Bioinformatics Approaches for Disulfide Connectivity Prediction

Chi-Hung Tsai[1], Chen-Hsiung Chan[1], Bo-Juen Chen[1], Cheng-Yan Kao[1], Hsuan-Liang Liu[2]* and Jyh-Ping Hsu[3]*

[1]*Bioinformatics Laboratory, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617;* [2]*Department of Chemical Engineering and Biotechnology and Graduate Institute of Biotechnology, National Taipei University of Technology, Taipei, Taiwan 10608;* [3]*Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan 10617*

Abstract: Protein structure prediction with computational methods has gained much attention in the research fields of protein engineering and protein folding studies. Due to the vastness of conformational space, one of the major tasks is to restrain the flexibility of protein structure and reduce the search space. Many studies have revealed that, with the information of disulfide connectivity available, the search in conformational space can be dramatically reduced and lead to significant improvements in the prediction accuracy. As a result, predicting disulfide connectivity using bioinformatics approaches is of great interest nowadays.

In this mini-review, the prediction of disulfide connectivity in proteins will be discussed in four aspects: (1) how the problem formulated and the computational techniques used in the literatures; (2) the effects of the features adopted to encode the information and the biological meanings implied; (3) the problems encountered and limitations of disulfide connectivity prediction; and (4) the practical usages of predicted disulfide bond information in molecular simulation and the prospects in the future.

Keywords: Protein engineering, protein folding, conformational space, disulfide connectivity, bioinformatics, molecular simulation.

INTRODUCTION

Disulfide bonds, formed by two cysteine residues with oxidized thiol groups, play important roles in proteins. Cysteine residues in proteins exhibit alternative states. Some cysteines form covalently-linked disulfide bonds, while others oscillate between reduced and oxidized forms. Disulfide bond is considered as a significant contributor to the stability of protein conformation, and may also be involved in the protein folding pathway. In some proteins, the oxidation and reduction of cysteine residues (and/or the formation of disulfide bonds) are the essential part of their catalytic functionalities. The oxidation of cysteine residues and formation of disulfide bonds take place outside the cytoplasm, where it provides a reducing environment. The thioredoxin/ thioredoxin reductase and glutathione/glutaredoxin pathways maintain the thiol reducing activity of *E. coli* cytoplasm [1]. In prokaryotes, disulfide bonds are mainly formed in the periplasmic space outside the membrane. In contrast, the formation of disulfide bonds takes place in endoplasmic reticulum (ER) in eukaryotes. As a result, proteins with stable disulfide bonds rarely reside in the cytoplasm.

Although *in vitro* protein folding experiments [2] suggest that disulfide bond formation may be a spontaneous event (but extremely slow); it is an enzyme-catalyzed process *in vivo*. Studies in both bacteria and yeast have revealed that enzymatic systems in the extracytoplasmic compartments are required for efficient and correct formation of disulfide bonds. The formation of correct disulfide connectivity patterns is also attributed to these catalytic enzymes. Since disulfide bond has great impact on the structure/function of proteins, correct connectivity patterns are required for normal activities of proteins with multiple disulfide bonds. The connectivity of peptide sequences can be determined and assigned with several experimental approaches. However, this can only be done after the sequences are synthesized or expressed. In the scenario of protein or peptide engineering, when cysteine residues are introduced in the hope of disulfide bond formation, the capability to evaluate or predict disulfide connectivity from the sequence should prove valuable.

With the knowledge of disulfide connectivity patterns, the conformation complexities of proteins can be largely reduced. The strong constraints imposed by disulfide bonds can be utilized to simplify the structure of proteins or maintain a specific designed conformation [3]. Disulfide bonds can stabilize protein native structures by lowering global free energy. Therefore, engineering of disulfide bonds may also increase the thermal stability of proteins. The development of bioinformatics approaches for disulfide connectivity prediction can be beneficial to protein structure/function prediction. Recently, it has been illustrated that some proteins in the cytoplasm may be activated or inactivated by disulfide formation under oxidative stress or other conditions. This type of 'switch' may be complement to other covalently-linked modifications, e.g. phosphorylation, methylation, car-

*Address correspondence to these authors at the Department of Chemical Engineering, National Taiwan University, Taipei, Taiwan 10617; Tel: 886-2-2363-7448; Fax: 886-2-2362-3040; E-mail: jphsu@ntu.edu.tw; f10894@ntut.edu.tw

1389-2037/07 $50.00+.00

boxylation, etc. The identities of these proteins are largely unknown. Using bioinformatics approaches, it is possible to identify these proteins and the disulfide bond switches in their sequences. The information added to the metabolic networks should provide a systematic view of the processes in living organism.

Bioinformatics approaches toward the prediction of disulfide connectivity pattern are mostly machine learning approaches. Protein descriptors are extracted from the sequences to compose the features or inputs to these learning machines. Various approaches have been employed, including statistical methods, neural networks (NN), and support vector machine (SVM). In this review, we will cover recent advances in the prediction of disulfide connectivity pattern, the approaches, feature selection, and other strategies for improving prediction accuracies. Before we go further into the topic of bioinformatics (computational) approaches, we will provide some biological background regarding the formation of disulfide bonds in prokaryotes and eukaryotes and the biological importance of disulfide bonds.

THE IMPORTANCE OF DISULFIDE BONDS

Disulfide bonds are mostly found in extracellular or secretory proteins. In bacteria, these proteins include toxins responsible for virulence; while in eukaryotes, these include growth factors and extracellular domains of various receptors. The covalent nature of disulfide bonds makes them strong constraints for protein conformations by bringing together sequentially distant regions in protein. Disulfide bonds also stabilize proteins from various aspects. Proteins containing disulfide bonds usually possess higher thermal and chemical stability. Besides, disulfide bonds also play important roles in protein folding pathways by guiding the folding pathway through disulfide bond containing intermediates [4-6]. The functions of various extracellular proteins can be controlled through reduction and formation of one or more disulfide bonds. Many functions of disulfide bonds in proteins are overlapping. For example, the stabilizing power of disulfide bonds may also play critical roles in protein folding pathways.

PROTEIN STABILITY

Disulfide bond has long been recognized as a major contributor to protein thermal stability [7]. A genome-wide survey has shown that hyperthermophile proteins tends to contain more disulfide bonds than proteins from thermophilic and mesophilic organisms [8]. Comparative studies on homologous proteins with different numbers of disulfide bonds have shown that disulfide bond indeed improve the thermal and chemical stabilities of proteins. For example, a comparison between cyclotide kalata B1 and B2 suggests that the cystine-knot structure (with three disulfide bonds) of kalata B1 is responsible for its higher stability comparing to that of kalata B2 (with two disulfide bonds) [9]. Protein engineering study of a thermolysin-like protease (TLP) from *Bacillus stearothermophilus* also confirmed that increasing the number of disulfide can significantly increase its thermal stability [10]. A double cysteine mutant was constructed to introduce a disulfide bridge on the surface of TLP, and the results showed a dramatical improvement of its thermal stability (longer half-life comparing to wild-type). In the case of endostatin, removal of disulfide bonds not only reduce the stability of protein, but also increase the helical contents of the backbone [11]. The loss of disulfide bonds also leads to the loss of protein activities, partly due to the loss of native structures. Introducing a disulfide bond to azurin also increase its thermal stability [12]. The stabilizing effect of introducing a disulfide bond is significantly higher than that of electrostatic interactions in azurin.

However, not all disulfide bonds are stabilizing. In a counter example illustrated by cold-adaptive α-amylase from *Pseudoalteromonas haloplanktis*, removal of four native disulfide bonds leads to a slight increase in stability, but a large decrease in activity [13]. Disulfide bonds in this α-amylase seem to maintain the activity of the enzyme by preventing the formation of electrostatic interactions in the active site. For cold-adaptive proteins, flexibility may be more critical than stability, since flexibility is essential for activity in cold environment [14]. Human neuroglobin and cytoglobin are both with hyperthermal stabilities. Reduction of a potential disulfide bond on distal loop near heme-binding site slightly increases the thermal stability [15]. This disulfide bond may have effect over the heme binding site indirectly, and have no effect on the entire structure. Overall, disulfide bonds contribute to protein stabilities. However, the engineering of disulfide bonds in attempt to increase thermal or chemical stabilities may not always lead to satisfactory results [16-19]. The locations of the disulfide bonds and their effects on the folding pathway of proteins should be considered.

PROTEIN FOLDING

The pioneering work by Anfinsen [2] has illustrated that ribonuclease A (RNase A) denatured by reduction of disulfide bonds can refold to its native structure. Further study has reveled that the folding of RNase A is highly cooperative and no misfolded disulfide bond species are observed [20]. These results imply that the formation of disulfide bonds is spontaneous during protein folding. However, the discoveries of a set of thiol/disulfide exchange enzymes in prokaryotes and eukaryotes have proved that the formation of disulfide bonds is a catalyzed process *in vivo* [21,22]. Genetic studies suggest that not all proteins readily fold with native disulfide bonds and disulfide bond isomerase (DsbC in bacteria and PDI in eukaryotes) is required to refold these proteins with non-native disulfide bonds [1].

In general, disulfide bonds increases the stability of native structure and folding intermediates [23]. Disulfide bonds introduced in different positions have different effect on folding pathways of CD2. For the folding of CD2, intermediates with non-native contacts are the lowest energy paths. Engineered disulfide bonds preventing the formation of non-native contacts slightly decrease in stability; whereas disulfide bonds promoting these contacts increase in stability [23]. In another case, some (not all) of the disulfide bonds present in the native state are observable in the intermediates [4,5]. *Momordica cochinchinensis* trypsin inhibitor-II is of circular form [5], but the folding pathway is different from another circular protein cyclotide kalata B1 [9], where the

two disulfide bond intermediate is not the direct precursor of the native protein [24].

Through these studies, one may conclude that disulfide bonds have profound influences on the folding of proteins. Disulfide bonds may either stabilize the transition states or the intermediates in the folding pathway, or they may stabilize the native conformation. However, these effects are subject to each individual protein. Therefore, introducing or removal of disulfide bonds through protein engineering is also an effective means to study the folding pathways of proteins [20]. The knowledge of disulfide connectivity patterns should serve as primers for experiment design in the studies of protein folding pathways.

PROTEIN FUNCTION

The effects of disulfide bonds on the stabilities and structures of proteins have direct impacts on the functions of proteins. Proteins with permanent disulfide bonds usually reside outside of the cytoplasm. It seems that the natural boundary for disulfide bonds comes from the reducing environment in the cytoplasm and the oxidative condition in periplasmic (prokaryotes) and endoplastic reticulum (eukaryotes). Since disulfide bonds are critical to maintain the structures (and functions) of proteins, loss of function due to disulfide bond reduction are not eligible as evidences for disulfide bond based functional switches. In most cytosolic proteins, gaining disulfide bonds may damage the proteins and leads to the inactivation of proteins.

However, in the past years, two proteins, OxyR [25] and HSP33 [26], are found to gain functionality through the formation of disulfide bonds. Some prokaryotes when exposed to oxidative stress will active the transcription of defense proteins (including superoxide dismutase and peroxidase) against the oxidation condition. It has been found that OxyR will form a disulfide bond under oxidation stress and be activated as a transcription factor for these defense proteins. These peroxidases will remove oxidants in the cytoplasm, relief the cell from oxidative stress. OxyR not only initiates the expression of peroxidases, but also promotes the transcription of disulfide reducing proteins, including glutathione reductase and glutaredoxin. These proteins will remove disulfide bonds from cytosolic proteins. These proteins induced by OxyR not only revert the cell back to reducing condition, but also re-active OxyR by removing the disulfide bonds.

Another mechanism against oxidative stress is triggered by HSP33, which is a chaperone protein. Under oxidative stress condition, four Zn chelating cysteines in HSP33 will from two disulfide bonds through the removal of Zn atom. The oxidized HSP33 will be activated as a chaperone. After the oxidizing stress disappears and the condition in the cell becomes reducing again, HSP33 will become inactive. With oxidation and removable of disulfide bonds as switches to anti-oxidant functions, the cell can response to oxidative stress in a rapid and efficient way.

Besides OxyR and HSP33, a number of proteins within the cytoplasm have been found to form disulfide bonds under oxidative stress [27]. Using 2D gel electrophoresis in conjunction with mass spectrometry, a number of proteins involved in chaperone function, glycolysis, cell growth, and signal transduction have been identified with disulfide formation [27]. This implies that using the formation of disulfide bonds as switches for protein functions may be more general than expected. So far, we have illustrated the importance of disulfide bonds. The capability to predict the formation of correct disulfide bonds in proteins could provide insights to the structures, stabilities, functions, and folding pathways. In the following sections, we will review the recent advances in the prediction of disulfide bond connectivity.

