國立交通大學

生物資訊所

博士論文

利用蛋白質序列預測雙硫鍵鍵結情形

Prediction of Disulfide Connectivity from Protein Sequences

研究生：陳玉菁（Yu-Ching Chen）

指導教授：黃鎮剛教授（Jenn-Kang Hwang）

中華民國九十六年十一月

利用蛋白質序列預測雙硫鍵鍵結情形

Prediction of Disulfide Connectivity from Protein Sequences

研 究 生：陳玉菁　　Student: Yu-Ching Chen

指導教授：黃鎮剛　　Advisor: Jenn-Kang Hwang

國立交通大學
生物資訊所
博士論文

A Dissertation
Submitted to Institute of Bioinformatics
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Ph.D.
in

Bioinformatics

November 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年十一月

利用蛋白質序列預測雙硫鍵鍵結情形

學生：陳玉菁　　指導教授：黃鎮剛博士

國立交通大學 生物資訊所 博士班

摘　　要

雙硫鍵對於蛋白質結構的穩定與蛋白質功能的調控有很大的影響力。目前蛋白質序列的資料量遠多於蛋白質結構的數目；因此若能發展計算的方法，從蛋白質的序列來預測雙硫鍵的配對情形(disulfide connectivity)，將有助於雙硫鍵蛋白質的研究。然而從蛋白質序列直接預測雙硫鍵配對情形的困難度在於雙硫鍵並不是序列上兩鄰近半胱氨酸(half-cystine) 的鍵結，而是鄰近空間中兩個半胱氨酸的鍵結，因此雙硫鍵配對的預測充滿著挑戰。然而科學家們也研究各式各樣的方法要從蛋白質序列來解開雙硫鍵配對情形的問題，但目前用來預測的方法都局限於在雙硫鍵個數小於等於五的蛋白質中。因為隨著蛋白質中雙硫鍵個數的增加，雙硫鍵配對情形的類別變多，預測更為困難。

在此研究中，開發了一個預測雙硫鍵配對情形的方法並命名為 S-S predictor，其結合了序列比對與機器學習法。一方面利用序列比對的優點，比對出與欲預測雙硫鍵配對情形蛋白質序列同源性且已知結構的雙硫鍵蛋白；如此整合了序列與結構的關連性達到預測目的。另一方面，當欲預測雙硫鍵之蛋白質無法比對出同源性蛋白質時，就使用支持向量法；本研究中找出有用的特徵值來做預測。例如利用兩兩半胱氨酸周圍胺基酸的演化資訊、兩兩半胱氨酸間在序列上的距離，還有整條蛋白質序列二十種胺基酸的變化。使用此方法，在序列相同程度小於 30%的蛋白質作預測，其正確率就雙硫鍵配對情形正確才算正確可達 0.81（Q_p），而雙硫鍵的正確率達 0.84（Q_c）；此正確率超越其它方法，且無雙硫鍵個數限制。S-S predictor 的網址是 http://140.113.239.214/~ssbond，對研究雙硫鍵配對情形的使用者來說，是一個方便實用的預測方法。

Prediction of Disulfide Connectivity from Protein Sequences

Student: Yu-Ching Chen Advisor: Dr. Jenn-Kang Hwang

Institute of Bioinformatics
National Chaio Tung University

Abstract

The disulfide bonds have great influences in stabilizing protein structures and regulating protein functions. At present there is a gap between protein sequences and protein structures; therefore, it would be a great help to predict disulfide connectivity from protein sequences. However, the difficulties in predicting disulfide connectivity from protein sequences lie in the nonlocal properties of the disulfide bridges that involve cysteine pairs at large sequence separation. Although many scientists develop various methods to solve this problem; it is still a challenge. These methods are limited by the number of disulfide bonds should equal or less than five, because as the increase of disulfide bonds in proteins the number of disulfide connectivity grows rapidly, and it is more difficult to predict disulfide connectivity.

In this research, I developed a method to predict disulfide connectivity and named S-S predictor; it combines sequence alignment method and machine learning method. The searching dataset of sequence alignment are disulfide proteins with known structures; therefore, the advantages of this method integrate sequence and structure information to predict disulfide connectivity. On the other hand, when homologs of query protein can not be found, the support vector machines are used to solve problem. I found some useful feature vectors in this research; such as the coupling evolutionary information between the local sequence environments of cysteine pairs, the cysteines sequence separations, and the global sequence descriptor, amino acid content. The performance of S-S predictor based on a dataset whose sequence identity between two proteins is lower than 30% is 0.81 and 0.84 in Q_p and Q_c, respectively. The accuracy of this method is higher than other method, and there is no limitation on the number of disulfide bond. S-S predictor is a useful and practical tool to study disulfide connectivity, and the website of S-S predictor is http://140.113.239.214/~ssbond.

Acknowledgement

似乎是求學生涯的最後階段，亦是生命中的一個段落，回首這些日子有甜有苦，但心中更是浮現許多要感謝的人。

感謝我的指導教授黃鎮剛教授，感謝老師在博士班生涯中所給予的指導、信任與幫助；從老師身上看到研究的熱忱與學習到對問題的探究、還有做研究的sense。此外，老師不光是指導我們，也總是盡其所能提供學生們良好的研究環境;使我們較一般學生有更多的機會與外界接觸、拓展視野。真的由衷感謝老師。

在此也要特別感謝林彩雲教授每每在論文上給予的指導與幫助，澎慧玲教授對研究與生活上的關心與建議，及有任何疑難雜症都可請問盧錦隆教授，還有楊進木教授願意提供的機會與幫忙。真的很感謝各位的指導，謝謝各位老師。

轉眼間要跟實驗室的大家告別了！謝謝景盛、勇欣、朝鈞、彥偉、志豪、少偉、小操、建華、志鵬、蔚倫、星男、啟文、小胖等學弟妹大家一路的陪伴；研究生活有大家相聚，真是熱鬧許多。也謝謝 kemp、丸子所推行的排球運動，與大家一同打球的日子，真是令我非常懷念。

另外也常想起與師母、欣怡、映羽、蔚倫、松桓還有啟文一起聚會的時光，謝謝師母與欣怡為大家擺上豐盛的午餐、絕佳的環境、悠揚的音樂，並帶領我們一同查經；分享著快樂與淚水，很是輕鬆、很享受更是好吃。學弟妹們若有興趣（吃）的話，可跟松桓報名。

也感謝梅竹小組的爸爸、媽媽們，許多年來謝謝你們的關心以及代禱；讓我與妹妹有機會認識神。總是在大家的身上看到謙虛、有愛與熱情，還有對聖經真理行在生活中的實際，很是感動。也難忘社青團契的弟兄姐妹，大家一同郊遊踏青、打球、吃飯、分享與成長。點點滴滴的生活實在難以數算；唯有感謝主。

還有要謝謝爸爸媽媽無微不至的愛，總是毫無條件的付出，常常有新鮮甜美的水果及宵夜可吃；也謝謝生命中有幽默的弟弟政賢、可愛的妹妹玉茹之同在，讓家裡總是充滿著溫暖的噪音與歡笑。謝謝你們！使我可以快樂、專心作研究，覺得好幸福！

最後感謝主給予豐富的生命，於研究或生活中軟弱時總是用話語來帶領、安慰我，使我有力量、有能力來面對；也藉著各樣人、事、物互相效益，叫愛神的人得益處。謝謝主讓我每每經歷祢自己，深知祢是又真又活的神。

將此論文獻給我主、我所愛與愛我的人。

Contents

Abbreviations

3D	: 3 dimension
CSP	: Cysteine separation profile
GA	: Genetic algorithm
GM	: Graph matching
MC	: Monte-carlo
MCGM	: Monte-carlo graph matching
NCBI	: National center for biotechnology information
NMR	: Nnclear magnetic resonance
Nm	: Nanometer
NN	: Neural network
NNGM	: Neural network graph matching
PDB	: Protein data back
PSI-BLAST	: Position specific iterative basic local alignment search tool
PSSM	: Position specific scoring matrix
RBF	: Radial basis function
RNN	: Recursive neural network
SVM	: Support vector machine
SVR	: Support vector regression
SWISS-PROT	: A protein sequence database of a high level of annotation

Chapter1

GENERAL INTRODUCTION

A protein sequence is constituted by twenty amino acids, and cysteine is one of the twenty. The particular side chain of cysteine is thiol group (-SH); however, the thiol-from cysteine is the most reactive amino acid under physiological conditions. When the oxidation of two thiol groups forms a disulfide bond (S-S) [Fig. 1], it is a covalent bond and can connect two distant cysteines. In fact, the formation of disulfide bond is one type of post-translational modification, and the protein is in a reducing redox environment. Therefore, most proteins containing stable disulfide bond in bacteria can be found in extracytoplasmic compartments or secreted into the external medium, disulfide proteins in eukaryotic cells are located in compartment such as the plasma membrane, the endoplasmic reticulum or secreted into external medium. However, oxidants and proteolytic enzyme in the extracellular environment can inactivate proteins; disulfide bond can protect proteins from damage and increase their half-life by stabilizing protein structure.[1] Generally, cysteine residues can be classified as free cysteine, ligand-bound cysteine, inter-chain half cystine, and intra-chain half cystine.

