson, R.M. (2006) Methods for the prediction of protein-ligand binding sites for Structure-Based Drug Design and virtual ligand screening, *Current Protein & Peptide Science*, **7**, 395-406.

Lawrence, J. (1994) *Introduction to Neural Networks*. California Scientific Software Press, Nevada City.

Leslie, C.S., Eskin, E., Cohen, A., Weston, J. and Noble, W.S. (2004) Mismatch string kernels for discriminative protein classification, *Bioinformatics*, **20**, 467-476.

Lewis, P.N. and Scheraga, H.A. (1971) Prediction of structural homology between bovine -lactalbumin and hen egg white lysozyme, *Arch Biochem Biophys*, **144**, 584-588.

Liao, L. and Noble, W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships, *J Comput Biol*, **10**, 857-868.

Lin, C.T., Lin, K.L., Yang, C.H., Chung, I.F., Huang, C.D. and Yang, Y.S. (2005) Protein metal binding residue prediction based on neural networks, *Int J Neural Syst*, **15**, 71-84.

Lovell, S.C., Davis, I.W., Arendall, W.B., 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S. and Richardson, D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation, *Proteins*, **50**, 437-450.

Lu, G.G. (2000) TOP: a new method for protein structure comparisons and similarity searches, *Journal of Applied Crystallography*, **33**, 176-183.

Madej, T., Gibrat, J.F. and Bryant, S.H. (1995) Threading a database of protein cores, *Proteins*, **23**, 356-369.

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection, *Nucleic Acids Res*, **30**, 4321-4328.

Maizel, J.V., Jr. and Lenk, R.P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences, *Proc Natl Acad Sci U S A*, **78**, 7665-7669.

McCall, K.A., Huang, C. and Fierke, C.A. (2000) Function and mechanism of zinc metalloenzymes, *J Nutr*, **130**, 1437S-1446S.

McPherson, A. (1999) *Crystallization of Biological Macromolecules* Cold Spring Harbor Laboratory Press

Menchetti, S., Passerini, A., Frasconi, P., Andreini, C. and Rosato, A. (2006) Improving Prediction of Zinc Binding Sites by Modeling the

Linkage Between Residues Close in Sequence. In Apostolico, A., Guerra, C., Istrail, S., Pevzner, P. and Waterman, M. (eds), *Proceedings of the Tenth Annual International Conference on Research in Computational Molecular Biology*. Springer Berlin / Heidelberg, Venice, Italy, 309-320.

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

Mittelman, D., Sadreyev, R. and Grishin, N. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments, *Bioinformatics*, **19**, 1531-1539.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. and Tramontano, A. (2007) Critical assessment of methods of protein structure prediction-Round VII, *Proteins*, **69 Suppl 8**, 3-9.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. and Tramontano, A. (2009) Critical assessment of methods of protein structure prediction - Round VIII, *Proteins*, **77 Suppl 9**, 1-4.

Moult, J., Fidelis, K., Rost, B., Hubbard, T. and Tramontano, A. (2005) Critical assessment of methods of protein structure prediction (CASP)--round 6, *Proteins*, **61 Suppl 7**, 3-7.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V, *Proteins*, **53 Suppl 6**, 334-339.

Muggleton, S., King, R.D. and Sternberg, M.J. (1992) Protein secondary structure prediction using logic-based machine learning, *Protein Eng*, **5**, 647-657.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, **247**, 536-540.

Nakata, K. (1995) Prediction of zinc finger DNA binding protein, *Comput Appl Biosci*, **11**, 125-131.

Noble, W.S. (2004) Support vector machine applications in computational biology. In Schölkopf, B., Tsuda, K. and Vert, J.-P. (eds), *Kernel methods in computational biology*. MIT Press, Cambridge, Mass., 71-92.

Oliva, B., Bates, P.A., Querol, E., Aviles, F.X. and Sternberg, M.J. (1997) An automated classification of the structure of protein loops, *J Mol Biol*, **266**, 814-830.

Orengo, C.A., Brown, N.P. and Taylor, W.R. (1992) Fast structure alignment for protein databank searching, *Proteins*, **14**, 139-167.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) CATH--a hierarchic classification of protein domain structures, *Structure*, **5**, 1093-1108.

Ouzounis, C., Perez-Irratxeta, C., Sander, C. and Valencia, A. (1998) Are binding residues conserved?, *Pac Symp Biocomput*, 401-412.

Passerini, A., Andreini, C., Menchetti, S., Rosato, A. and Frasconi, P. (2007) Predicting zinc binding at the proteome level, *Bmc Bioinformatics*, **8**, 39.

Passerini, A., Punta, M., Ceroni, A., Rost, B. and Frasconi, P. (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks, *Proteins*, **65**, 305-316.