DISULFIDE CONNECTIVITY PREDICTION AS A BIOINFORMATICS ISSUE

With the prior knowledge of the protein sequence, the problem of disulfide bond prediction is to infer the locations and the connectivity of disulfide bridges. Fig. (**1**) illustrates the disulfide connectivity pattern of human urokinase (UniProt ID: UROK_HUMAN; PDB ID: 1U6Q). This problem can be subdivided into four related sub-problems. First of all, does this protein chain contain any disulfide bridges (chain classification)? Since only a part of proteins contain disulfide bridges, the most straightforward impulse is to discriminate them from others and perform further analysis. Usually, binary classifiers are trained with various machine learning techniques to classify protein chains into those containing disulfide bridges and those devoid of disulfide bridges. Second, does this cysteine residue involve in the formation of a disulfide bond (cysteine bonding state prediction)? For proteins with disulfide bridges, it happens that some of the cysteines are not oxidized with another one to form a disulfide bond. Similarly, cysteines are classified into the classes either bonded or non-bonded with all kinds of classifiers. However, different from the first sub-problem, the flanking sequences around cysteine residues are usually used as input instead of the entire protein sequence.

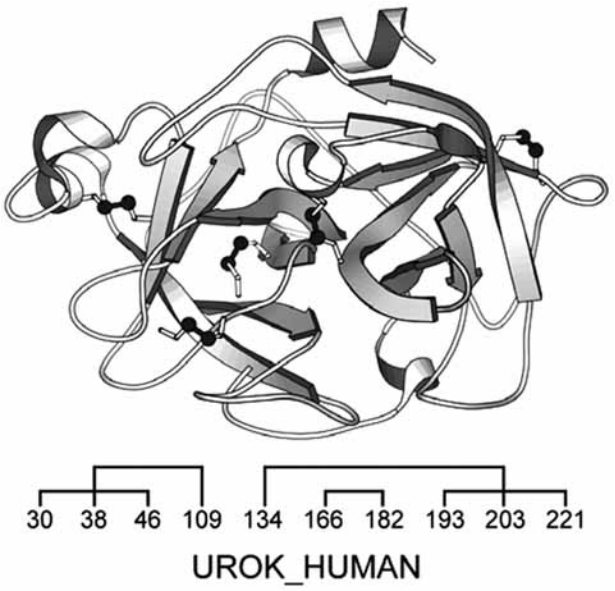

Fig. (1).The 3D structure and the disulfide connectivity pattern of UROK_HUMAN (PDB ID: 1U6Q).

Third, given a pair of cysteines, do they form a disulfide bond (bridge classification) or how likely they are linked together (bonding probability estimation)? Theoretically, it should be a binary classification problem just like previous two sub-problems. However, it sometimes happens that a cysteine residue is predicted to form disulfide bonds with multiple partners, which is unfeasible. To decide the final answer from the candidate combinations, the estimation of bonding probabilities between cysteine pairs is required in such situation. Fourth and also the last, what is the connectivity pattern of bonded cysteines to form the correct disulfide bridges (disulfide connectivity prediction)? This is the real problem that is to be answered. In some studies, this sub-problem is addressed together with the third one with the assumption that the correct oxidation states of cysteines are already known. This problem is difficult and challenging due to its essence and complexity, and the accuracy is so far limited.

In the following sections, approaches addressing different sub-problems will be discussed. Some of them may provide solutions for more than one sub-problem (mostly the predictions of disulfide bonding state and connectivity pattern). The overview of their performances using different features are shown in (Figs. **2** and **3**), and their strength and weakness will also be discussed.

MEASURES OF PERFORMANCE

For the methods addressing cysteine oxidation state prediction, the prediction quality is usually examined using the *n* fold cross-validation. During the cross-validation, the dataset is split into *n* subsets (*n* = 4~20), and each subset is singled out in turn as a test set with the remaining subsets used as a training set. The efficiency of the predictors is scored using the statistical indexes defined as follows:

$$Q_2 = \frac{P}{N} \tag{1}$$

where P is the total number of correctly predicted cysteines, and N is the total number of cysteines. If a protein is treated as a unit, and only those proteins whose cysteines are all correctly predicted are counted. The prediction accuracy per protein is:

$$\mathrm{Q_{2prot}} = \frac{P_p}{N_p} \tag{2}$$

Where P_p is the number of correctly predicted proteins and N_p is the total number of proteins.

Some studies also provide the specificity and the sensitivity for their predictors:

$$\text{specificity} = \frac{TP_x}{TP_x + FP_x}, \ \text{sensitivity} = \frac{TP_x}{TP_x + FN_x} \tag{3}$$

where TP_x is the number of true positives, FP_x is the number of false positives, and FN_x is the number of false negatives of state x (bonded or non-bonded). Another measure of accuracy for oxidation state prediction is Matthew's correlation coefficient (MCC):

$$\mathrm{MCC} = \frac{TP_x TN_x - FP_x FN_x}{\sqrt{(TP_x + FN_x)(TP_x + FP_x)(TN_x + FP_x)(TN_x + FN_x)}} \tag{4}$$

where TN_x is the true negatives of state x.

To evaluate the accuracy of disulfide connectivity prediction (the fourth sub-problem), two indexes Q_p and Q_c are wildly used:

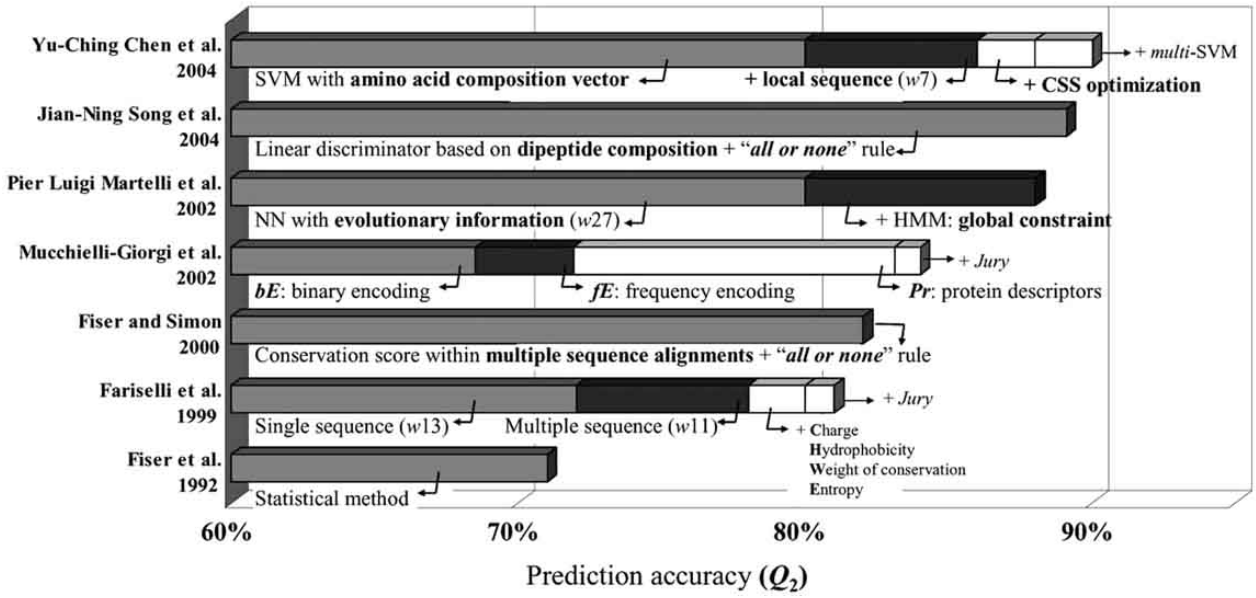

Fig. (2). The overview of the prediction accuracy (Q_2) for the prediction of disulfide bonding state with various combination of features by different methods [29, 31, 37, 38, 40, 42, 78]. To be noticed, it is unfair to compare these results directly since different datasets and experimental protocols were used. However, the tendency of the effects for each feature can be observed. The annotation starts with a "+" sign means that the result is obtained by using this feature in addition to former ones. For the sliding windows used to extract sequence data, the window sizes are labeled in parentheses (e.g. *w*11 as window size 11).

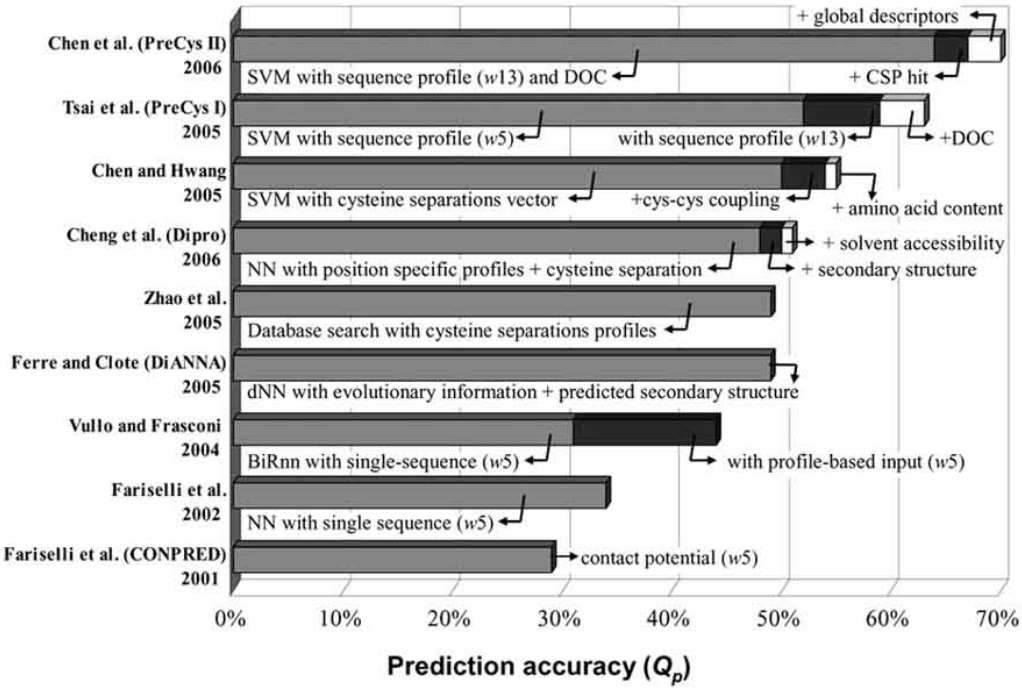

Fig. (3). The overview of the prediction accuracy (Q_p) for the prediction of disulfide connectivity with various combination of features by different methods [28, 46, 47, 53, 55, 56, 58, 64]. To be noticed, it is unfair to compare these results directly since different datasets and experimental protocols were used. However, the tendency of the effects for each feature can be observed. The annotation starts with a "+" sign means the result is obtained by using this feature in addition to former ones. For the sliding windows used to extract sequence data, the window sizes are labeled in parentheses (e.g. *w*5 as window size 5).

$$Q_p = \frac{\delta_p}{N_p},\ Q_c = \frac{\delta_c}{N_c} \tag{5}$$

Where δ_p and δ_c are the total numbers of correctly predicted patterns and cysteine pairs, respectively; and N_p and N_c are the total numbers of proteins and bonding cysteine pairs in the dataset. The index Q_p is pattern-based and estimates predictive performance at the protein level. The index, Q_c, on the other hand, is couple-based and measures the fraction of correctly predicted disulfide bridges. Sometimes the performances of proposed methods are compared to that of a random predictor [28], which has equal probabilities to perform the prediction of every possible connectivity pattern.

PREDICTION OF DISULFIDE BONDING STATE

Incorporating Evolutionary Information in Predicting the Disulfide-Bonding State of Cysteine in Proteins

In the early 90s, statistical [29] and neural network-based [30] methods were proposed to predict the cysteine bonding state with flanking residues. However, the evolutionary information was not considered, and its effect was not investigated until 1999 [31]. In the study of Fariselli *et al.* neural network-based predictors were trained to predict the bonding states of cysteines in proteins. Standard feed-forward neural networks were implemented with a back-propagation algorithm as learning procedure [32]. The network architecture consisted of a perceptron with two output nodes, which discriminates the bonded and free cysteine propensities, respectively, with no hidden layers. Each cysteine residue, flanked by symmetrical segments, was classified as bonded (disulfide bridge-forming) or free, according to the output values of the neurons. Eight different input coding to the networks based on single-sequence input or multiple-sequence profile were considered. Moreover, the charges in the cysteine environment, the hydrophobicity profile, the conservation weight, and the relative entropy of residues were also used to increase the input information.