Disulfide bonds have great influences in determining protein structure and mediating biological function. In structure, disulfide bonds play a vital role in the folding process of many proteins.[2,3] Anfinsen reduced whole disulfide cystines of pancreatic bovine ribonuclease in vitro, then re-oxidized them, correct disulfide connectivity and native structure was restored.[4] Since disulfide bond has its specific $C_\beta - S_\gamma$ and $S_\gamma - S_\gamma$ bond lengths, 1.81 and 2.04 Å, respectively, $C_\beta - S_\gamma - S_\gamma$ bond angle of 104.15°[5], and a single disulfide bridge can stabilize the protein by 2-5 kcal/mol.[6,7] According to disulfide bond number and location in a protein, they can contribute to the thermodynamic stability of the 3D structure and increase protein stability. Because disulfide bonds add strong structural constraints, reduce the search in the conformational space.[8,9] Proteins containing disulfide bond have diverse functions such as hydrolase, inhibitor, hormones and toxins. For example, ligand-bound cysteines fix heme in cytochromes.[10] Enzyme thioredoxin, which is related with photosynthesis, seed germination, transcription, acts as a regulatory switch of target proteins by reducing their disulfide bonds.[11] Cone snails are venomous mollusks and their venoms contain disulfide-rich peptide conotoxin having two disulfide bonds.[12] Thus, the knowledge of the disulfide connectivity is vital in the study of structure and function of proteins.

At present, there are several kinds of methods, experimental determination and

machine learning prediction, to solve disulfide connectivity in proteins. The experimental methods include chemical methods, nuclear magnetic resonance (NMR) spectroscopy, and X-Ray crystallography. From chemical experiments disulfide connectivity can be inferred by series chemical reactions. On the other hand, structures of disulfide proteins can be resolved by NMR spectroscopy[13] and X-Ray crystallography; consequently, disulfide connectivity can be detected from protein structures. These methods can offer correct and more confident disulfide connectivity.

Nevertheless, the sequence-structure gap is widened rapidly as a consequence of the large-scale whole genome projects. In the absence of an experimentally determined structure, protein sequences do not report reliable information relating either the oxidized form of cysteines or disulfide bridge locations. Therefore, it is necessary and helpful to predict disulfide connectivity from protein sequence by machine learning method. First, disulfide connectivity is predicted based on graph representation of difulfide bridges, where vertices are oxidized cystines and edges represent a pair of cystines calculated from contact potential optimization.[14] Next, neural network predictions were used to replace contact potential optimization for increasing predictive power.[15] In a subsequent improvement[16], a recursive neural networks and evolutionary information were used, and cysteine separation profiles (CSPs)[17] of proteins were adopted for the prediction of disulfide connectivity.

Furthermore, secondary structure information and diresidue frequencies based on neural network were designed to solve disulfide connectivity.[18] Meanwhile, pattern-wise method using SVM based on feature vectors such as coupling between the local sequence environments of cysteine pairs, the cysteine separations, and the amino acid content were used to predict disulfide connectivity.[19] On the other hand, sequential distance between oxidized cysteines combined with SVM was also applied to determine disulfide connectivity.[20] Then, two-level models integrate SVM models and cysteins separation search to tackle the problem.[21] Nevertheless, SVM coupled with genetic algorithm (GA) for feature selection to remove noisy or irrelevant features was applied to infer disulfide connectivity.[22] Recently, support vector regression (SVR) based on multiple sequences feature vectors and predicted secondary structures are used to infer disulfide connectivity. The performance of these methods ranged from 29% to 74%, and 38% to 79% for Q_p and Q_c respectively.

In this thesis, I proposed a hybrid system, S-S Predictor, which combines PSI-BLAST method and machine learning method to study disulfide connectivity. PSI-BLAST method is used to search homologous disulfide proteins with known structures; meanwhile, structures record the information of disulfide connectivity and in general similar sequences may have analogous structures. Therefore, utilizing sequence-to-structure mapping can infer the disulfide connectivity of query proteins.

However, not all of the query proteins can find homologs, if homologs can not be found; feature vectors from protein sequence are prepared to feed into SVM to infer disulfide connectivity. The flowchart is shown in Figure 2.

Chapter 2

Prediction of disulfide connectivity using blast method

INTRODUCTION

Disulfide bonds are known to play an important structural role in stabilizing protein conformations by reducing the number of unfolded conformations.[3,23-28] Since disulfide bonds impose geometrical constraints on the protein backbones, the disulfide patterns may well dictate to a certain degree the overall three-dimensional (3D) protein structures. Indeed, recent works[29-32] have shown that the disulfide patterns are closely related to protein structures. There are a number of efforts[8,9,33-40] to model disulfide bridges or disulfide-rich systems either from protein sequences or from 3D structures. On the other hand, disulfide bonds are more than just inert structural motifs – it is known that the functions of some secreted soluble proteins and cell-surface receptors depend on the cleavage of their disulfide bonds.[1] Therefore, the knowledge of the disulfide patterns is vital in the study of structure and function of proteins.

Nowadays many computational approaches[5,14,15,17-19,21,22,41-43] are used in predicting disulfide connectivity. They in terms of contact potential[14], neural network[16,18], support vector machine[19,22], cysteine separation profile computations[17] and genetic algorithms methods[22] etc., utilize restricted training data for prediction to prevent overtraining. Now the better average performance can reach 74% in Q_p and 79% in Q_c within B

= 2 …5[22]. On the other hand, more and more protein structures are solved in the Protein Data Bank (PDB), which is the worldwide depository of information about the three-dimensional structures of large biological molecules including proteins and nucleic acids.[44] When the protein structure is built, disulfide connectivity is also determined; therefore, sequence-to-structure mapping techniques can be used to identify potential disulfide bonds. The purpose of this research is to take the advantages of huge PDB data, and extracts abundant information from it. Hence, most information can be used to predict disulfide connectivity, and the performance of this work can be complementary with machine learning method.

In this work, homologous search tool, PSI-BLAST, was used to search evolutionarily related disulfide proteins with PDB structure; further, according to evolutionary relationship within PDB proteins infers the disulfide connectivity of query proteins. On the other hand, the contribution of residues between disulfide cysteines and the geometrical pattern of disulfide cysteines are investigated, too.

MATERIALS AND METHODS

Disulfide connectivity prediction

PSI-BLAST method was used to find homologs of query protein, and the searching dataset includes proteins with disulfide bonds from PDB structure. There are three kinds of methods to infer the disulfide connectivity based on the PSI-BLAST consensus, such as, disulfide pair method which applies cystein pair as a unit, disulfide pattern method which uses disulfide pattern as a unit, and a hybrid method which combines disulfide pair method and disulfide pattern method to predict disulfide connectivity.

Disulfide pair method

I used the notation $\phi = \{C_1, C_2\}$ to denote the cysteine pair comprising C_1 and C_2. For each cysteine pair, there are two possible bonding states: $\sigma_1 = C_1 \oplus C_2$, where $\oplus$ denotes a disulfide bridge between C_1 and C_2, and $\sigma_2 = C_1 \otimes C_2$, where $\otimes$ denotes no disulfide bridge between C_1 and C_2. In this way, I can define the disulfide connectivity patterns in terms of the bonding states.

The connectivity matrix M is defined in terms of the bonding states, $C_p \Theta C_q$, which are predicted by PSI-BLAST method. The initial matrix elements M_{pq} are set to 0, p and q that represent the order of cysteines in query protein sequence. The rules to construct the matrix are:

$$M_{pq} = M_{pq} + 1, \text{ if } \Theta = \oplus \qquad (1)$$

$$M_{pq} = M_{pq}, \text{ if } \Theta = \otimes \qquad (2)$$

The score Ω_T of the disulfide connectivity pattern T was computed from M by

$$\Omega_T = \sum_{i<j}^{T} {}' M_{ij} \qquad (3)$$

Where $\sum'$ indicates that any two index pairs (i, j), and (i', j') under the summation sign should satisfy the requirements $i \neq i'$ and $j \neq j'$. The disulfide pattern with the maximal score, i.e. $\max\{\Omega_T\}$, is taken as the prediction.

Disulfide pattern method

For a disulfide protein with n cysteines (i.e., $c_1, c_2, ..., c_n$), its disulfide pattern is denoted by $(c_i c_j, c_{i'} c_{j'}, ...)$, where $c_i c_j$ designates a disulfide bridge formed between cysteines i and j, and the number of possible disulfide pattern is $N_p = \frac{\binom{2B}{2}\binom{2B-2}{2}\binom{2B-4}{2}\cdots\binom{2}{2}}{B!} = (2B-1)!! = \prod_{i \le B}(2i-1)$. However, the disulfide pattern is taken as a unit and can be directly predicted by PSI-BLAST method. For example, homologs of query protein can be found after PSI-BLAST; examine homologs to see if disulfide cysteines of the query protein are all aligned with disulfide cysteines of homologs. Therefore, the frequency of each disulfide pattern, T, was calculated from homologs whose disulfide cysteines can be aligned with all disulfide cysteines of the query protein and the frequency was assigned as score Ω_T

of the disulfide connectivity pattern T. Finally, the disulfide pattern with the maximal score, i.e. $\max\{\Omega_T\}$, was taken as the result.