Pauling, L., Corey, R.B. and Branson, H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain, *Proc Natl Acad Sci U S A*, **37**, 205-211.

Pavlidis, P., Wapinski, I. and Noble, W.S. (2004) Support vector machine classification on the web, *Bioinformatics*, **20**, 586-587.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A*, **85**, 2444-2448.

Presnell, S.R., Cohen, B.I. and Cohen, F.E. (1992) A segment-based approach to protein secondary structure prediction, *Biochemistry*, **31**, 983-993.

Qi, Y., Sadreyev, R.I., Wang, Y., Kim, B.H. and Grishin, N.V. (2007) A comprehensive system for evaluation of remote sequence similarity detection, *Bmc Bioinformatics*, **8**, 314.

Ramachandran, G.N. and Sasisekharan, V. (1968) Conformation of polypeptides and proteins, *Adv Protein Chem*, **23**, 283-438.

Rangwala, H. and Karypis, G. (2005) Profile-based direct kernels for remote homology detection and fold recognition, *Bioinformatics*, **21**, 4239-4247.

Rost, B. (1997) Protein structures sustain evolutionary drift, *Fold Des*, **2**, S19-24.

Rost, B. (1999) Twilight zone of protein sequence alignments, *Protein Eng*, **12**, 85-94.

Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy, *J Mol Biol*, **232**, 584-599.

Rost, B., Sander, C. and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction, *J Mol Biol*, **235**, 13-26.

Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance, *J Mol Biol*, **326**, 317-336.

Schymkowitz, J.W., Rousseau, F., Martins, I.C., Ferkinghoff-Borg, J., Stricher, F. and Serrano, L. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field, *Proc Natl Acad Sci U S A*, **102**, 10147-10152.

Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Eng*, **11**, 739-747.

Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L. and Jones, D.T. (2004) Predicting metal-binding site residues in low-resolution structural models, *J Mol Biol*, **342**, 307-320.

Soding, J. (2005) Protein homology detection by HMM-HMM comparison, *Bioinformatics*, **21**, 951-960.

Teodorescu, O., Galor, T., Pillardy, J. and Elber, R. (2004) Enriching the sequence substitution matrix by structural information, *Proteins*, **54**, 41-48.

The-UniProt-Consortium (2009) The Universal Protein Resource (UniProt) 2009, *Nucleic Acids Res*, **37**, D169-174.

Tupler, R., Perini, G. and Green, M.R. (2001) Expressing the human genome, *Nature*, **409**, 832-833.

Wang, G. and Dunbrack, R.L., Jr. (2003) PISCES: a protein sequence culling server, *Bioinformatics*, **19**, 1589-1591.

Wang, Y., Sadreyev, R.I. and Grishin, N.V. (2009) PROCAIN: protein profile comparison with assisting information, *Nucleic Acids Res*, **37**, 3522-3530.

Vapnik, V.N. (2000) *The Nature of Statistical Learning Theory*. Springer.

Vesterstrom, J. and Taylor, W.R. (2006) Flexible secondary structure based protein structure comparison applied to the detection of circular permutation, *J Comput Biol*, **13**, 43-63.

Windsor, L.J., Bodden, M.K., Birkedal-Hansen, B., Engler, J.A. and Birkedal-Hansen, H. (1994) Mutational analysis of residues in and around the active site of human fibroblast-type collagenase, *J Biol Chem*, **269**, 26201-26207.

Wood, M.J. and Hirst, J.D. (2004) Predicting protein secondary structure by cascade-correlation neural networks, *Bioinformatics*, **20**, 419-420.

Yang, A.S. and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance, *J Mol Biol*, **301**, 665-678.

Ye, Y. and Godzik, A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists, *Bioinformatics*, **19 Suppl 2**, II246-II255.

Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory, *J Mol Biol*, **315**, 1257-1275.

Zhang, J., Bloedorn, E., Rosen, L. and Venese, D. (2004) Learning rules from highly unbalanced data sets. *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*. AOL Inc., Dulles, VA, USA;, 571- 574.

Zhang, Y. (2008) Progress and challenges in protein structure prediction, *Curr Opin Struct Biol*, **18**, 342-348.

Zhang, Y. (2009) Protein structure prediction: when is it useful?, *Curr Opin Struct Biol*, **19**, 145-155.

Zhang, Y., Hubner, I.A., Arakaki, A.K., Shakhnovich, E. and Skolnick, J. (2006) On the origin and highly likely completeness of single-domain protein structures, *Proc Natl Acad Sci U S A*, **103**, 2605-2610.

Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality, *Proteins*, **57**, 702-710.

Zhi, D., Krishna, S.S., Cao, H., Pevzner, P. and Godzik, A. (2006) Representing and comparing protein structures as paths in three-dimensional space, *Bmc Bioinformatics*, **7**, 460.