Training was performed using 2,452 cysteine containing segments extracted from 641 non-homologous proteins of well-resolved three-dimensional structure. Using protein single sequences, the prediction accuracy Q_2 was as high as 72% with cross-validation procedure. The addition of evolutionary information significantly increased the prediction efficiency up to 78% (Fig. **2**). This improvement proves that the evolutionary information is relevant in predicting disulfide-bonding state as well. Furthermore, an improvement of 2% was obtained when the conservation weight and relative entropy were also used, whereas the charge and the hydrophobicity profile seemed to contain no useful information. Finally, a jury of networks improved the prediction accuracy up to 81%, as previously shown by other studies [33,34].

Predicting the Oxidation State of Cysteines by Multiple Sequence Alignment and the "All or None" Rule

Thornton [35] mentioned that the free thiols are unstable outside the cell, i.e. cysteine residues predominantly occur in disulfide bridges [36]. In the intracellular environment, the thiols are kept reduced by glutathione. However, they are

very reactive outside the cell, and may even cause polymerizations. The redox state of the cysteines in a protein is highly correlated to its cellular location. Fiser and Simon [37] assessed the differences between the statistical frequencies of oxidized and reduced cysteines on the surface and in the interior of proteins, and also analyzed the correlation between the cellular location of a protein and the oxidation state of its cysteines. Proteins were grouped into three subgroups according to their cellular locations (intracellular, extracellular, and periplasmic), and checked the occurrence of different types of cysteine in the subgroups. The cellular location showed a high correlation with the oxidation state of the cysteines, but this correlation was not exclusive. The frequencies of different oxidation states of cysteine in secondary structural elements were also analyzed, as well as the types of secondary structural elements linked by a disulfide bridge. The role of disulfide bonds in structure stabilization is supported by the observation that they most often connect two coil regions or a coil with a regular secondary structure.

In the two datasets analyzed, only a few percent of the proteins (2-4%) contain both redox types of cysteines in the same molecule. Nevertheless, the free cysteines in these proteins are often suspected for such as inter-domain links, heavy atom binding sites, active site, etc. so that the cysteine is also oxidized. These bonded cysteines, either liganded cysteines or half cysteines, usually maintain important functions of a protein, and are highly conserved compared with free cysteines. These observations strongly suggest that cysteines tend to occur in the same oxidation state within the same protein. Base on this "all or none" rule, Fiser and Simon [37] proposed a new method for predicting the oxidation state of cysteine residues by conservation scores derived from multiple sequence alignments. A database of 81 protein alignments was used in their analysis. Conservation scores based upon the physico-chemical properties of amino acids were calculated for each position in each alignment. For each position in each protein, this score was then divided by the average conservation of the protein to give a relative conservation score C_r. Therefore, if a larger fraction of the predicted cysteines belongs to one oxidation state (bonded or free), the rest cysteine residues in the same molecule can be assumed to be in the same oxidation state. In cases where the number of reduced and oxidized cysteines predicted in a protein is equal, the averages of relative conservation scores are compared.

The efficiency of the prediction was tested by the jackknife procedure, and the prediction accuracy of the redox state of cysteines was above 82%. The results suggest that the natural borderline lies between the different oxidation states of cysteine rather than between the half cysteines and cysteines. However, there are exceptions (e.g. superoxide reductase, SOD), and for the proteins contain both forms of cysteines, the prediction error can occur. Since the prediction of this method is either all oxidized or all reduced, the practical usage in protein engineering is relatively limited.

Predicting the Disulfide Bonding State of Cysteines Using the Amino Acid Content of the Protein

Previous methods all adopted the information gained from the aligned sets of sequences in predicting the oxidation state of cysteines. Base on the observation that the cysteines of a same protein tends to be found in the same oxidation state, Fiser *et al.* [37] proposed a prediction scheme at the protein level (all or none), and reached up to 82% of success. Such strategy deviates from predicting the oxidation state of cysteines starting from their local sequence environments. It is wondered how informative the amino acids flanking the cysteines are compared with the whole amino acid content of the protein.

In 2002, Mucchielli-Giorgi *et al.* [40] surveyed the efficiency of different descriptors in predicting the cysteine disulfide bonding states. The relative contributions of the segments flanking the cysteines and of the overall amino acid composition of the protein were investigated. For the local environment of cysteines, two different ways of encoding were considered. A window centered on the cysteine of interest was used to define the environment of the cysteine. For each residue in the window, the "binary" description (enote as *bE*) consisted of 0 and 1 according to the type of the particular amino acid encoded. On the other hand, the "frequency" description (denoted as *fE*) which reflects the under- or over-representation of some amino acid types at the different locations in the window was defined as:

$$\log\left(f_k^l / f^l\right) \tag{6}$$

where f_k^l corresponds to the occurrence frequency of the type of amino acid l observed at a given position of the window for the protein k, and f^l is the occurrence frequency of the same type of amino acid in the whole database. As regards to global information, the protein descriptors (denoted as *Pr*) were composed of overall amino acid composition of the protein, along with the normalized protein size and cysteine occurrence.

Simple predictors based on the training of a logistic function were adopted. It is possible to analyze the weights associated with each input, and thus to extract information about the parameters that contribute to the prediction. Given a logistic function and its associated vector of weights, the decision rule was simply comparing the output of the logistic function with a given threshold. If the output value reaches the threshold, the cysteine of interest was predicted as involved in a disulfide bond; otherwise, it was predicted as the reduced state.

The methods was evaluated on a dataset consist of 559 proteins with 5 fold cross-validation. The prediction accuracy obtained with the *bE* encoding was 68%, which is slightly lower than the 72% obtained by Farselli and coworkers [31] with neural networks and single sequences. The use of the *fE* encoding leads to better results, but still on the order of 70% because of the absence of evolutionary information. Significantly higher score (83%) was obtained for the *Pr* encoding, using only the set characterizing the protein composition. These results suggest that, for the disulfide bonding state, the information on the residues flanking the cysteines is less informative than the amino acid content of the whole protein. Additionally, using a combination of logistic functions learned with subsets of proteins homogeneous in terms of their amino acid contents, the prediction accuracy can even reach 84% (Fig. **2**).

Predicting the Disulfide-Bonding State of Cysteines Using Local Context and Global Information with HNN

The importance of the local context in predicting a cysteine-bonding state has been demonstrated using statistical methods [29] and neural networks [30,31]. The accuracy of 81% obtained by Fariselli *et al.* [22] shows that the local context of the central cysteine determines the correct bonding state, and a NN is capable of capturing the relevant interaction within the input window conducive to the bonding or nonbonding state. However, these methods are unable to capture the global information since they predict one cysteine at a time, and this is performed without keeping records of the different predictions associate to a given sequence. That is to say, when a cysteine is predicted in a chain, none of the information about the presence of other cysteines or their predicted bonding states is taken into consideration.

Since each disulfide bond requires two cysteine residues, it is obvious that the total number of cysteines in the disulfide-bonding state of a protein chain should be even. Martelli *et al.* proposed an approach to take advantage of both local and global characteristics of the protein chains, as well as incorporating the trivial "even bonded cysteines" constraint in predicting the cysteine-bonding state [38,39]. This task is addressed by implementing a hybrid system that combines a neural network and a hidden Markov model (hidden neural network) (Fig. **4a**). The local information is extracted by a feed-forward NN, with a sliding window of 27-long centered on the cysteine. For each protein, the evolutionary information based on a multiple sequence alignment is used as NN input. And for each cysteine with local context (W), the estimated probabilities for the bonding and nonbonding states are generated by NN. The global information and the "even bonded cysteines" constraint are then introduced by a hidden Markov model. The NN outputs are used as emission probabilities of a four-state hidden Markov model. The allowed transitions are indicated by arrows, and the bonding and nonbonding cysteine states are respectively indicated by circles of red and blue colors. The path can end only from an even state, which guarantees that only even number of cysteines can be assigned as bonded state.

Training was performed using 4,136 cysteine-containing segments extracted from 969 nonhomologous proteins of well-resolved three-dimensional structure. After a 20-fold cross-validation procedure, the average accuracies of Q_2=~80% and Q_{2prot} =57% were obtained when only the NN-based predictor was adopted. This is similar to the results of Fariselli *et al.* (see Fig. **2**), with NN using a smaller set of proteins [31]. However, when the hybrid system (HNN) was tested on the same protein set, Q_2 increased up to 88% and Q_{2prot} remarkably improved by at least 27%. The improvement is seemingly caused by the introduction of the global constraint by the regular grammar implemented in the HMM, which not only captures the number of cysteines in a chain but also keeps track of the bonding states of all the cysteines in the same chain.

Prediction of the Disulfide-Bonding State of Cysteines in Proteins Based on Dipeptide Composition

It has been revealed that cysteines tends to occur in the same oxidation state within the same protein, and the amino acid composition of proteins with oxidized and reduced cysteines shows clear differences [37]. This is an obvious cooperation phenomenon that cannot be elucidated by only local context near cysteines. Mucchielli-Giorgi *et al.* [40] used

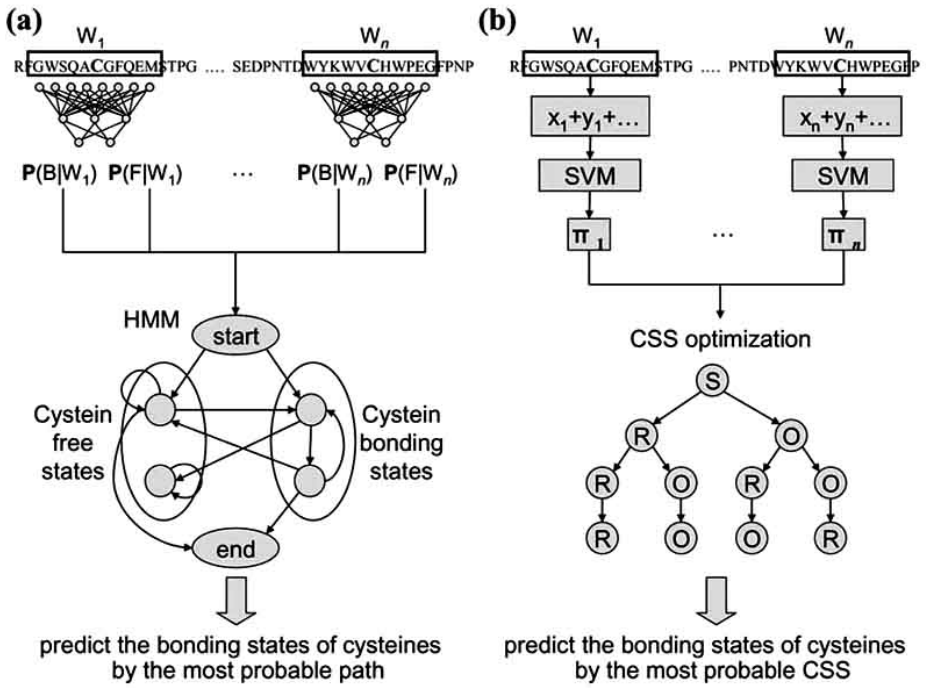

Fig. (4). The flowcharts of the methods for disulfide bonding state prediction by (a) NN+HMM [38]. P(B|W_i) and P(F|W_i) are the probabilities for the bonding (B) and nonbonding (F) states, respectively, at a given local context (W_i); (b) SVM+CSS optimization [42]. The notations S, O, and R denote the initial, bonding, and nonbonding states, respectively. Only CSS of disulfide proteins with 3 cysteines is shown.

logistic functions learned with subsets of proteins with similar amino acid compositions to predict the disulfide-bonding state and reached success rates close to 84%. However, predictions based only on single amino acid composition may lose the information contained in the sequence-order. Incorporating such information may further improve prediction performance. For this reason, dipeptide composition can be considered as another simple representative form of proteins, which incorporates information of global composition and local sequence-order at the same time. A two-class predictor, exploring the dipeptide composition of protein sequence, was proposed by Song *et al.* [41] for predicting the oxidation state of cysteines in proteins by means of a linear discriminator.

For a protein k in the dataset, a characteristic index Q_k is defined as +1 if protein k has at least one disulfide bridge (denoted as OXICYS); otherwise -1 if protein k has no intra-chain disulfide bridge (denoted as REDCYS). The main idea of this method is to predict the characteristic index Q_k of protein k by means of its 400 dipeptide composition $p_{ab}^{(k)}$. A simple linear function of $p_{ab}^{(k)}$ is used to approximate Q_k:

$$Q_k = \sum_{ab} v_{ab} p_{ab}^{(k)} \tag{7}$$

where a and b are amino acid types (they can be the same), and ab stands for one dipeptide. The parameters v_{ab} are constants for all proteins, and the summation runs over all the 400 types of dipeptides. The parameters v_{ab} are chosen that best fit the dataset by minimizing the fitting error. With these parameters, the characteristic index for a given protein can be calculated with dipeptide content by Eq. (7), and the classification (either OXICYS or REDCYS) can be decided by comparing to a critical value Q_c.