Hybrid method

This method combines disulfide pair method and disulfide pattern method. When the disulfide connectivity of a query protein can be predicted from disulfide pattern method, the disulfide pair method will be neglected. On the other hand, if whole disulfide cysteines of the query protein are not totally aligned with homologs, the pair method uses disulfide pairs to present and predict disulfide connectivity.

Performance Indices

To evaluate the performance of the classifiers, use two assessment of measures:[14,16] Q_c, a disulfide-bridged measure of the fraction of the correctly predicted disulfide bridges, and Q_p, a protein-based measure of the fraction of proteins whose global disulfide pattern is correctly predicted. Q_p is the more stringent performance index. Specifically, they are defined as

$$Q_c = \frac{1}{N_c}\sum_{i=1}^{N_c}\delta_{c_i} \qquad (4)$$

$$Q_p = \frac{1}{N_p}\sum_{i=1}^{N_p}\delta_{p_i} \qquad (5)$$

where δ_{c_i} is defined for the i^{th} disulfide bridge as

$$\delta_{c_i} = \begin{cases} 1, \text{ if the } i^{th} \text{ predicted disulfide bridge is correct} \\ 0, \text{ if the } i^{th} \text{ predicted disulfide bridge is incorrect} \end{cases},$$

and N_c is the total number of disulfide bridges. Similarly, δ_{p_i} is defined for the i^{th} disulfide proteins as

$$\delta_{p_i} = \begin{cases} 1, \text{ if the predicted connectivity pattern of the } i^{th} \text{ protein is correct} \\ 0, \text{ if the predicted connectivity pattern of the } i^{th} \text{ protein is incorrect} \end{cases},$$

and N_p is the total number of disulfide proteins.

There is another assessment of measure: Q_s, a pattern-based measure of the number of proteins whose global disulfide pattern is correctly predicted over the number of proteins which disulfide connectivities are predicted by PSI-BLAST method. It is defined as

$$Q_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \delta_{s_s} \qquad (6)$$

where δ_{s_i} is defined for the i^{th} disulfide protein as

$$\delta_{s_i} = \begin{cases} 1, \text{if the predicted connectivity pattern of the } i^{th} \text{ protein is correct} \\ 0, \text{if the predicted connectivity pattern of the } i^{th} \text{ protein is incorrect} \end{cases},$$

and N_s is the total number of disulfide proteins whose disulfide connectivities are predicted by blast method.

Datasets

There are three datasets used in this research; one named SP39-ID30, another named CYS040307-NR, and the other named CYS090506.

SP39-ID30

In order to compare my methods with previous works[14,16], I followed the same criteria in selecting the sequences from the SWISS-PROT database release No. 39.[45] The constructed dataset contains only the sequences with experimentally verified intra-chain disulfide bridge annotations, and excludes the sequences whose disulfide bonds are assigned as 'probable', 'potential' or 'by similarity'. I consider the sequences with 2 to 5 disulfide bridges ($B = 2,\ldots,5$), which account for more than 80% of SWISS-PROT sequences. The final dataset contains 482 sequences, of which 168 are with two disulfide bonds ($B = 2$), 177 three ($B = 3$), 95 four ($B = 4$) and 42 five ($B = 5$). Then group the sequences into 4 sets according to their disulfide bond number as predicting set– each set was selected in such a way that sequence homology among the sets is less than 30%, and the number of sequences of each set is approximately equal. Further, these sets are used for the 4-fold cross validation procedures as in the previous work.[14,16]

On the other hand, there are two kinds of training set for PSI-BLAST searching datasets; one is intra-B SP39-ID30 and the other is inter-B SP39-ID30. When the

disulfide bond number in training set is the same with predicting set, and these training sets are named intra-B SP39-ID30; however, disulfide proteins with $B = 2,\ldots,5$ in training set is named as inter-B SP39-ID30.

CYS040307-NR

The cys proteins in CYS040307-NR are extracted from Non-redundant PDB set, nrpdb.040307, in National Center for Biotechnology Information (NCBI), and this dataset is used in PSI-BLAST method for BLAST database. There are total 91,398 proteins by chain in nrpdb.040307, after filtering out non disulfide bond proteins and selecting non redundant proteins; there are 5,913 disulfide proteins, and is named CYS040307-NR. The dataset contains sequence with 1-26 disulfide bonds: 1,850 sequences with one disulfide bonds ($B = 1$), 1,673 two ($B = 2$), 983 three ($B = 3$), 651 four ($B = 4$), 231 five ($B = 5$), 178 six ($B = 6$), 138 seven ($B = 7$), 75 eight ($B = 8$), 38 nine ($B = 9$), 20 ten ($B = 10$), 10 eleven ($B = 11$), 7 twelve ($B = 12$), 3 thirteen ($B = 13$), 5 fourteen ($B = 14$), 4 fifteen ($B = 15$), 16 sixteen ($B = 16$), 16 seventeen ($B = 17$), 4 eighteen ($B = 18$), 4 nineteen ($B = 19$), 1 twenty-two ($B = 22$), 2 twenty-three ($B = 23$), 3 twenty-five ($B = 25$), and 1 twenty-six ($B = 26$). Figure 3(A) shows the distribution of disulfide proteins versus their disulfide bond number B.

CYS090506

The cys proteins in CYS090506 are extracted from Non-redundant PDB set, nrpdb.090506, in National Center for Biotechnology Information (NCBI), and this dataset is used in PSI-BLAST method for BLAST database. There are total 81,641 proteins by chain in nrpdb.090506, after filtering out non disulfide bond proteins; the set of 15,252 disulfide proteins, is named CYS090506. The dataset contains sequence with 1-26 disulfide bonds: 5,537 sequences with one disulfide bonds ($B = 1$), 4,021 two ($B = 2$), 2,054 three ($B = 3$), 1,634 four ($B = 4$), 635 five ($B = 5$), 560 six ($B = 6$), 334 seven ($B = 7$), 149 eight ($B = 8$), 87 nine ($B = 9$), 36 ten ($B = 10$), 8 eleven ($B = 11$), 15 twelve ($B = 12$), 5 thirteen ($B = 13$), 12 fourteen ($B = 14$), 7 fifteen ($B = 15$), 37 sixteen ($B = 16$), 66 seventeen ($B = 17$), 6 eighteen ($B = 18$), 1 nineteen ($B = 19$), 1 twenty ($B = 20$), 1 twenty-one ($B = 21$), 2 twenty-two ($B = 22$), 2 twenty-three ($B = 23$), 5 twenty-five ($B = 25$), and 1 twenty-six ($B = 26$). Figure 3(B) shows the distribution of disulfide proteins versus their disulfide bond number B.

Furthermore, in order to explore the influence of residues nearby cysteines; disulfide proteins in CYS090506 were modified and named CYS090506(w), and w represents window size around cysteine. In this dataset, the residues within window size around disulfide cysteines are kept, and other residues are replaced by symbol X in protein sequence. Symbol X has a special function in PSI-BLAST; it is used to filter out noisy residues in protein sequence and the coupling scores with other amino

acids are assigned in scoring matrix.

RESULTS AND DISCUSSION

Distribution of cysteine pairs

Figure 4 show the distribution of disulfide bridges for $B = 2,\ldots,8$ in the dataset CYS090506. When the circles locate near the diagonal mean the positions of two cysteines are nearby in the sequence, the circles deviate diagonal represent the location of cystein pairs are distant in the sequence. The positions of a disulfide pair are nearby in the sequence in the case of B = 2, 5, and 8 [Fig. 4(A, D, G)]; however, the positions of cysteines are distant in B = 4, 6, and 7 [Fig. 3(C, E, F)]. In B = 3 [Fig. 4(B)], some locations of cysteine pairs are close in sequence; some are distant. In the case B = 2 [Fig. 4(A)], two of the most popular disulfide pairs are C_1C_2 and C_3C_4, and the dominant disulfide connectivity is $[C_1C_2, C_3C_4]$. As the increase of disulfide bond number, the distributions of disulfide pairs become more complicated [Fig. 4(B-G)]. When B = 3 [Fig. 4(B)], disulfide pairs C_1C_2, C_3C_4 and C_5C_6 are the most dominant, C_2C_4 and C_3C_6 are second dominant, and the frequent disulfide pattern is $[C_1C_2, C_3C_4, C_5C_6]$. For B = 4 [Fig. 4(C)], the disulfide pairs C_2C_7 and C_1C_8 are most dominant ones, and the others are C_3C_5, C_4C_5 and

C_4C_6, then the most popular disulfide connectivity is [C_1C_8, C_2C_7, C_3C_5, C_4C_6]. For B = 5 [Fig. 4(D)], most popular disulfide pairs are C_9C_{10}, the others are C_7C_8, C_6C_7, C_1C_3, C_1C_4, and C_5C_6; two dominant disulfide connectivity are [C_1C_3, C_2C_4, C_5C_6, C_7C_8, C_9C_{10}] and [C_1C_2, C_3C_4, C_5C_6, C_7C_8, C_9C_{10}]. In B = 6 [Fig. 4(E)], the dominant disulfide pairs are C_1C_6, C_2C_3, C_4C_{12}, C_5C_{10}, C_7C_8, and C_9C_{11}; dominant disulfide connectivity is [C_1C_6, C_2C_3, C_4C_{12}, C_5C_{10}, C_7C_8, C_9C_{11}]. In the case B = 7, C_2C_4, C_7C_9, C_8C_9, C_1C_{13}, C_3C_{12}, C_5C_{14}, C_6C_{11}, and C_9C_{11} are dominant disulfide pairs, and [C_1C_{13}, C_2C_4, C_3C_{12}, C_5C_{14}, C_6C_{11}, C_7C_9, C_8C_{10}] and [C_1C_8, C_2C_{14}, C_3C_5, C_4C_{13}, C_6C_{12}, C_7C_{10}, C_9C_{11}] are dominant disulfide connectivities. Finally, in the case B = 8, the dominant disulfide pairs are C_2C_3, C_1C_4, C_9C_{11}, and $C_{12}C_{13}$; dominant disulfide connectivity are [C_1C_3, C_2C_4, C_5C_7, C_6C_8, C_9C_{11}, $C_{10}C_{12}$, $C_{13}C_{15}$, $C_{14}C_{16}$], [C_1C_4, C_2C_3, C_5C_{13}, C_6C_{16}, C_7C_{10}, C_8C_{12}, C_9C_{11}, $C_{14}C_{15}$] and [C_1C_6, C_2C_3, C_4C_5, C_6C_{15}, C_7C_{14}, C_8C_9, $C_{10}C_{11}$, $C_{12}C_{13}$].

Disulfide connectivity prediction

After homology search, there are three methods to determine disulfide

connectivity. One is disulfide pair method uses disulfide pairs to represent disulfide connectivity; the other one is disulfide pattern method directly uses disulfide connectivity as prediction unit, and finally is hybrid method. The performances of three methods are very similar no matter in Q_P [Fig. 5(A)] or Q_c [Fig. 5(B)]; however, the hybrid method is equal or slightly better than other methods. Therefore, the hybrid method is used for following experiments. It is reasonable use disulfide pattern method first, because whole disulfide cysteines are all aligned at once means this alignment is conserved and significant.

The disulfide number in searching dataset

In order to compare the performance with previous results, four cross-validations are used by PSI-BLAST in SP39-ID30. There are two different searching databases, one is intra-B dataset where the disulfide number, B, in searching database is the same with query proteins. The other is inter-B dataset whose disulfide number, B, in searching database is not only the same with query proteins but also includes other disulfide bond number proteins expect query proteins. The average accuracies by using intra-B dataset [Fig. 6(A)] is 0.22 and 0.24, and inter-B dataset [Fig. 6(B)] is 0.47 and 0.49 for Q_p and Q_c, respectively. There is an over 20% improvement with

inter-B dataset as searching dataset, implying the existence of sub patterns in disulfide connectivity. For example the homologs of protein KLK_PIG are list in table 4, and the native disulfide connectivity of KLK_PIG is [26_42, 121_190, 155_169, 180_205] where 26_42 represents the positions of cysteines in sequence, 26 and 42, forms a disulfide bond. The disulfide connectivity of KLK_PIG is constituted by piecing up the sub patterns from homologs; however, the sub patterns of KLK_PIG can be classified as four groups such as [121_190, 155_169, 180_205], [26_42, 121_190, 155_169], [26_42, 155_169], and [26_42]; they are named according to the locations of disulfide bonds in the query protein. Some protein structures were further examined [Fig. 10(A-C)] based on different sub patterns of query protein. In figure 10(A), 2kaiA offers a sub pattern [26_42] and the color of the disulfide bond is purple; besides, 2kaiB provides a sub pattern [121_190, 155_169, 180_205] and the color of disulfide bonds are green, blue and red, respectively. Combination the sub patterns of 2kaiA and 2kaiB, the native disulfide connectivity of query protein is completed. In figure 10(B), the sub pattern of 1zr0C is [26_42, 121_190, 155_169]; the colors of these disulfide bonds are purple, green, and blue, and the sub pattern of 1gvzA is [26_42, 155_169], and the colors of disulfide bonds are purple and blue. The differences between figure 10(A) and figure 10(B) are 2kai contains two protein chains, 2kaiA and 2kaiB; moreover, there is an inter-disulfide bond between china A

and chain B, and the color of this bond is CPK; also has another intra-disulfide bond [94_119] (red) connect loops. Examining figure 10(B), there is only one protein chain, 1zr0C, and has two additional intra-disulfide bonds (CPK). One locates similarly with inter-disulfide bond in 2kai, and the other lies ahead on the big helix; however, there is no intra-disulfide bond connect loops which is relative to red disulfide bond in figure 10(A). On the other hand, 1gvzA [Fig. 10(C)] involves one protein chain, when it compares with figure 2kai; it lack two disulfide bonds green and red in 2kai, and it also contains a intra-disulfide bond, which locates similarly with inter-disulfide bond in 2kai. Although the number of disulfide bonds and the locations of these disulfide bonds within these four protein chains are different, their protein structures are similar. It also implies that disulfide connectivity might be modified during evolution to maintain the protein structure to keep protein function.

Another example is HGF_HUMAN; its native disulfide connectivity is [70_96, 74_84, 128_106, 149_189, 177_201], and the sub patterns are list in table 5; however, there are three types of sub patterns, [70_96, 74_84], [149_189, 177_201], and [128_206, 149_189, 177_201], and their structures are shown in figure 11(A), (B), and (C), respectively. The structure of 2hgfA [Fig. 11(A)] is similar to the upper part of 1gmnA [Fig. 11(B)]; however, there is a sub pattern in 2hgfA, and there is no disulfide bonds in that part of 1gmnA. Furthermore, there is an overlap of sub patterns between

1gmnA and 1i71A [Fig. 11(C)], the structure of 1i71A is analogous with the lower part of 1gmnA, and there is a one more disulfide bond in 1i71A. Therefore, sub patterns of specific disulfide connectivity might carry their own structures, be modified during evolution, reveal evolutionary information, and might be used in help constructing phylogenetic tree.

In fact, not all disulfide connectivity of query proteins can be predicted by PSI-BLAST method; it needs to have homologous proteins. Therefore, a performance index Q_s is used to evaluate the accuracy based on the disulfide proteins whose disulfide connectivity can be predicted by PSI-BLAST. According to figure 7, the performance index Q_s is higher than Q_p not only in intra-B dataset [Fig. 7(A)] but also in inter-B dataset [Fig. 7(B)]. However, it means if the disulfide connectivity of a query protein can be predicted by PSI-BLAST method, the average accuracy Q_s is 0.5 and 0.72 in intra-B dataset and inter-B dataset, respectively. For this reason, homologous search is a confident method to predict disulfide connectivity.

Performance based on different searching datasets

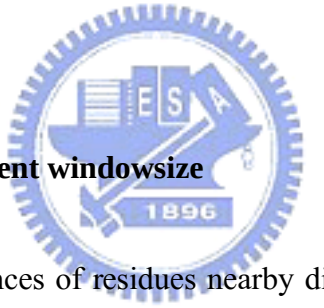
In Table 2, the average accuracy in SP39-ID30 is 0.47 and 0.49 for Q_p and Q_c,

respectively, in CYS040307-NR is 0.90 and 0.91; meanwhile, in CYS090506 is 0.95 and 0.97. As the searching datasets are larger, it includes various disulfide proteins and involves more disulfide information; therefore, the accuracy is greatly improved. It also implies when PSI-BLAST method are used to predict disulfide connectivity, the abundance of disulfide protein dataset is necessary. Otherwise, as the disulfide bond number is raised, the accuracy is decreased; it is reasonable because aligning more cysteines with homologs are getting harder and disulfide environment is slightly different in proteins with distinct disulfide bond number.

Performance based on different sequence identity

The accuracies rise in Q_p [Fig. 8(A)] and Q_c [Fig. 8(B)] because of higher similar of sequence identity. When the sequence identity among query protein and target proteins is between 0 to 20%, it doesn't have any contribution in predicting disulfide connectivity. However, in 0 to 30% the performances of Q_p and Q_c in $B = 5$ are 0.02 and 0.15 respectively; average performances are 0.29 and 0.30. Generally 30% sequence identity is a threshold to find a template in homology modeling; similarly 40% sequence identity is a threshold in disulfide connectivity prediction. Therefore, it is consistent more alike between two sequences are more

parallel in their properties such as structures, functions and so on. Furthermore, when sequence identities raised from a range of 0 to 30% to 0 to 40% , the accuracies showed sharp increases. For example, Q_p and Q_c are equal or below 40% versus disulfide bonds B in sequence identity 0-30%, and are equal or below 60% in sequence identity 0-40%. Nevertheless, the accuracies predicted from sequence identities ranging from 0-40% to 0-90% are gradually increased and sharply increased from 0-100% identity. When query and target proteins are more similar, the more reliable information can be used for prediction disulfide connectivity.