The performance of this method was measured with 8,114 cysteine-containing segments extracted from 1,856 non-homologous proteins. The 459 OXICYS proteins have 2,719 cysteines, of which 2,354 take part in intra-chain disulfide bonds. In other words, most cysteines (up to 87%) in OXICYS proteins are involved in the formation of disulfide bridges; while 5,395 cysteines in 1,397 REDCYS proteins are all in reduced state. This distribution is similar to the observation of other researchers [37,40]. Using the jackknife procedure, the prediction accuracy of the cysteine oxidation states is as high as Q_2=89.1%. Comparing to the result obtained by Mucchielli-Giorgi *et al.* [22] using amino acid composition, this comparatively higher prediction accuracy shows that the information incorporating "all or none" rule and sequence-order can be well described by protein's dipeptide composition.

Prediction of Bonding States of Cysteines Using the Support Vector Machine Based on Multiple Feature Vectors and Cysteine State Sequence

In previous sections, various approaches for predicting the bonding states of cysteines were described. It is noticeable that the major breakthroughs come from the utilization of evolutionary information [31, 37-39], the incorporation of local context and global composition [40,41], and the introduction of global constraints [38,39]. In 2004, Chen *et al.* [42] attempted to integrate these ideas based on the support vector machine (SVM), and achieved an extraordinary accuracy up to 90%. Their approach consisted of two stages: 1) the generation of state probability for each cysteine by SVM, and 2) the optimization of cysteine state sequence (Fig. **4b**). In the first stage, SVM was used to predict the bonding states of cysteines. In addition to the local sequences defined by flanking residues of the interested cysteines, the amino acid composition was also adopted as input vector. The decision value obtained from SVM was further normalized by the arctan transfer function, and used as the state probability of each cysteine state in the next stage. In the second stage, the cysteine state sequences (CSS) for a protein with n cysteines was described by the vector s_n= (σ_0, σ_1, σ_2 … σ_n, σ_{n+1}), where $\sigma_i \in \{O, R\}$ [43], the bonding state (O) or the nonbonding state (R), and σ_0 and σ_{n+1} denoting the initial and the final state of the sequence, respectively. Take the proteins with 3 cysteines as an example, there are 4 possible CSSs [i.e., (*OOR*), (*ORO*), (*ROO*), and (*RRR*)]. Each CSS provides a transition path of a particular type of the bonding states of all cysteines in a chain. For each s_n, there is an associated transition probability vector (τ_0, τ_1 …, τ_n), where τ_i is the state-to-state transition probability from σ_i to σ_{i+1}. The state probability of the ith cysteine in state σ_i obtained from SVM is denoted as π_{σ_i}. The overall transition probability of the CSS vector s_n is then given by:

$$p(s_n) = \tau_0 \prod_{i=1}^{n} \pi_{\sigma_i} \tau_i \tag{8}$$

For each protein, the probability of the particular transition path can be easily computed by Eq. (8). Comparison of the probabilities of the transitions paths allows one to predict the most probable CSS. The branch-and-bound algorithm is used to optimize the probability of CSS, while the constraint of even number of cysteines is also applied.

The method was evaluated by the data set comprising 4,136 cysteine-containing segments extracted from 969 non-homologous proteins. Without the optimization of CSS, the SVM based on local sequences or global amino acid compositions yielded similar prediction accuracies (about 80%). However, when multiple feature vectors (combining local sequences and global amino acid compositions) were used, significantly improvement (from 80% to 86%) of the prediction accuracy was obtained. While coupled with CSS optimization, SVM based on multiple feature vectors can yield 90% in overall prediction accuracy (Fig. **2**). This demonstrates that CSS description provides additional information about the bonding states of the cysteines in proteins.

PREDICTION OF DISULFIDE CONNECTIVITY PATTERNS

The Prediction of Disulfide Connectivity as Maximum-Weighted Graph Matching

In 2001, several studies [29-31, 37] tackled the prediction of the disulfide bonding state of cysteines in proteins, and the prediction accuracy of the state-of-the-art at that time

was also acceptable (about 80%). However, this is not the end of the story since the correct connectivity of the bonded cysteines remains unknown.

This problem was firstly addressed by Fariselli and Casadio with stochastic optimization methods [28]. In their work, a classical mapping of disulfide connectivity was proposed, which is extensively adopted in the later researches. Given an even number of cysteines ($2B$, with $B \in N$) forming disulfide bonds, the problem is to determine the correct connectivity pattern among all the possible alternatives. An undirected weighted graph G, which consists of $|V|=2B$ vertices and $|E|=2B(2B-1)/2$ undirected edges, was used to represent the entire system (Fig. **5**). Each vertex represents an oxidized cysteine in the protein, and each edge between two vertices represents the strength of the interactions among the corresponding cysteine pair. With this graph G, the problem of finding the correct connectivity pattern is equivalent to the problem of computing the maximum-weight perfect matching [44], which can be interpreted as the connectivity pattern with highest overall interaction potential.

Subsequently, given a protein sequence containing B disulfide-bond candidates, a rule to assign the bonding potential between two cysteines is required. In the prediction of disulfide bonding state, the local sequence environments of the free and bonded cysteines are quite different as highlighted by neural network-based predictors [30,31]. It is wondered whether there are also characteristic marks within cysteine pairing that can be captured and used to indicate the formation of disulfide bridges. To explore the information in the local cysteine environment, Fariselli and Casadio [28] made a very basic assumption: the nearest neighbors of each cysteine, according to the sequence, all make contacts. Following this assumption, the weight of the edge between vertex i and j is computed as:

$$w_{i,j} = \sum_{k \in S_i} \sum_{l \in S_j} U(k,l) \tag{9}$$

where $U(k, l)$ is the contact potential between amino acids of residue type k and l, and S_i and S_j are the nearest neighbor residues of cysteines i and j, respectively, defined by a symmetric five residue-long window centered on the cysteines. After the edge weights are computed by Eq. (9), the Edmonds–Gabow's algorithm (EG) is then used to solve the connectivity pattern with the maximum weight [45].

Different residue contact potentials were also developed and tested by Fariselli and Casadio [28]. The best performing one was the contact potential derived by Monte-Carlo simulated annealing, and the accuracy was 17 times higher than that of a random predictor in the case of proteins with four disulfide bonds. However, the overall accuracy is only 29% (Fig. **3**), which is far from practical usage. In the subsequent improvement [46], neural network predictions were used instead of assigning edge weights. Satisfactory results were achieved for the simplest cases (B=2 and 3), but the overall accuracy remains low (34%).

Disulfide Connectivity Prediction Using Recursive Neural Networks and Evolutionary Information

Fariselli *et al.* used stochastic optimization [28] and neural network [46] to estimate the interaction potential of cysteine pairs, and predicted the connectivity pattern by finding the maximum weight perfect matching. Vullo and Frasconi [47], on the other hand, adopted a very different strategy to

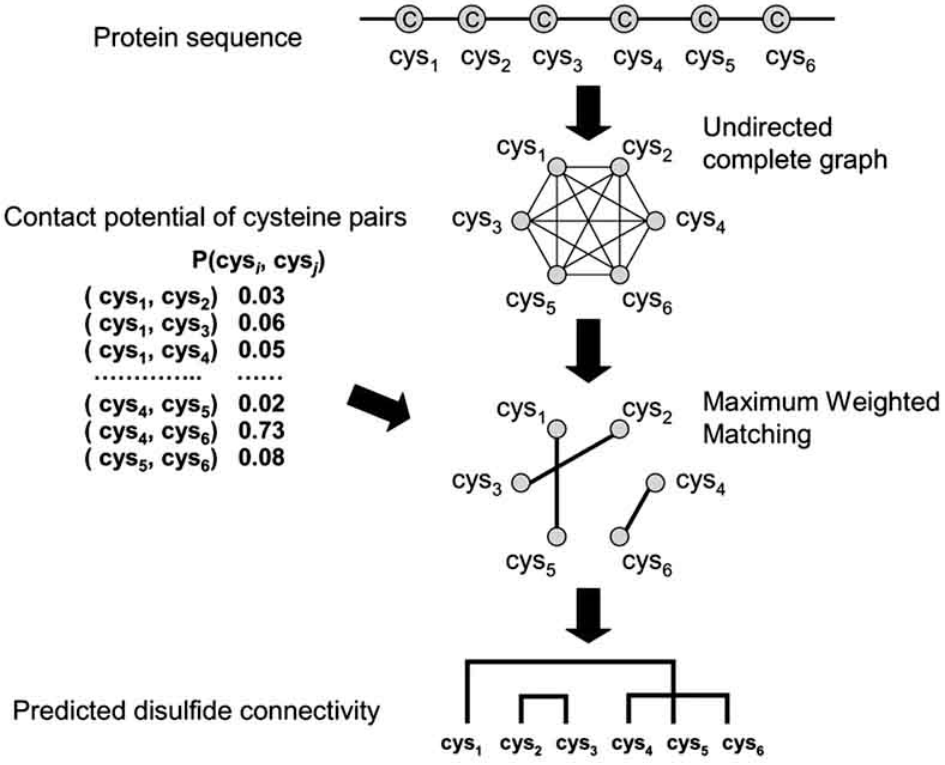

Fig. (5). The transformation of the disulfide connectivity prediction to a maximum weighted matching problem.

tackle this problem. Instead of focusing on a cysteine pair each time, recursive neural networks were trained to score the entire connectivity pattern, and the space of candidate patterns was exhaustively explored to find the one with the highest score. As described in the earlier section, a feasible disulfide connectivity pattern can be represented as a perfect matching in an undirected graph $G = (V, E)$. That is to say, with B disulfide bonds in a chain, $|V|=2B$, $|E|=B$ and degree(v)=1 for any $v \in V$. Let $G^* = (V, E^*)$ denote the graph with target connectivity pattern E^*, the function $s^*(E)$ which maps the undirected graphs of feasible connectivity patterns to real numbers is then defined as:

$$s^*(E) = \frac{|E \cap E^*|}{|E|} \quad (10)$$

It can be easily shown that the function represents the fraction of correct pairs in the candidate solution, and $s^*(E) = 1$ if-and-only-if E is the correct pattern. With this scoring function, the original problem of predicting the correct connectivity pattern can be formulated as finding the pattern with the highest score among the candidates. However, the answer E^* is obviously unknown at prediction time, thus bi-recursive neural network (BiRNN) [48,49] is adopted to learn the scoring function $s(E, V)$ from examples of known disulfide bond patterns. Each vertex in a connectivity graph contains information describing the local environment of the corresponding bonded cysteine. Data encodings based on single sequence and multiple sequence alignments profile were used, respectively, with a window of size 5 amino acids centered on the bonded cysteine. Two additional features, the normalized sequence position of cysteines and the relative sequence length of the protein, were also adopted to enrich the label vectors.

Prediction results on the SWISS-PROT 39 dataset, which was also adopted by Fariselli *et al.*[28,46], were reported. The prediction accuracy Q_p of the BiRNN with single-sequence based encoding was 31%, which is similar to that obtained by Fariselli *et al.* using NN [28,46]. Using multiple alignment profiles instead, the prediction accuracy significantly can be improved to 44% (Fig. **3**), indicating that the evolutionary information is also crucial to the prediction of disulfide connectivity.

Cysteine Separation Profiles on Protein Sequences Infer Disulfide Connectivity

In the study of Chuang *et al*. [50], it has been shown that similar disulfide connectivity patterns infer similar protein structures regardless of sequence identity. Two proteins, tick anticoagulant peptide (PDB ID: 1TAP) [51] and cacicludine (PDB ID: 1BF0) [52], with the same disulfide connectivity patterns were used to demonstrate this discovery. Although they share sequence identity of only 18.2%, their structures are remarkably similar with a C_α root-mean-square deviation (RMSD) of 3.6Å. Moreover, the positions and separations of cysteine residues are also similar for these two proteins (the sequential indexes of cysteines in the two proteins are [5, 15, 33, 39, 55, 59] and [7, 16, 32, 40, 53, 57], respectively).