Performance based on different windowsize

Only consider the influences of residues nearby disulfide cysteine and locations of disulfide cysteines in protein sequence to disulfide connectivity prediction; therefore, reserve the residues with windowsize *w* around disulfide cysteines and replace other resides as X. From figure 9 the performances in Q_p [Fig. 9(A)] and Q_c [Fig. 9(B)] are constantly upward according to the increase of windowsize until *w* = 13, and the performances are 0.84 and 0.84 in $B = 2$ for Q_p and Q_c, 0.84 and 0.85 in $B = 3$, 0.91 and 0.93 in $B = 4$, 0.86 and 0.88 in $B = 5$, average performance are 0.85 and 0.87. However, when $w = 25$ Q_p and Q_c are 0.92 and

0.92 in B = 2, 0.88 and 0.88 in B = 3, 0.94 and 0.95 in B = 4, 0.88 and 0.90 in B = 5, finally average performances are 0.90 and 0.91 for Q_p and Q_c respectively. Therefore, the average performances can reach 0.91 and 0.92 for Q_p and Q_c when $w \geq 25$.

Table 3 summarizes the performances based on different windowsize in CYS090506, and the performances in CYS090506 are better than CYS090506(25) in $B = 2...5$; therefore, the overall performances of CYS090506 are higher than CYS090506(25), the improvement are about 5% and 4% in Q_p and Q_c, respectively. It demonstrates that residues nearby disulfide cysteines contain sufficient information and play an important role in disulfide bonds forming; meanwhile, other residues carry less related information. Therefore, the formation of disulfide connectivity is mostly determined on the local residues around disulfide cysteines, the suitable windowsize is 25.

CONCLUSION

The better performance of PSI-BLAST method is established by abundant information of searching dataset; therefore, searching dataset CYS090506 is used for

this method. Meanwhile, the accuracy of PSI-BLAST method using inter-B searching dataset is better than using intra-B searching dataset; it reveals that there might exist sub patterns of specific disulfide connectivity. Furthermore, those proteins with sub patterns of specific disulfide connectivity might be evolutionary related, and these sub patterns might play an important role in protein structures or protein functions during evolution. This research also indicates sequence identity is higher between query protein and homologs, and then the performance of disulfide connectivity prediction is better; however, the lowest threshold of sequence identity is 30%. Moreover, local residues with windowsize 25 around disulfide cysteines carry significant messages to direct the formation of disulfide connectivity, and the overall performances are 0.90 and 0.91 for Q_p and Q_c, respectively. This is consistent with previous studies that local information brings sufficient knowledge to predict disulfide bonds. However, whole residues are utilized in the prediction, the average performances reach 0.95 and 0.95 for for Q_p and Q_c, respectively. It describes that local resides and non-local residues possess different information, and there information are complementary. The limitation of this prediction is that when a novel disulfide protein appears, the homologs of the query protein may not be found; therefore, machine learning method will be applied, this is in chapter 3.

Chapter 3

Prediction of disulfide connectivity from protein sequences

INTRODUCTION

Recently, computational biology has made significant progress in the prediction of the bonding states from protein sequences.[46-49] A number of approaches based on neural networks[47,49], statistical analysis[48] or support vector machines[46] have been shown to be quite effective in predicting the bonding state of cysteine (around 81-90% prediction accuracy). However, predicting disulfide connectivity from protein sequences remains a challenging problem in computational biology. This is because the disulfide bridges are non-local in nature (i.e., though the two cysteines that form the disulfide bridge are close in 3D space, they may be far apart from each other in the sequence). Hence, the prediction of disulfide connectivity requires extracting information about spatial proximity of cysteine pairs from one-dimensional protein sequences. The problem is further complicated by the rapid increase of possible disulfide patterns as the number of disulfide bridges increases. For example, when the number of disulfide bridges is 2, there are 3 possible disulfide patterns; but when the number of disulfide bridges increases to 5, the possible number of disulfide patterns rapidly increases to 945.

In general, disulfide-predicting approaches can be classified as two kinds, pattern-wise method and pair-wise method; furthermore, pattern-wise method utilizes

disulfide connectivity as unit and pair-wise method is in terms of cystein pair as a base to predict disulfide connectivity. In fact, pair-wise method emphasize local environment of two disulfide cysteines, use local sequence information as an input. To the best of my knowledge, the first attempt to predict the locations of disulfide bridges directly from protein sequences based on pair-wise method was done by Fariselli and Casadio.[14] They reduced disulfide connectivity to the graph matching (GM) problem in which the graph vertices are equivalent to the residues of cysteine-forming disulfide bridges, and the weight edges contact potentials. Then, the Monte-Carlo (MC) simulated annealing method is used to optimize the weights and the disulfide bridges are then identified by finding the maximal weight perfect matching. This method will be referred as MCGM. Further, Fariselli *et al.*[15] improved their results by using the NN to predict the cysteine pair wise interactions. This method will be referred to as NNGM. Next, Ferre and Clote[18] capture secondary structure information and diresidue frequencies based on neural network. Furthermore, Tsai *et al.*[20] apply sequential distance between oxidized cysteines using SVM to generate bonding potentials of cysteine pairs.

In the researches of pattern-wise method, Vullo and Frasconi[16] use an *ad hoc* recursive neural network (RNN) to predict disulfide connectivity. Later, Zhao *et al.*[17] compare cysteine separation profiles from testing and template dataset to solve this

problem. Next, Song *et al.*[43] use multiple sequence feature vectors such as cysteine-cysteine coupling pair; amino acid compositions etc. and secondary structure rely on support vector regression (SVR) to predict disulfide connectivity. On the other hand, Chen *et al.*[21] develop a two-level hierarchical framework combining pair-wise and pattern-wise method.

In general, these approaches predict 29-74% of the disulfide patterns for a dataset sharing less than 30% sequence identity, after a 4-fold cross validation procedure. In this research[19], I use SVMs based on feature vectors such as the coupling between the local sequence environments of cysteine pairs, the cysteine separations, and the amino acid content. Later, this method is modified[22] as disulfide pairs to present disulfide connectivity and genetic algorithm are used for optimizing the parameters.

MATERIALS AND METHODS

The support vector machine

Support Vector Machine (SVM)[50] has found many applications[46,51-53] in computational biology and has been shown to be a quite effective machine-learning method. Its basic idea is to map data into a high dimensional space and find a

separating hyperplane with the maximal margin between two kinds of data. Since this method is quite well known, we give only a brief description of the basic theory behind the SVM. The SVM is basically a binary classifier. Given training vectors $x_i,\ i=1,...,l$ and a vector y defined as: $y_i=1$ if x_i is in class I, and $y_i=-1$ if x_i is in the class II. The support vector technique tries to find the separating hyperplane $w^T x_i+b=0$ with the largest distance between two classes, measured along a line perpendicular to this hyperplane, which is equivalent to solving the following problems:

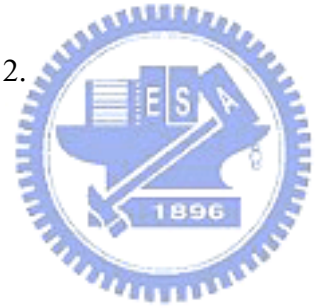
$$\min_{w,b,\xi}\frac{1}{2}w^T w+C(\sum_{i=1}^{l}\xi_i) \text{ and } y_i\left[\left(w^T\phi(x_i)\right)+b\right]\geq 1-\xi_i \qquad (7)$$

Constraints $y_i\left[\left(w^T\phi(x_i)\right)+b\right]\geq 1-\xi_i$ allow that training data may not be on the correct side of the separating hyperplane $w^T x+b=0$. C is the penalty parameter to be optimized. In practice, the explicit form of $\phi(x)$ is not required, and we only need to calculate the kernel function given by $K\left(x_i,x_j\right)\equiv\phi\left(x_i\right)^T\phi\left(x_j\right)$. We use the Radial Basis Function (RBF) kernel given by $e^{-\gamma\left\|x_i-x_j\right\|^2}$ for all the computations, where γ is the kernel parameter. All the SVM calculations are performed using LIBSVM.[54] For SVM training, a few parameters such as the penalty parameter C and the kernel parameter γ of the RBF function must be determined in advance. Choosing optimal parameters for support vector machines is an important step in SVM design. In this

work, we use the cross validation on different parameters for the model selection.[55]

Data sets

SP39-ID30 is used for comparing this method with previous works[14,16], and is classified as four sets to perform 4-fold cross validation procedures. Dataset, SP39-ID30, is extracted from the SWISS-PROT database release No. 39[45], and there are total 482 proteins with with 2 to 5 disulfide bridges ($B = 2,\ldots,5$); furthermore, the sequence homologous within proteins are less than 30%. The details of SP39-ID30, please see methods in chapter 2.



The feature vectors

The selection of relevant features in large and complex biological data sets significantly affects the effectiveness of the SVM method. We select three types of feature vectors: the coupling between the local sequence environments of cysteine pairs, the cysteine sequence separations, and the amino acid content.

The cysteine-cysteine coupling

A sequence window of size $2l+1$ amino acids centered on the cysteine is used to describe the neighboring sequence environment of the cysteine. Evolution

information of the protein sequence is included in the window by using the sequence profile generated by PSI-BLAST,[56] i.e., the position specific substitution matrix (PSSM). The use of the PSSM has the advantage of avoiding the time-consuming multiple-sequence alignment procedures. The PSSM of a protein sequence is a $L \times 20$ matrix, where L is the sequence length and 20 is the number of amino acid types (the amino acid type is numbered from 1 to 20). The matrix element p_{ij} of the PSSM represents the log-odds score of the i^{th} amino acid of type j. Each 20-element row vector of the PSSM represents the distribution of the occurrences of 20 amino acid types at the specific position. Let $\mathbf{w}_i = (a_{i-l},\ldots,a_{i-1},a_i,a_{i+1},\ldots,a_{i+l})$ denote the sequence window of size $2l+1$ centered around the bonded cysteine at the i^{th} position, where a_k is the k^{th} amino acid. A 20-element vector $\mathbf{v}_{w_i} = (v_1^{w_i}, v_2^{w_i}, \ldots v_{20}^{w_i})$ associated with the sequence window $\mathbf{w}_i$ is defined, where $v_k^{w_i}$ is the PSSM element of the amino acid type k. If the amino acid of a given type occurs more than once within the window, $v_k^{w_i}$ is the sum of the associated PSSM elements. The coupling between the i^{th} and j^{th} cysteines is computed by $\mathbf{s}_{ij} = c_i'\mathbf{v}_{w_j} + c_j'\mathbf{v}_{w_i}$, where c_k' is the PSSM element of cysteine type at the k^{th} row. For a given disulfide pattern, sum up all the possible cysteine pairs to get $\mathbf{s} = \sum_{ij} \mathbf{s}_{ij}$. The symbol S is used to denote the cysteine-cysteine coupling of disulfide patterns. After preliminary

experiment, set the window size to be 21 for $B=3$ and 5, 7 for $B=2$, and 27 for $B=4$.

Cysteine-spacing patterns

For a disulfide protein with n cysteines, i.e., $c_1,c_2,\ldots,c_n$, its disulfide pattern is denoted by $(c_ic_j,c_{i'}c_{j'},\ldots)$, where c_ic_j designates a disulfide bridge formed between cysteine i and j. For a given disulfide pattern $(c_ic_j,c_{i'}c_{j'},\ldots)$, there is an associated cysteine spacing pattern given by $(d_{ji},d_{j'i'},\ldots)$, where d_{ji} is the sequence spacing c_i and c_j. An example is given in figure 12. For a protein with four cysteines $c_1c_2c_3c_4$, which form two disulfide bonds, there will be three possible disulfide configurations: $C_1=(c_1c_2,c_3c_4)$, $C_2=(c_1c_3,c_2c_4)$ and $C_3=(c_1c_4,c_2c_3)$. The three corresponding cysteine spacing patterns are given by $D_1=(d_{12},d_{34})$, $D_2=(d_{13},d_{24})$ and $D_3=(d_{14},d_{23})$. The symbol D is used to denote the cysteine separation vector.

Amino acid content

Amino acid composition has been shown to be a useful global sequence descriptor in fold recognition,[52] and in the prediction of the bonding states of cysteines[46] and protein subcellular localization.[53] Amino acid composition is represented by the composition vector $A=(a_1,a_2,\ldots,a_{20})$, where $a_k=n_k/n_0$. Here 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. The notation A is used to denote the encoding of the amino acid composition.

Performance Indices

Q_p and Q_c are used to evaluate the performance of disulfide connectivity, and the definitions can be seen in Methods of Chapter 2.

RESULTS AND DISCUSSIONS

Table 6 summarizes the performances of SVMs based on various encodings. The results computed from the random predictor is also listed, and referred to as R, as the reference of the base performance. The Q_p and Q_c of the random predictor are given by $1/(2B-1)!!$ and $1/(2B-1)$, respectively.[14] In general, the pattern-based Q_p is lower than the disulfide bridge-based Q_c, since the former counts only those proteins whose complete disulfide patterns are correctly predicted. In the case of $B=2$, both D and S classifiers perform similarly (67%). However, it is interesting to note that the much simpler A classifier, which uses only global sequence information of amino acid composition, gives fairly good results (61%). In the case of $B=3$, the

differences in the predictive performance among the classifiers start to show themselves. The D classifier performs significantly better, and, in terms of the more stringent Q_p, it is 16% and 7% higher than A and S, respectively. Note that the D encoding does not contain any information about the explicit amino acid sequence other than the cysteine separations. This is consistent with previous works[29,32] indicating that disulfide patterns and cysteine separations are closely related to each other and that disulfide patterns can be effectively used to detect remote homologues undetectable by the sequence alignment methods. In the case of $B = 4$ and 5, the prediction accuracies of the SVMs, though significantly better than those of the random predictor, are not yet practical at present. The poor results for these cases are due to the relatively smaller number of the reliably annotated proteins with higher number of disulfide bridges in the dataset (see the Datasets Section). However, the situation is expected to improve when more known structures are available in the future. On the other hand, when comparing the results of the D classifiers with those of the random predictor R, there is a phenomenon that, the ratios of Q_p between D and R are 28 and 120 for $B = 4$ and 5, respectively, indicating that the SVM is still effective in these cases.

Using multiple feature vectors can improve on the performance of the SVM classifiers based on a single feature vector type has previously shown in many

biological applications [46,52,53]. I selected the following linear combinations: $D + w_A A$, $D + w_S S$ and $D + w_A A + w_S S$, where w_d is the weight associated with the d encoding. After preliminary experiment, we set the weights to be $w_A = 1$ and $w_S = 0.001$. For the sake of simplicity, I use the simpler notations $D + A$, $D + S$ and $D + A + S$, with the understanding that w_A and w_S are omitted from the notations. Table 7 compares the performances of the SVMs based on the multiple feature vectors. As expected, the SVMs based on the multiple feature vectors in general perform better than those based on a single feature vector type.

Figure 13 shows some typical examples of the predictions by the $D + A + S$ classifier. Figure 13(A) shows the case of B = 3, 1tpa:I,[57] which is a bovine pancreatic trypsin inhibitor, Figure 13(B) the case of B = 4, 1afh,[58] a nonspecific lipid transfer protein, and Figure 13(C), the case of B = 5, 1pcn,[59] a porcine pancreatic procolipase. In these cases, the disulfide bridges are all perfectly predicted. The number of incorrectly predicted disulfide bridges, if any, will be either greater than or equal to 2, since one incorrectly predicted disulfide bridge will necessarily give rise to another one. An example is given in Figure 13(D). The observed and the predicted disulfide patterns of 1phm[60] (peptidylglycine -hydroxylating monooxygenase) are [1-6,2-4,3-5,7-10,8-9] and [1-6,2-4,*3-8*,7-10,*5-9*], respectively (the incorrect predictions are in italics). Hence, in the case of B = 2, the cysteine pair-based measure

Q_c of a protein is either 0 or 1, while in the case of $B = 3$, Q_c is 1, 1/3 or 0.

Table 8 compares the results of the $D+A+S$ with those of other methods. Using 50% of accuracy of Q_p as threshold, the overall prediction accuracy of $D+S+A$ is above 50% ($Q_p = 0.55$ and $Q_c = 0.57$), which is higher than those from methods such as MCGM, NNGM, RNN, DiANNA, and CSP, which give 0.29–0.49 in Q_p and 0.38–0.52 in Q_c. Besides, these methods are all pair-wised method, except $D+A+S$. MCGM applies contact potential to predict bonding state of disulfide pair, NNGM uses neural network, RNN uses evolutionary information based on recursive neural networks, DiANNA employs secondary information and diresidue frequencies with neural networks and CSP calculate the smallest divergence value between cysteine separation profiles of query protein and templates. The feature vectors of these methods focus on local environment of disulfide pairs; or protein length. On the other hand, performances of some methods are better than this work such as pairSVM, 2-level SVM, SVM_GA, SVR; furthermore, pairSVM and the first level of 2-level SVM utilize local amino acids and distance between disulfide pair as feature vectors, second level of 2-level SVM uses output of first level, cysteine separation profile, and protein length as input of SVM. Their performances are 0.63 and 0.70 in Q_p for pairSVM and 2-level SVM, respectively. However, the performances of SVM_GA are based on my feature vectors, SVR also utilizes my

feature vectors and others, and their average performances are all 0.74 in Q_p. The feature vector cysteine-cysteine coupling possess the local environment information of disulfide connectivity, and evolutionary interaction between two disulfide cysteines; amino acid contents show the global information of whole protein sequence; furthermore, cysteine spacing patterns represent the relative distances between disulfide cysteines. It demonstrates that these feature vectors are important and useful in prediction of disulfide connectivity.

CONCLUSION

Though the SVM is known to be a powerful machine learning method, due to the complexity of biological data, the identification and selection of relevant biological features becomes an important issue in the applications of SVMs to biological problems. In this work, I tested SVMs in the prediction of disulfide connectivity using biological features characteristic of disulfide bridges. My results indicate that both cysteine-cysteine sequence couplings and cysteine separations are important features in predicting disulfide connectivity. This is consistent with the previous studies[29,32] indicating that a close relationship exists between cysteine separations and disulfide patterns, and that such a relationship can be utilized to identify the remote homologs

undetectable by sequence alignments. I showed that the SVM based on the cysteine separations give the best predictive performance among the SVMs based on the single feature vector. I also showed that the SVMs based on the multiple feature vectors out-performs those based on the single feature vector.