Enlightened by these observations, Zhao *et al.* [53] made an rational assumption that cysteine separation patterns are highly related to disulfide connectivity patterns. That is to say, proteins with similar cysteine separations should exhibit the same disulfide connectivity pattern. According to this assumption, a quite straightforward approach was proposed to predict disulfide connectivity (Fig. **6**). For a protein with B disulfide bridges, the information of the separations among cysteine residues is encoded as a vector denoted as cysteine separation profiles (CSPs):

$$\text{CSP} = (S_1, S_2, \ldots, S_{2B-1}) = (C_2-C_1, C_3-C_2, \ldots, C_{2B}-C_{2B-1}) \quad (11)$$

Where C_i is the position of ith cysteine residue in the given protein and S_i is the separation between cysteine C_i and C_{i+1}. Comparing the CSPs of two proteins x and y, the divergence, D, is defined as follows:

$$D = \sum_i |S_i^x - S_i^y| \quad (12)$$

where S_i^x and S_i^y are the ith separations for CSPs for proteins x and y, respectively. For each test protein, the CSP was compared with all CSPs of template proteins in the database. The disulfide connectivity pattern of the test protein can be predicted as that of the template protein with the most similar CSP, i.e. with the smallest divergence value D.

Previous works on disulfide connectivity predictions have used graphs to represent disulfide connection patterns. In addition, protein sequences, contact potentials, and evolutionary information have been well used to score various connection patterns [28, 47]. In Zhao's research, on the other hand, only the vectors of separations between oxidized cysteine residues were adopted to predict disulfide connectivity, and the method was also extremely simple. However, the prediction accuracy of 4-fold cross validation on the non-redundant dataset derived from Swiss-Prot 39 was 49%, which is comparable or higher than that of graph-based methods with local context inputs. The results confirmed the assumption that the cysteine separations do contain significant information for the perdition of disulfide connectivity.

With the aid of SVM, Chen *et al.* also utilized the feature of cysteines sequence separations, as well as the coupling between the local sequence environments of cysteine pair and amino acid content [54]. The prediction accuracy of the SVM with cysteine separation inputs was 50%, which is higher than the 45% and 39% obtained by the SVMs with inputs of cysteine coupling and amino acid content, respectively. These results can be explained by the non-local properties of the disulfide bridges that involve cysteine pairs at large sequence separation. Since cysteine coupling contains only local information, and amino acid composition is unable to describe the relation between cysteine pairings. The best accuracy of 55% was obtained by incorporating all these three features in their work (Fig. **3**).

Improving Disulfide Connectivity Prediction with Additional Features

The incorporation of other features is then investigated, since with local sequence context only seems to fail to discriminate the correct disulfide connectivity from others. Three approaches, DiANNA [55], DIpro [56,57] and PreCys

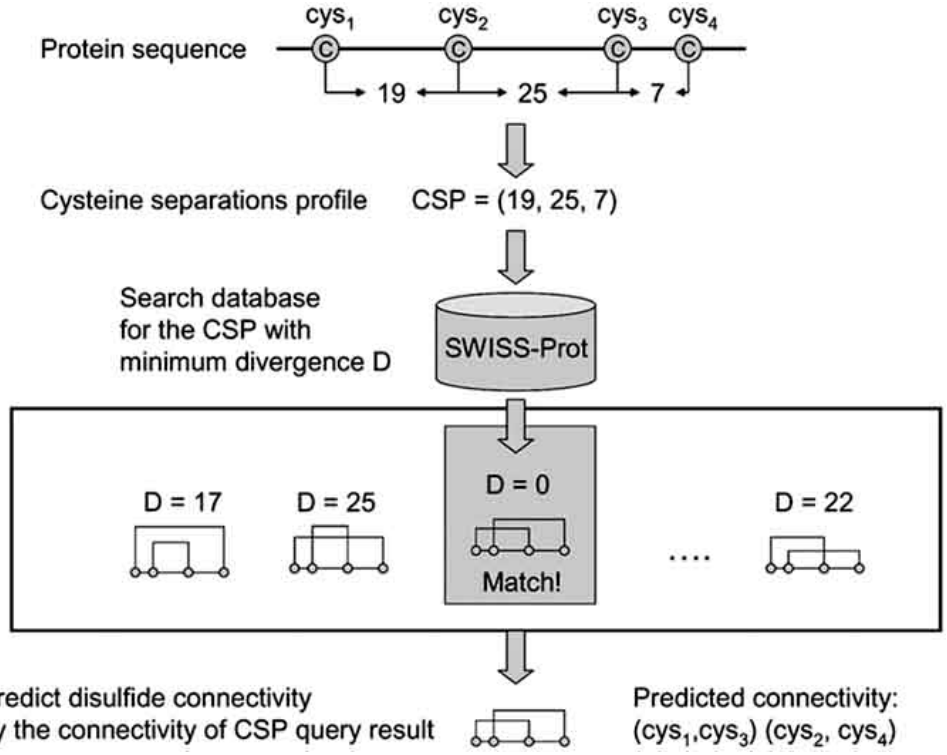

Fig. (6). The flowchart of the disulfide connectivity prediction by CSP search.

[58], based on the weighted graph transformation proposed by Fariselli *et al.* [28] were developed.

Motivated by the observation of a bias in the secondary structure preferences of free cysteines and half-cystines [59], a stand-alone program called DiANNA (*DiA*minoacid *N*eural *N*etwork *A*pplication) was developed by Ferre and Clote [55]. Flanking sequences of cysteines, along with the secondary structure predicted by PSIPRED [60] and the evolutionary information generated by PSI-BLAST [61], were used to train a diresidue neural network for predicting disulfide bond partners in a protein given only amino acid sequence. The maximum weight matching algorithm of Edmonds-Gabow [45] was then applied to assign disulfide bond partners, with the weighted complete graph whose nodes were half-cysteines and weights were values output from the neural network. As calibrated by receiver operating characteristic curves [62] from 4-fold cross-validation, the conditioning on secondary structure showed a marked improvement over the previous works [28,46,49]. Nevertheless, a slight drop in performance occurs when secondary structure is predicted rather than being derived from three-dimensional protein structures.

DIpro [56], on the other hand, can be applied both in situations where the bonded state of each cysteine is known, or in *ab initio* mode where the state is unknown. Firstly, the kernel methods were used to predict whether a given protein chain contains intra-chain disulfide bridges or not, and after that recursive neural networks were applied to predict the bonding probabilities of each pair of cysteines in the chain. These probabilities in turn led to an accurate estimation of the total number of disulfide bridges, and the global disulfide bridge connectivity pattern was inferred by solving the weighted graph matching problem. Local sequence with evolutionary information derived from multiple sequence alignment and the linear sequence separation between the two cysteines were used as input encodes. Secondary structure (SS) and solvent accessibility (SA) information were also added to the input to study their influence upon the prediction. Evaluated on the dataset derived from PDB, the overall accuracy with no additional input information was 48% and reached 51% using true SS and SA as inputs. Using true SS and SA information slightly improved performance, but SS or SA information predicted by the predictors at this stage seemed too noisy to be helpful. This result is consistent with the observation obtained by Ferre and Clote [55].

In the work of Tsai *et al.* [58], the influence of the descriptor derived from sequential distance between oxidized cysteines (denoted as DOC) was full explored. An approach, denoted as PreCys, using SVM based on weighted graph matching was developed to predict the disulfide connectivity pattern in proteins. Different window sizes were used to extract sequence profiles when building SVM models. Using the same window size of 5 as used in previous works [28,46,47,56,57], similar accuracy of 52% was also obtained using PreCys. The accuracy increased with enlarging window size and peaked at 13 (Q_P = 59%), which is better than those obtained in previous works. This may also benefited from the generality of SVM, which avoided overfitting during the training process. When DOC was applied, the SVM models could achieve prediction accuracy of 63%. These results imply that the formation of disulfide linkages between cysteines is determined not only by the local information of cysteines but also by the relationships among them.

The above methods (DiANNA, DIpro and PreCys) described in this section are all available at the respective web servers listed in Table **1**.

Table 1. The Disulfide State/Connectivity Prediction Methods with Available Web-Interfaced Implementations

Authors	Bonding state[a]	Bonding pattern[b]	Method	Web-tool available
Fiser and Simon	SS or S[c]		MSA	http://guitar.rockefeller.edu/~andras/cyspred.html[d]
Mucchielli-Giorgi *et al.* (CysState)	√		Logistic function	http://bioserv.rpbs.jussieu.fr/cgi-bin/CysState
Vullo and Frasconi		√	RNN	http://neural.dsi.unifi.it/cysteines[e]
O'Connor and Yeates	√	√	GDAP	http://www.doe-mbi.ucla.edu/~boconnor/GDAP/
Jianlin Cheng *et al.* (DIpro)	√	√	RNN+WGM	http://www.igb.uci.edu/servers/psss.html
Ferre and Clote (DiANNA)	√	√	dNN+WGM	http://clavius.bc.edu/clotelab/DiANNA
Zhao *et al.* (CSP)		√	CSP search	http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/
Tsai *et al.* (PreCys I)		√	SVM+WGM	http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/[f]
Chen BJ *et al.* (PreCys II)		√	two level SVM	http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/

[a.] the methods addressed on disulfide bonding state prediction.
[b.] the methods addressed on disulfide connectivity prediction.
[c.] the method developed by Fiser and Simon predicted all the cysteines in a protein either oxidized (SS) or reduced (S).
[d,e] these two links seems dead.
[f.] a simple bonding state prediction method is also available on this website.

A Two-Level Model to Predict Disulfide Connectivity

Previous methods for disulfide connectivity prediction either infer the bonding potential of cysteine pairs [28, 46, 55-58] or rank alternative disulfide bonding patterns [47, 53, 54]. Thus, these methods encode data according to cysteine pairs (pair-wise) or disulfide bonding patterns (pattern-wise). However, using either encoding scheme alone can not fully utilize the local and global information of proteins. In the prediction of secondary structures, Rost and Sander [63] designed a two-layer neural network to integrate local and global information, and obtained significant improvements in prediction accuracy. In the prediction of disulfide bonding state, successful strategies with similar two-level models were also proposed by Martelli *et al.* [38,39] and Chen *et al.* [42].

Encouraged by these examples, a two-level framework (Fig. **7**) was adopted to predict disulfide connectivity in the subsequent work of PreCys, which is denoted as PreCys II [64]. With this framework, both the pair-wise and pattern-wise encoding schemes were considered. In the first level, SVM models inferred the bonding potential between two cysteines (Fig. **7**, top). Given a protein with oxidized cysteines, the data were first encoded with respect to each possible cysteine pairs. Two descriptors were considered: 1) local sequence profiles (evolutionary information) around target cysteines from multiple sequence alignments and 2) the sequential distance between oxidized cysteines (DOC). In the second level, pattern-wise encoding was adopted to tackle this problem from a global perspective (Fig. **7**, bottom). For each protein, all possible disulfide bonding patterns were generated for encoding. Three descriptors were considered to encode disulfide bonding patterns: 1) the confidence scores derived from the outputs of level-1 SVM, 2) the results of CSP search [53], and 3) the cysteine ordering and the protein length. With these encodings, SVM was trained to distinguish the correct pattern from the connectivity pattern space. This method was also validated on the datasets used in previous works [47, 56-58], and the overall accuracy of 70% showed that the combination of both local and global information can further improve the prediction performance (Fig. **3**).

Genome-Wide Protein Disulfide Bond Prediction

With the advancements of sequencing techniques, the number of sequenced complete genomes increases explosively. The coverage of Protein Data Bank (PDB) database, on the other hand, is still very low for any particular microbial genome, with many completely sequenced genomes having even no structure available at all. However, the considerable size of complete genomes and the increasing growth speed provide an opportunity for systematic computational predictions. Different from the machine learning methods described previously, the Genomic Disulfide Analysis Program (GDAP) website provides an application which maps the queried sequence to known structures, and uses specific residue distance criteria to identify potential disulfide bonds.