CHAPTER 4

S-S Predictor: a disulfide connectivity prediction server

INTRODUCTION

Cysteines stand an important role in protein sequence; not only two oxidized cysteines can form a disulfide bond but also some cysteines modulate the protein functions.[3,23,25,26,28,61,62] Therefore, prediction the bonding state of cysteines can help in determination protein structure and infer the influence of cysteine in protein function. As we known, disulfide bond is a distant bond; it connects two distant cysteines; therefore, constraints the searching space of protein sequence from denatured state to native state. On the other hand, ligand-bound cysteines are involved in protein function, for example, inner mitochondrial membrane protein sco 1p that contains a CxxxC motif[63,64], and this motif is involved in copper transport. In fact, a number of computational approaches[46-49,65-67] are developed to predict the bonding states of cysteines. Chen *et al.*[46] develop a method to predict the bonding states of cysteines using SVM based on multiple feature vectors and the cysteine state sequences; consequently, the performance is 90% in overall prediction accuracy and 0.77 Matthews correlation coefficient.

Hence, according to the previous researches, disulfide connectivity can be predicted in terms of the knowledge of bonding state cysteines. Disulfide bonds play an important role in stabilizing protein structure and regulating protein function. Therefore, the ability to infer disulfide connectivity directly from protein sequence is

valuable in both structural modeling and functional analysis. The previous study[29] showed that disulfide proteins with the same disulfide connectivity are usually having similar folds, even if these proteins have very low sequence identities. The disulfide connectivity prediction has been investigated by a variety of computational methods[14,16-22,42,43] with the prior knowledge of bonding states of cysteines. However, Lu *et al*. use the cysteine-cysteine coupling, cysteine spacing patterns, and amino acid content feature vectors[19] based on SVM and adjust parameters by GA , the average performance is 74% in Q_p and 79% in Q_c .

In general, there are two stages in predicting disulfide connectivity, first predict the bonding states of cysteines, and then infer the disulfide connectivity from oxidized cysteines. However, during the two-part predictions, the overall performance will decrease. In my research, homologous search is used to solve this problem; it can directly predict disulfide connectivity by sequence-to-structure mapping without knowing oxidized cysteines in advance, but not all query proteins can find homologs. Therefore, I develop an approach to predict disulfide connectivity from protein sequences, which is referred as S-S Predictor (http://140.113.239.214/~ssbond). The S-S Predictor is a hybrid method based on both sequence alignment and machine learning method. It first performs PSI-BLAST search to identify the sequence homologs that have known disulfide connectivity. The S-S Predictor then predicts the

disulfide pattern of the query sequence based on the similarity of the cysteine separation vectors. If no homologs of known disulfide patterns are found, the S-S Predictor switches to the support vector classifier to predict the disulfide patterns.

The S-S Predictor integrates both sequence alignment and the machine learning method to predict disulfide connectivity. This method will be useful to biologists interested in the study of disulfide proteins. The S-S Predictor server can be accessed from http://140.113.239.214/~ssbond.

MATERIALS AND METHODS

Dataset

There are two dataset used in this research; one is SP39-ID30, which contains 482 disulfide proteins, and is divided as four sets for 4-fold cross validation; the other searching dataset for PSI-BLAST is CYS090506, which contains 15,252 disulfide proteins with known structures. The details of these dataset are described in the Methods in Chapter 2.

Disulfide connectivity prediction

According to the PSI-BLAST consensus, the hybrid method was applied to

predict disulfide connectivity. (Please see the Methods in Chapter 2.)

Support vector machine

(Please see the Methods in Chapter 3.)

Feature vectors

Cysteine-cysteine coupling, cysteine spacing patterns, and amino acid content feature vectors are used in this research (Please see the Methods in Chapter 3.), and SVM_GA[22] is adopted as the machine learning method. Therefore, disulfide connectivity is presented as cysteine pairs in machine learning part of this predictor, and there are two kinds of cysteine pairs CP_1 and CP_2 . CP_1 uses one cysteine pairs, and then CP_2 uses two cysteine pairs as a unit to display disulfide connectivity. For example, the native disulfide connectivity is $[C_1C_3, C_2C_4]$; however, CP_1 method shows this disulfide connectivity as $[C_1 \otimes C_2, C_1 \oplus C_3, C_1 \otimes C_4, C_2 \otimes C_3, C_2 \oplus C_4, C_3 \otimes C_4,]$, and CP_2 method shows this disulfide connectivity as $[C_1 \otimes C_2 - C_3 \otimes C_4, C_1 \oplus C_3 - C_2 \oplus C_4, C_1 \otimes C_4 - C_2 \otimes C_3]$, where $\oplus$ denotes bonding state between two cysteines, and $\otimes$ denotes non-bonding state between two cysteines. Furthermore, the bonding states of disulfide pairs are predicted by SVM. Finally, cysteines are identified as vertices and disulfide bonds are presumed as edges; therefore, finding correct disulfide connectivity is

treated as computing the maximum-weight perfect matching [68].

Feature selection

Genetic algorithm (GA)[22] is used to optimize feature selection such as an *m*-dimensional vector, parameter *C* and the kernel parameter γ of the SVM. There are three steps in GA: selection operator, mutation operator, and crossover operator; meanwhile, the prediction accuracy of disulfide connectivity is defined as fitness function. N solutions are produced in initial population, and denotes as 0^{th}; half of N, $n_{1,...,N/2}$, is indicated as father population, and the others of N, $n_{N/2,...N}$, is indicated as mother population. In the step of selection operator, it determines the best solutions of father and mother in χ^{th} population based on fitness function. Furthermore, there are two types of mutations in mutation operator, first every bit in vectors is mutated in $n_{1,...,N/2}$, if the mutation rate is less than a mutation threshold $\mu_0 = 0.1$; second randomly choose a bit from each feature vector to be mutated in $n_{N/2,...N}$. Finally, crossover operators are performed between n_{2p-1} and n_{2p}, where $p = 1,...,N/2$, and if the crossover rate is less than crossover threshold $\mu_1 = 0.5$, one-point crossovers are implemented.

S-S Predictor process

The flowchart of S-S predictor is described in figure 2. First, the query protein searches for homologous disulfide proteins in dataset CYS090506 by PSI-BLAST. If whole bonding cysteines in a query protein are totally aligned with some homologous disulfide protein, then calculate the frequencies of disulfide connectivity appeared in the matched homologs and the number of all possible disulfide connectivity is $(2B-1)!!$ where B is the number of disulfide bond. Finally, assign maximal number of disulfide connectivity as a result. In fact, not all of bonding cysteines in query protein can be totally fit with cysteines in homologs; therefore, treat disulfide pair as a unit, count the number of various disulfide pairs, $B(2B-1)$, in partial matched homologs, and assign disulfide connectivity with maximal number of disulfide pairs as the result. When homologs of the query protein can not be found by PSI-BLAST, feature vectors of a query protein are prepared and feed into SVM with GA optimization to predict disulfide connectivity. In summary, according to the aligned level of bonding cysteines between query protein and homologs, if total bonding cysteines are aligned with homologs, the PSI-BLAST method in terms of disulfide connectivity as basis to predict disulfide connectivity, or in terms of disulfide pair as basis. Consequently, when no homologs can be found with query protein, the SVM_GA is used to predict disulfide connectivity.

RESULTS AND DISCUSSION

Interface

The website of the S-S predictor is http://140.113.239.214/~ssbond, and the input interface is shown in figure 14. The main options in the interface are as follow.

Predict options The default value of the options is "let me guess the positions of oxidized cysteines"; S-S predictor can predict the bonding states and disulfide connectivity at the same time by PSI-BLAST method. If the oxidized cysteines are understood in advance, then users can check the option of "input the positions of oxidized cysteines", and input the positions of oxidized cysteines in sequence at the same time; however, the positions of oxidized cysteines should be separated by a blank space.

Query sequence A query protein sequence is presented as standard amino acid one-letter codes, and FASTA format is accepted in this predictor; therefore, users can write annotation of the query protein after the sign of ">". Meanwhile, spaces and newline will be automatically stripped.

Upload file Offer another choice for users, not only can copy paste a query protein in the interface, but also can upload query protein sequence from text file as FASTA format.

When users choose proper options from input interface, push "Submit" button to

send the query information. Consequently the output interface is shown as figure 15, the main contents includes:

Prediction method It can tell users the disulfide connectivity is predicted by blast method or SVM method; furthermore, users can click Blast method to see the result of alignment.

Sequence label The FASTA format is accepted in input interface; therefore, what users write after ">" will be recorded in this part for labeling the query sequence.

Sequence length It records the number of amino acids in a query protein.

Number of bonded cysteines It records the number of cysteines which are related with disulfide bonds in a query protein.

Number of free cysteines It records the number of cysteines which do not participate the forming of disulfide bonds in a query protein.

Number of disulfide bonds It records the number of disulfide bonds in a query protein.