The GDAP is an implementation and extension of a previously proposed disulfide bond prediction technique (Fig. **8**) [65]. Each protein of unknown disulfide bridges was queried against PDB [66] to identify homologous structures, using BLAST [67], PSI-BLAST [61], and Sequence Derived Properties (SDP) [68] in succession. If no homolog could be found among the known structures using all of these algorithms, prediction of disulfide bonding for this protein was not possible. On the other hand, if a match was found between a genomic protein sequence and a PDB structure's sequence, they were aligned together using a local alignment algorithm [69]. This alignment was used as the basis of the sequence-to-structure mapping. Cysteines were then identified in the query sequence, and the coordinates of the corresponding residues in the aligned structure were also ex-

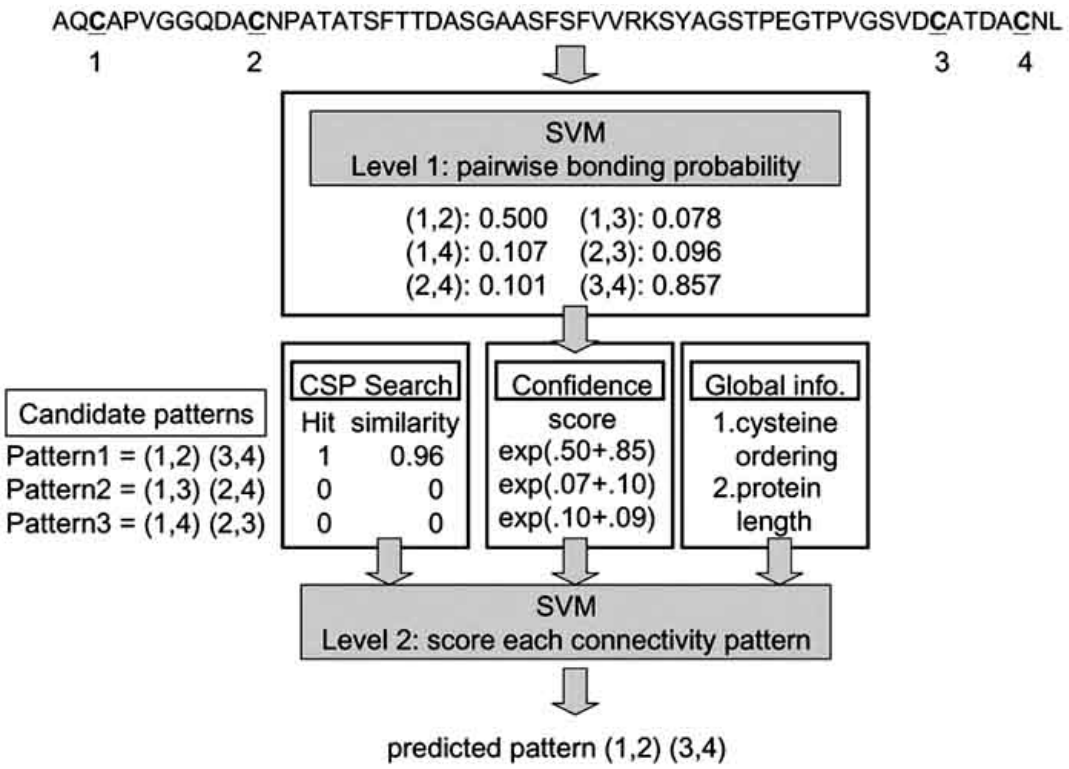

Fig. (7). The flowchart of PreCys II, a disulfide connectivity predictor with two level SVM. At level 1, the bonding probability corresponding to each cysteine pair is estimated by first SVM. At level 2, each candidate connectivity pattern is scored by the second SVM with the confidence score, CSP search result, and the global information of the protein as inputs.

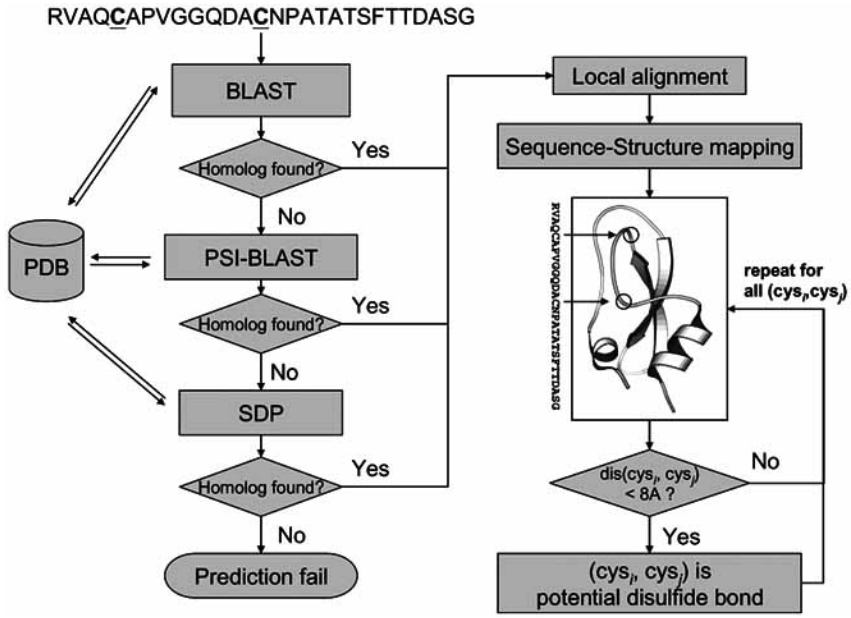

Fig. (8). The flowchart of the Genomic Disulfide Analysis Program (GDAP).

tracted. Two cysteines were considered to be potentially disulfide bonded if the distance between their C_α positions was within 8Å [65] when mapped onto a homologous structure. Every combination of cysteine pairings was examined with this criterion.

In addition to predicting specific disulfide bridges, the GDAP website also actively annotates the query genomic sequences. This information is provided to help in the analysis of disulfide bond-containing proteins. Moreover, proteins that contain known metal binding motifs are flagged. The

identification of proteins with suspected and known metal binding motifs is important because clusters of cysteine residues in these proteins might actually serve catalytic or metal binding roles, rather than forming a structural disulfide bond. Together, the annotations allow the user to restrict the disulfide bond predictions to proteins that meet any combination of these three criteria.

EFFECTS OF THE FEATURES ADOPTED TO ENCODE THE INFORMATION AND THE BIOLOGICAL MEANINGS IMPLIED

Although different methods have been proposed to predict the location of disulfide bridges, their basic idea is to extract information from the databases. The features selected to encode the information are crucial to the extraction of knowledge. Some features are highly correlated to the problem whereas some features are nearly irrelevant, and their effects are usually expressed on the performance of prediction. In the following sections, some frequently used features are described and their effects under different conditions are also discussed.

Describing the Environment of Neighborhood with Local Sequence Context

Since the protein sequence is known before the prediction, the sequences of local fragments are commonly used to encode the local environment for the prediction. This is usually done with input windows of variable lengths (ranged from 5 to 27 residues) centered at target cysteines (Fig. **9a**) [30,31]. Only the flanking residues are taken into consideration while the cysteine is removed from the center, because cysteine always present in the central position of the segment and does not carry any information. An all-zero-but-one binary vector of 21 elements is used to encode the residue type of each residue, with all elements set to 0 but one, set to 1, whose position in the vector identifies the particular residue type [31,47]. The twenty elements correspond to 20 amino acids, while the last one provides a signal when the sequence window encounters either the C or N terminus of protein. In the study of Chen *et al.* [42], sequences with different window sizes (from 5 to 21) were tested for the prediction of oxidation states. The predictive performances steadily increases in accordance with the window size, and reach the maximum (Q_2=81%) when window size is 15. Similar prediction accuracy was also observed in studies using other approaches based on local sequence feature vectors [31,39]. The prediction accuracy remained relatively the same as the window size increases, since only the residues nearby is relevant to the oxidation state of cysteine and those in a distance contain few information.

The Conserved Patterns Hidden in the Evolutionary Information

In addition to the binary encoding of single sequence, the profile-based encoding which incorporates evolutionary information is an alternative way to describe the local environment [47,55]. In the work of Rost and Sander [33], the incorporation of evolutionary information has been shown to substantially increase the accuracy of neural networks for protein secondary structure prediction. For the prediction of cysteine oxidation state (Fig. **2**) and disulfide connectivity (Fig. **3**), improvements have also been obtained using evolutionary information [33, 70-72]. Similarly, each residue is encoded by a 21-element vector as in the former case. However, each of the first 20 elements represents the relative frequencies of the 20 amino acids in the sequence profile which extracted from multiple sequence alignment [31] or from the position-specific substitution matrix (PSSM, Fig. **9b**) [54,55] generated by PSI-BLAST [61,73] (Fig. **9c**). When profile-based encoding is used, the central cysteine is also taken into account since it contains evolution information as well. Tsai *et al.* [58] used sequence profiles of different window sizes to build SVM models for disulfide connectivity prediction. Using the same window size of 5 as used by Vullo and Frasconi (2004), similar Q_P of 52% was also obtained. Furthermore, the overall Q_P increases with enlarging window size and peaks at 13.

Some general characteristics of the local contexts conducive to disulfide bond formation were captured by the weight values of perceptrons without hidden layers [31]. When a segment of 11 residues centered in the cysteine to be predicted was considered, the following features were observed: 1) The presence of cysteine residues in the environment of the central cysteine strongly favors the disulfide bond formation. However, if the cysteine appears at 3 residues away from the central cysteine (e.g. the central cysteine at position *i*, and the other cysteine at position *i*+3 or *i*-3), it is significantly conducive to non-bonding state. This is in good agreement with the observation that in proteins, metal-binding cysteines are typically found within the pattern CXXC (Fig. **10**). That is, two cysteines separated by any other two residues. Moreover, these cysteines, whether involved in a disulfide bridge or a metal-binding site, are both highly conserved during the evolutionary process. 2) Hydrophilic and/or charged residues in the environment are highly conducive toward disulfide-bond formation compared with that of hydrophobic residues, which are poorly conducive.

The Correlation Between the Oxidation Tendency and the Amino-Acid Composition of a Protein

For a protein, the amino acid content is usually represented by a composition vector $A = (a_1, a_2, \ldots, a_{20})$, where $a_k = n_k/n_0$, and n_k is the number of occurrences of the amino acid of type k, and n_0 is the total number of amino acids of the query sequence. It has been shown that amino acid content is a useful global sequence descriptor in several fields, e.g. fold recognition [74] and protein sub-cellular localization [75]. Fiser and Simon [37] found that the cysteines in a protein tend to be in the same oxidation state, i.e. either oxidized or reduced at the same time. In the analysis on different datasets, only a few percent (2-4%) of proteins contain both redox types of cysteines in the same molecule, and the non-disulfide cysteines in these proteins are often suspected to be involved in the formation of inter-domain links, heavy atom binding sites, or active site. In the study of Passerini and Frasconi [76], machine learning techniques were used to discriminate between ligand-bound and disulfide-bound cysteines.

Furthermore, amino acid composition of proteins with oxidized and reduced cysteines shows clear differences.

MYSFPNSFRFGWSQAGFQCEMSTPGSEDPNTWYKWVHDPENNGPGYWG

(a) Single sequence (binary encoding)

F Q C E M

FQCEM: 00001000000000000000, 00000000000010000000, Skip, 00010000000000000000, 00000000010000000000

(b) Position Specific Scoring Matrix

	...	A	G	F	Q	C	E	M	S	T	P	G	S	...
A		0.08	0.02	0.09	0.08	0.03	0.08	0.03	0.09	0.01	0.05	0.05	0.00	
C		0.00	0.12	0.03	0.10	0.01	0.06	0.06	0.07	0.11	0.04	0.02	0.07	
D		0.06	0.07	0.03	0.07	0.07	0.04	0.06	0.04	0.04	0.04	0.08	0.10	
E		0.00	0.01	0.01	0.00	0.00	0.01	0.06	0.06	0.08	0.04	0.05	0.04	
F		0.09	0.07	0.08	0.01	0.08	0.05	0.01	0.02	0.05	0.07	0.08	0.03	
G		0.01	0.03	0.02	0.03	0.01	0.01	0.10	0.04	0.01	0.03	0.04	0.08	
H		0.05	0.07	0.03	0.01	0.03	0.03	0.03	0.07	0.08	0.09	0.04	0.00	
I		0.07	0.03	0.05	0.01	0.06	0.06	0.07	0.03	0.02	0.07	0.01	0.00	
K		0.06	0.10	0.01	0.05	0.09	0.09	0.02	0.07	0.06	0.02	0.01	0.07	
L		0.00	0.01	0.04	0.11	0.10	0.04	0.09	0.03	0.07	0.05	0.02	0.09	
M		0.06	0.09	0.09	0.01	0.08	0.06	0.05	0.01	0.01	0.09	0.06	0.03	
N		0.06	0.01	0.02	0.10	0.03	0.06	0.09	0.02	0.11	0.02	0.08	0.08	
P		0.09	0.09	0.07	0.01	0.10	0.10	0.03	0.07	0.02	0.04	0.03	0.09	
Q		0.07	0.04	0.07	0.11	0.03	0.03	0.04	0.09	0.08	0.03	0.04	0.07	
R		0.09	0.05	0.09	0.06	0.09	0.10	0.07	0.03	0.06	0.06	0.08	0.06	
S		0.01	0.01	0.00	0.06	0.04	0.05	0.05	0.07	0.03	0.08	0.04	0.02	
T		0.06	0.05	0.04	0.05	0.00	0.08	0.04	0.02	0.07	0.02	0.06	0.02	
V		0.06	0.10	0.05	0.02	0.09	0.00	0.01	0.08	0.01	0.04	0.07	0.03	
W		0.02	0.04	0.09	0.03	0.04	0.03	0.01	0.05	0.00	0.06	0.07	0.10	
Y		0.04	0.02	0.11	0.08	0.03	0.02	0.08	0.05	0.07	0.05	0.08	0.02	

(c) Sequence profile

F Q C E M

FQCEM: {.09, .03, ... .09, .11} {.08, .10, ... .03, .08} {.03, .01, ... .04, .03} {.08, .06,... .03, .02} {.03, .06, ... .01, .08}

Fig. (9). The illustration of (a) single sequence encoded by binary encoding; (b) the position specific scoring matrix derived from multiple sequence alignments; and (c) the sequence profile encoding for a local sequence segment of window size 5 centered at cysteine residue.

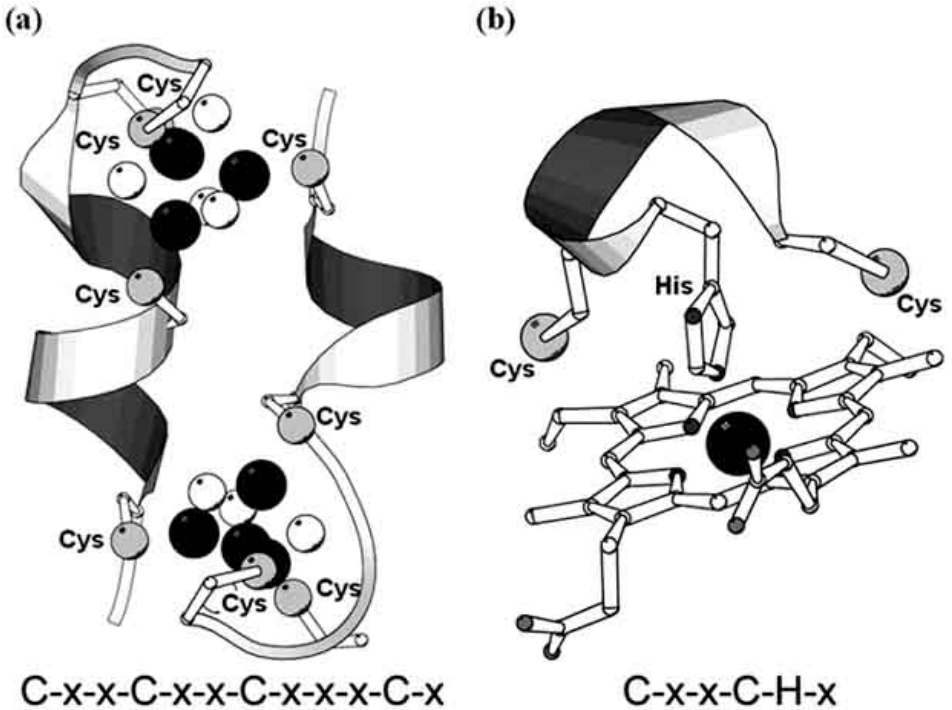

Fig. (10). The ligand binding motifs of (a) 4Fe4S binding site (C-2x-C-2x-C-3x-C-x) and (b) heme groups binding (C-2x-C-H-x).

From the results of several studies [37,40,77,78], the occurrence of serine, threonine, and glutamine is highly favored for oxidized cysteines (Table **2**), while in the case of reduced cysteines the occurrence of glutamate, histidine, lysine, and arginine are higher (Table **3**). The former group includes hydrogen bond forming residues, while positively charged residues are prominent in the latter group. These observations are in good agreement with earlier results obtained by Fiser *et al.* [29] where the sequence environment of cysteine was analyzed.

Table 2. The Amino Acid Types Contribute Positively to Disulfide Bonding State [37,40,77,78]

	Ser	Thr	Gln	Asp	Asn	Tyr
Fiser and Simon (2000)	√	√	√			
Mucchielli-Giorgi *et al.*[a]	√	√		√	√	√
Abkevich and Shakhnovich	√	√	√		√	
JN Song *et al.* (2004)[b]			√		√	√

[a]. The amino acid types of Pro and Phe are also reported by Mucchielli-Giorgi *et al.*
[b]. The dipeptide pattern of XA is also reported by Song *et al.* where A for alanine and X for any of 20 amino acid except A, Q, W, and Y.

Table 3. The Amino Acid Types Contribute Negatively to Disulfide Bonding State [37, 40, 77, 78]

	Glu	His	Lys	Arg	Met	Val
Fiser and Simon (2000)	√	√	√	√		
Mucchielli-Giorgi *et al.*		√			√	√
Abkevich and Shakhnovich[a]	√		√	√		
JN Song *et al.* (2004)[b]					√	

[a]. The amino acid type of Asp is also reported by in the study of Abkevich and Shakhnovich.
[b]. The dipeptide patterns of XC, XG, XI and XT are also reported by Song *et al.* where C for cysteine, G for glycine, I for isoleucine, T for threonine and X for any of 20 amino acid with some exceptions.

The Conservation of Cysteine Separation Patterns for Disulfide Bridges

Disulfides bonds help stabilize the structures of protein fragments between the connected cysteine residues [79]. Protein folding simulations [3,80,81] incorporating disulfide-bond constraints also confirmed this observation. Since the stability of folded structure concerns the function of a protein, half-cystines that forming disulfide bonds are usually highly-conserved through the evolution process. Moreover, Harrison and Sternberg [82] reported that the regularities in disulfide-bridged β-sheets and in cysteine clusters can be used to classify the folds of disulfide-rich proteins. In the study of Chuang *et al*. [50] and van Vlijmen *et al*. [83], they also showed that proteins with similar disulfide bonding patterns share similar folds.

For a protein with B disulfide bridges (i.e. $2B$ oxidized cysteines), the number of possible disulfide connectivity patterns N_p can be formulated as follows:

$$N_p = (2B\text{-}1)!! \quad (13)$$

Theoretically, the number of existing disulfide connectivity patterns should increase rapidly with the number of disulfide bridges B. However, according to the statistical data obtained from SSDB database [50], the observed numbers of patterns in PDB peaked at 5 disulfide bridges, but only 45 patterns were observed while 945 possible patterns were expected. These results imply that the disulfide connectivity pattern of a protein sequence could be predicted from a limited set of templates.

Since the cysteines are highly conserved for disulfide bridges and the connectivity is extremely biased, the separation pattern between cysteine residues could be used as a descriptor to describe the global information of a protein. Using cysteine separations only, Zhao *et al.* [53] and Chen *et al.* [54] obtained comparable or higher accuracy than those local-sequence based approaches. Their results suggest that, for disulfide connectivity prediction, the cysteine separation pattern is as informative as the sequence context of a protein.

Additional Features: Secondary Structure, Solvent Accessibility, Protein Length

In addition to the sequence-based features described above, some structural features such as secondary structure, solvent accessibility, protein length are also frequently used. However, these features are either less informative or also unknown when tackling the problem. As a result, incorporating these features leads to no crucial effects for the prediction of disulfide bridges, although sometimes it slightly improves the prediction accuracy.

The secondary structure annotations are usually derived from the Dictionary of Secondary Structure of Protein (DSSP) [84], and reduced into three classes: 1) helix (H)-alpha helix, 3/10 helix and pi helix; 2) coil (C)-hydrogen bonded turn, bend and coil; and 3) sheet (E)-beta-bridge and extended strand. In the work of Fiser and Simon [37], the correlation between the oxidation state of cysteines residues and their distribution in secondary structural elements was assessed. It seems that the high frequency of disulfide bridges between linked coils (C-C) or a coil and a regular structure (C-H or C-E) are most frequently observed. This explains the important role of disulfide bridges for the protein three dimensional structures by stabilizing the folded structure covalently. In the works of Ferre *et al.* [55] and Cheng *et al.* [56], using DSSP-derived secondary structure

annotation significantly improved the prediction accuracy. However, if PSIPRED-predicted secondary structures were used instead, there was no noticeable improvement. Nevertheless, the effects of solvent accessibility and protein length are seldom mentioned or discussed, except Mucchielli-Giorgi *et al.* reported that a large normalized size of protein is a negative parameter to the formation of disulfide bridges [40].

LiMITATIONS OF DISULFIDE CONNECTIVITY PREDICTION

There are two major categories for the methods of disulfide connectivity prediction: 'pattern-wise' [47,53,54] and 'pair-wise' [55-58] approaches. The "pattern-wise" methods take the whole protein as a unit directly and rank alternative connectivity patterns, and can easily include global information. On the other hand, the 'pair-wise' methods lack the overview of the whole protein and are usually limited to the scope of local environments of cysteines. However, the pattern-wise methods often suffer from the problem of insufficient data, particularly when the number of disulfide bonds increases. For proteins with 5 disulfide bonds, there are some patterns that only have one instance in the dataset. These patterns are not likely to be predicted correctly by pattern-wise methods because there is no enough information for model training. On the other hand, pair-wise methods can still predict the connectivity of these proteins correctly, since the pattern can be assembled by the bonding pairs predicted separately. Moreover, the imbalance situation between the positive and negative data differs for pair-wise and pattern-wise methods. As to a protein with B disulfide bonds, the positive/negative ratio is 1: (2B-2) for pair-wise encoding. However, for the pattern-wise encoding, the imbalance is more severe, since there is only one correct pattern among the (2B-1)!! generated entries. Taking B=5 for example, the positive/negative ratio is only 1: 8 in pair-wise encoding. With the same bond number B in pattern-wise encoding, there are 945 entries where the positive/negative ratio is 1:944. Such severe imbalance can bias the learning process and result in poor models.

CONCLUSIONS AND DISCUSSION

From the viewpoint of bioinformatics, the prediction task could be transformed as the extraction of knowledge from existing data. Thus the performance of predictor is determined by 1) the way to formulate the problem, 2) the information techniques used to extract knowledge, 3) the descriptor used to encode data, and 4) the domain knowledge incorporated during the mechanism.

Take the prediction of disulfide connectivity for example, infer the connectivity by weighted graph matching or rank between the alternative patterns decides the way of data encoding (pattern-wise or pair-wise), while each encoding schemes have its strength and weakness. The information technique used, surprisingly, is a relatively irrelevant factor determining the prediction performance. Indeed, NN and SVM seem to be better tools for capturing the information in data comparing with statistical methods or stochastic optimizations. However, it is very difficult to tell which one is more powerful among machine learning methods such as NN and SVM. To be noticed, a jury system above multiple predictors usually leads to a slightly improvement.

The most crucial factor in prediction is the feature used. Generally, features containing more information lead to higher accuracy. For the prediction of disulfide bonding states and connectivity, methods using sequence profiles usually achieve higher accuracy than using single sequence only, since evolutionary information is also contained. However, when the amino acid composition is used in these two problems, the effects are quite different. For bonding state prediction, the amino acid composition is a good descriptor to discriminate between the proteins tends to be oxidized or reduced, and the prediction accuracy is satisfactory even with this feature only. On the other hand, the connectivity prediction involves several cysteine pairs at a time, but the amino acid composition does not contain specific information for each cysteine pair. Thus, only minor improvements are obtained when incorporating amino acid composition in the prediction of disulfide connectivity. Besides, the attribute of feature also matters. Local features and global features carry the information of different aspects, and are usually complementary to each other. The best performance is always achieved when features of local and global information are well integrated.

At last, the domain knowledge, e.g. the conservation of cysteines, the "all or none" rule, the C-xx-C motif, and the even number constraint of oxidized cysteines, is also an essential factor to perform a good prediction. Domain knowledge is usually obtained from observations or the essence of the problem, and could be used to refine the prediction results. Without domain knowledge, the prediction usually becomes a pure machine learning project, and sometimes the results obtained are biologically unfeasible.

In this paper, studies addressed on the prediction of disulfide bridges are introduced. The strategies adopted and the effects of different features are discussed. The basic ideas of these methods and the observations from their results may enlighten further studies when solving biological problems with bioinformatics approaches. Several web-interfaced tools are also available for these methods. The predicted disulfide information may be useful for advanced researches in protein structure prediction, protein structure modeling, and protein engineering.

REFERENCES

[1] Nakamoto, H. and Bardwell, J. C. A. (**2004**) *Biochim. Biophys. Acta*, 1694, 111-119.
[2] Anfinsen, C. B. (**1973**) *Science*, *181*, 223-230.
[3] Huang, E. S., Samudrala, R. and Ponder, J. W. (**1999**) *J. Mol. Biol.*, *290*, 267-281.
[4] Arolas, J. L., Popowicz, G. M., Bronsoms, S., Aviles, F. X., Huber, R., Holak, T. A. and Ventura, S. (**2005**) *J. Mol. Biol.*, *352*, 961-975.
[5] Cemazar, M., Daly, N. L., Haggblad, S., Lo, K. P., Yulyaningsih, E. and Craik, D. J. (**2006**) *J. Biol. Chem.*, In press.
[6] Robinson, A. S. and King, J. (**1997**) *Nat. Struct. Biol.*, *4*, 450-455.
[7] Vieille, C. and Zeikus, G. J. (**2001**) *Microbiol. Mol. Biol. Rev.*, *65*, 1-43.
[8] Beeby, M., Connor, B. D., Ryttersgaard, C., Boutz, D. R., Perry, L. J. and Yeates, T. O. (**2005**) *PLoS Biol.*, 3, e309.
[9] Colgrave, M. L. and Craik, D. J. (**2004**) *Biochemistry*, *43*, 5965-5975.
[10] Mansfeld, J., Vriend, G., Dijkstra, B. W., Veltman, O. R., Van den Burg, B., Venema, G., Ulbrich-Hofmann, R. and Eijsink, V. G. H. (**1997**) *J. Biol. Chem.*, *272*, 11152-11156.

[11] Zhou, H., Wang, W. and Luo, Y. (**2005**) *J. Biol. Chem.*, *280*, 11303-11312.
[12] Tigerstrom, A., Schwarz, F., Karlsson, G., Okvist, M., Alvarez-Rua, C., Maeder, D., Robb, F. T. and Sjolin, L. (**2004**) *Biochemistry*, *43*, 12563-12574.
[13] Siddiqui, K. S., Poljak, A., Guilhaus, M., Feller, G., D'Amico, S., Gerday, C. and Cavicchioli, R. (**2005**) *J. Bacteriol.*, *187*, 6206-6212.
[14] Zavodszky, P., Kardos, J., Svingor, A. and Petsko, G. (**1998**) *Proc. Natl. Acad. Sci. USA*, *95*, 7406-7411.
[15] Hamdane, D., Kiger, L., Dewilde, S., Uzan, J., Burmester, T., Hankeln, T., Moens, L. and Marden, M. C. (**2005**) *FEBS J.*, *272*, 2076-2084.
[16] Villafranca, J. E., Howell, E. E., Oatley, S. J., Xuong, N. and Kraut, J. (**1987**) *Biochemistry*, *26*, 2077-2082.
[17] Wells, J. A. and Powers, D. B. (**1986**) *J. Biol. Chem.*, *261*, 6564-6570.
[18] Mitchinson, C. and Wells, J. A. (**1989**) *Biochemistry*, *28*, 4807-4815.
[19] Van den Burg, B., Dijkstra, B. W., van der Vinne, B., Stulp, B. K., Eijsink, V. G. H. and Venema, G. (**1993**) *Protein Eng.*, *6*, 521-527.
[20] Wedemeyer, W. J., Welker, E., Narayan, M. and Scheraga, H. A. (**2000**) *Biochemistry*, *39*, 4207-4216.
[21] Kadokura, H., Katzen, F. and Beckwith, J. (**2003**) *Annu. Rev. Biochem.*, *72*, 111-135.
[22] Rietsch, A. and Beckwith, J. (**1998**) *Annu. Rev. Genet.*, *32*, 163-184.
[23] Mason, J. M., Cliff, M. J., Sessions, R. B. and Clarke, A. R. (**2005**) *J. Biol. Chem.*, *280*, 40494-40499.
[24] Daly, N. L., Clark, R. J. and Craik, D. J. (**2003**) *J. Biol. Chem.*, *278*, 6314-6322.
[25] Zheng, M., Slund, F. and Storz, G. (**1998**) *Science*, 279, 1718.
[26] Barbirz, S., Jakob, U. and Glocker, M. O. (**2000**) *J. Biol. Chem.*, *275*, 18759-18766.
[27] Cumming, R. C. andon, N. L., Haynes, P. A., Park, M., Fischer, W. H. and Schubert, D. (**2004**) *J. Biol. Chem.*, *279*, 21749-21758.
[28] Fariselli, P. and Casadio, R. (**2001**) *Bioinformatics*, *17,* 957-964.
[29] Fiser, A., Cserzo, M., Tudos, E. and Simon, I. (**1992**) *FEBS Lett.*, *302*, 117-120.
[30] Muskal, S. M., Holbrook, R. S. and Kim, S. H. (**1990**) *Protein Eng.*, *3*, 667-672.
[31] Fariselli, P., Riccobelli, P. and Casadio, a. R. (**1999**) *Proteins*, *36*, 340-346.
[32] Rumelhart, D., Hinton, G. and Williams, R. (**1986**) *Nature*, *323*, 533-537.
[33] Rost, B. and Sander, C. (**1994**) *Proteins: Struct. Funct. Genet.*, *19*, 55-72.
[34] Krogh, A. and Sollich, P. (**1997**) *Phys. Rev.*, E55, 811-825.
[35] Thornton, J. M. (**1981**) *J. Mol. Biol.*, *151*, 261-287.
[36] Fahey, R. C., Hunt, J. S. and Windham, G. C. (**1977**) *J. Mol. Evol.*, 10, 155-160.
[37] Fiser, A. and Simon, I. (**2000**) *Bioinformatics*, *16*, 251-256.
[38] Martelli, P. L., Fariselli, P., Malaguti, L. and Casadio, R. (**2002**) *Protein Sci.*, *11*, 2735-2739.
[39] Martelli, P. L., Fariselli, P., Malaguti, L. and Casadio, R. (**2002**) *Protein Eng.*, *15*, 951-953.
[40] Mucchielli-Giorgi, M. H., Hazout, S. and Tuffery, P. (**2002**) *Proteins*, *46*, 243-249.
[41] Song, J.-N., Wang, M.-L., Li, W.-J. and Xu, W.-B. (**2004**) *Biochem. Biophys. Res., Commun.*, *318*, 142-147.
[42] Chen, Y.-C., Lin, Y.-S., Lin, C.-J. and Hwang, J.-K. (**2004**) *Proteins*, *55*, 1036-1042.
[43] Abe, Y., Odaka, M., Inagaki, F., Lax, I., Schlessinger, J. and Kohda, D. (**1998**) *J. Biol. Chem.*, *273*, 11150-11157.
[44] Papadimitrou, C. H. and Steiglitz, K., *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs, NJ: Prentice-Hall, 1982.
[45] Gabow, H., "Implementation of Algorithms for Maximum Matching on Nonbipartite Graphs," in *Computer Science Department*: Stanford University, 1973.
[46] Fariselli, P., Martelli, P. L. and Casadio, R., "A neural network-based method for predicting the disulfide connectivity in proteins," in *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies (KES 2002)*, vol. 1, E. Damiani, Ed.: IOS Press, 2002, pp. 464-468.
[47] Vullo, A. and Frasconi, P. (**2004**) *Bioinformatics*, *20*, 653-659.
[48] Frasconi, P., Gori, M. and Sperduti, A. (**1998**) *IEEE Trans. Neural Networks*, *9*, 768-786.
[49] Vullo, A. and Frasconi, P. (**2003**) *J. Bioinformatics Comput. Biol.*, *1*, 411-431.
[50] Chuang, C.-C., Chen, C.-Y., Yang, J.-M., Lyu, P.-C. and Hwang, J.-K. (**2003**) *Proteins*, *53*, 1-5.
[51] Antuch, W., Guntert, P., Billeter, M., Hawthorne, T., Grossenbacher, H. and Wuthrich, K. (**1994**) *FEBS Lett.*, *352*, 251-257.
[52] Gilquin, B., Lecoq, A., Desne, F., Guenneugues, M., Zinn-Justin, S. and Menez, A. (**1999**) *Proteins: Struct. Funct. Genet.*, *34*, 520-532.
[53] Zhao, E., Liu, H.-L., Tsai, C.-H., Tsai, H.-K., Chan, C.-h. and Kao, C.-Y. (**2005**) *Bioinformatics*, *21*, 1415-1420.
[54] Chen, Y.-C. and Hwang, J.-K. (**2005**) *Proteins*, *61*, 507-512.
[55] Ferre, F. and Clote, P. (**2005**) *Bioinformatics*, *21*, 2336-2346.
[56] Cheng, J., Saigo, H. and Baldi, P. (**2006**) *Proteins: Struct. Funct. Bioinformatics*, *62*, 617-629.
[57] Baldi, P., Cheng, J. and Vullo, A., *Large-scale prediction of disulphide bond connectivity*. Cambridge: MIT Press, 2005.
[58] Tsai, C.-H., Chen, B.-J., Chan, C.-h., Liu, H.-L. and Kao, C.-Y. (**2005**) *Bioinformatics*, *21*, 4416-4419.
[59] Petersen, M. T. and al., e. (**1999**) *Protein Eng.*, *12*, 535-548.
[60] Jones, D. T. (**1999**) *J. Mol. Biol.*, *292*, 195-202.
[61] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (**1997**) *Nucleic Acids Res.*, *25*, 3389-3402.
[62] Gribskov, M. and Robinson, N. (**1996**) *Comput. Chem.*, *20*, 25-34.
[63] Rost, B. and Sander, C. (**1993**) *J. Mol. Biol.*, *232*, 584-599.
[64] Chen, B.J., Tsai, C.H., Chan, C.H. and Kao, C.Y. (**2006**) *Proteins: Struct. Funct. Bioinformatics*, in press.
[65] Mallick, P., Boutz, D. R., Eisenberg, D. and Yeates, T. O. (**2002**) *Proc. Natl. Acad. Sci.*, *99*, 9679-9684.
[66] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (**2000**) *Nucleic Acids Res.*, *28*, 235-242.
[67] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (**1990**) *J. Mol. Biol.*, *215*, 403-410.
[68] Fischer, D. and Eisenberg, D. (**1996**) *Protein Sci.*, *5*, 947-955.
[69] Smith, T. F. and Waterman, M. S. (**1981**) *J. Mol. Biol.*, *147*, 195-197.
[70] Rost, B. and Sander, C. (**1994**) *Proteins: Struct. Funct. Bioinformatics*, *20*, 216-226.
[71] Rost, B., Casadio, R., Fariselli, P. and Sander, C. (**1995**) *Protein Sci.*, *4*, 521-533.
[72] Rost, B., Casadio, R. and Fariselli, P. (**1996**) *Protein Sci.*, *5*, 1704-1718.
[73] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (**1997**) *Nucleic Acids Res.*, *25*, 3389-3402.
[74] Yu, C., Wang, J., Yang, J., Lyu, P., Lin, C. and Hwang, J. (**2003**) *Proteins: Struct. Funct. Genet.*, *50*, 531-536.
[75] Yu, C., Lin, C. and Hwang, J. (**2004**) *Protein Sci.*, *12*, 1402-1406.
[76] Passerini, A. and Frasconi, P. (**2004**) *Protein Eng.*, *17*, 367-373.
[77] Abkevich, V. I. and Shakhnovich, E. I. (**2000**) *J. Mol. Biol.*, *300*, 975-985.
[78] Song, J.-N., Wang, M.-L., Li, W.-J. and Xu, W.-B. (**2004**) *Biochem. Biophys. Res. Com.*, *318*, 142-147.
[79] Anfinsen, C. and Scheraga, H. (**1975**) *Adv. Protein Chem.*, *29*, 205-299.
[80] Abkevich, V. and Shaknovich, E. (**2000**) *J. Mol. Biol.*, *300*, 975-985.
[81] Skolnick, J., Kolinski, A. and Ortiz, A. R. (**1997**) *J. Mol. Biol.*, *265*, 217-241.
[82] Harrison, P. M. and Sternberg, M. J. E. (**1996**) *J. Mol. Biol.*, *264*, 603-623.
[83] Vlijmen, H. W. T. v., Gupta, A., Narasimhan, L. S. and Singh, J. (**2004**) *J. Mol. Biol.*, *335*, 1083-1092.
[84] Kabsch, W. and Sander, C. (**1983**) *Biopolymers*, *22,* 2577-2673.
[85] Per J. Kraulis, (**1991**) *J. Appl. Cryst.*, *24*, 946-950.

Received: April 20, 2006 **Revised: May 08, 2006** **Accepted: May 15, 2006**