Predicted disulfide connectivity It presents the disulfide connectivity of a query protein; meanwhile, disulfide bond between two bonded cysteines is connected by "-" and every disulfide bond is separated by a blank space.

Output sequence It shows the query protein sequence and the format is one row contains fifty amino acids and every ten amino acids are separated as a blank

space. Furthermore, the positions of disulfide cysteines are labeled and every pair of disulfide bonds is linked by a red line. Therefore, it is clear and convenient for users to identify the locations of disulfide bonds and positions of disulfide cysteines among the query sequence.

Performance

Compare the performances of PSI-BLAST method, SVM_GA, and S-S predictor [Table 9] according to dataset SP39-ID30 with 4-fold cross validation, and the overall accuracies of SVM_GA are 27% and 30% better than PSI-BLAST in Q_p and Q_c; however, this is because the advantages of PSI-BLAST method are based on sequence homologous, the similarity of every two sequences within SP39-ID30 is equal or lower than 30%. Otherwise, SVM_GA can capture sufficient information to predict disulfide connectivity no matter what the sequence similarities are between query protein and training dataset. Consequently, the performances of S-S predictor, a hybrid method of PSI-BLAST and SVM_GA, are 0.90 and 0.90 in B=2, 0.80 and 0.84 in B=3, 0.75 and 0.84 in B=4, 0.60 and 0.76 in B=5, and finally the overall accuracy are 0.81 and 0.84 for Q_p and Q_c, respectively. Therefore, the information carried by PSI-BLAST method and SVM_GA for disulfide connectivity predictions is complementary.

Table 10 compares the results of S-S predictor with other methods. The S-S predictor is the only one method that gives the overall prediction accuracy above 80% (Q_p= 0.81, and Q_c= 0.84), while the other methods give 0.29-0.74 in Q_p and 0.38-0.78 in Q_c. Therefore, S-S predictor is useful and helpful for disulfide connectivity prediction.

CONCLUSION

The S-S predictor is a hybrid system to predict disulfide connectivity; it first performs PSI-BLAST method; however, if homologs can not be found, and then implements machine learning method. S-S predictor also provides a website for large-scale disulfide connectivity prediction; furthermore, the input interface of S-S predictor is clear and easy for user to paste or upload protein sequence with FASTA format, and to choose options to guess the positions of oxidized cysteines or directly input the positions of oxidized cysteines. On the other hand, the output interface not only offers the disulfide connectivity, but also presents the protein sequence with positions of disulfide cysteines and indicating the disulfide bonds. This information is convenient for users to obtain the locations of disulfide connectivity in protein sequence, and understand the local environment of disulfide cysteines. The of S-S predictor is an outstanding method to predict disulfide and its performances are 0.81

and 0.84 in Q_p and Q_c, respectively.

References

1. Hogg PJ. Disulfide bonds as switches for protein function. Trends Biochem Sci 2003;28:210-214.
2. Gilbert HF. Protein chaperones and protein folding. Curr Opin Biotechnol 1994;5:534-539.
3. Wedemeyer WJ, Welker E, Narayan M, Scheraga HA. Disulfide bonds and protein folding. Biochemistry 2000;39:7032.
4. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223-230.
5. Dombkowski AA. Disulfide by Design: a computational method for the rational design of disulfide bonds in proteins. Bioinformatics 2003;19:1852-1853.
6. Creighton TE. Disulphide bonds and protein stability. Bioessays 1988;8:57-63.
7. Tidor B, Karplus M. The contribution of cross-links to protein stability: a normal mode analysis of the configurational entropy of the native state. Proteins 1993;15:71-79.
8. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267-281.
9. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217-241.
10. Anderson JL, Chapman SK. Ligand probes for heme proteins. Dalton Trans 2005:13-24.
11. Yano H, Kuroda S, Buchanan BB. Disulfide proteome in the analysis of protein function and structure. Proteomics 2002;2:1090-1096.
12. Han YH, Wang Q, Jiang H, Liu L, Xiao C, Yuan DD, Shao XX, Dai QY, Cheng JS, Chi CW. Characterization of novel M-superfamily conotoxins with new disulfide linkage. Febs J 2006;273:4972-4982.
13. Fadel V, Bettendorff P, Herrmann T, de Azevedo WF, Jr., Oliveira EB, Yamane T, Wuthrich K. Automated NMR structure determination and disulfide bond identification of the myotoxin crotamine from Crotalus durissus terrificus. Toxicon 2005;46:759-767.
14. Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. Bioinformatics 2001;17:957-964.
15. Fariselli P, Martelli PL, Casadio R. A neural network based method for predicting the disulfide connectivity in proteins. Knowledge Based Intelligent

Information Engineering Systems and Allied Technologies (KES). Amsterdam: IOS Press 2002:464-468.
16. Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 2004;20:653-659.
17. Zhao E, Liu HL, Tsai CH, Tsai HK, Chan CH, Kao CY. Cysteine separations profiles on protein sequences infer disulfide connectivity. Bioinformatics 2005;21:1415-1420.
18. Ferre F, Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 2005;21:2336-2346.
19. Chen YC, Hwang JK. Prediction of disulfide connectivity from protein sequences. Proteins 2005;61:507-512.
20. Tsai CH, Chen BJ, Chan CH, Liu HL, Kao CY. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. Bioinformatics 2005;21:4416-4419.
21. Chen BJ, Tsai CH, Chan CH, Kao CY. Disulfide connectivity prediction with 70% accuracy using two-level models. Proteins 2006;64:246-252.
22. Lu CH, Chen YC, Yu CS, Hwang JK. Predicting disulfide connectivity patterns. Proteins 2007;67:262-270.
23. Abkevich VI, Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioninformatics analysis. J Mol Biol 2000;300:975-985.
24. Anfinsen CB, Scheraga HA. Experimental and theoretical aspects of protein folding. Adv Protein Chem 1975;29:205-300.
25. Clarke J, Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. Biochemistry 1993;32:4322-4329.
26. Clarke J, Hounslow AM, Bond CJ, Fersht AR, Daggett V. The effects of disulfide bonds on the denatured state of barnase. Protein Sci 2000;9:2394-2404.
27. Harrison PM, Sternberg MJ. Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. J Mol Biol 1994;244:448-463.
28. Yokota A, Izutani K, Takai M, Kubo Y, Noda Y, Koumoto Y, Tachibana H, Segawa S. The transition state in the folding-unfolding reaction of four species of three-disulfide variant of hen lysozyme: the role of each disulfide bridge. J Mol Biol 2000;295:1275-1288.
29. Chuang CC, Chen CY, Yang JM, Lyu PC, Hwang JK. Relationship between protein structures and disulfide-bonding patterns. Proteins 2003;53:1-5.

30. Harrison PM, Sternberg MJ. The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. J Mol Biol 1996;264:603-623.
31. Mas JM, Aloy P, Marti-Renom MA, Oliva B, Blanco-Aparicio C, Molina MA, de Llorens R, Querol E, Aviles FX. Protein similarities beyond disulphide bridge topology. J Mol Biol 1998;284:541-548.
32. van Vlijmen HW, Gupta A, Narasimhan LS, Singh J. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. J Mol Biol 2004;335:1083-1092.
33. Dani VS, Ramakrishnan C, Varadarajan R. MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. Protein Eng 2003;16:187-193.
34. Hazes B, Dijkstra BW. Model building of disulfide bonds in proteins with known three-dimensional structure. Protein Eng 1998;2:119-125.
35. Kong L, Lee BT, Tong JC, Tan TW, Ranganathan S. SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. Nucleic Acids Res 2004;32:W356-359.
36. Mas JM, Aloy P, Marti-Renom MA, Oliva B, de Llorens R, Aviles FX, Querol E. Classification of protein disulphide-bridge topologies. J Comput Aided Mol Des 2001;15:477-487.
37. Pabo CO, Suchanek EG. Computer-aided model-building strategies for protein design. Biochemistry 1986;25:5987-5991.
38. Sowdhamini R, Srinivasan N, Shoichet B, Santi DV, Ramakrishnan C, Balaram P. Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. Protein Eng 1989;3:95-103.
39. Thangudu RR, Vinayagam A, Pugalenthi G, Manonmani A, Offmann B, Sowdhamini R. Native and modeled disulfide bonds in proteins: knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. Proteins 2005;58:866-879.
40. Vinayagam A, Pugalenthi G, Rajesh R, Sowdhamini R. DSDBASE: a consortium of native and modelled disulphide bonds in proteins. Nucleic Acids Res 2004;32:D200-202.
41. Cheng J, Saigo H, Baldi P. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins 2006;62:617-629.
42. Ferre F, Clote P. DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res 2005;33:W230-232.

43. Song J, Yuan Z, Tan H, Huber T, Burrage K. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. Bioinformatics 2007;23:3147-3154.
44. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.
45. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45-48.
46. Chen YC, Lin YS, Lin CJ, Hwang JK. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 2004;55:1036-1042.
47. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins 1999;36:340-346.
48. Fiser A, Simon I. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 2000;16:251-256.
49. Martelli PL, Fariselli P, Malaguti L, Casadio R. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng 2002;15:951-953.
50. Cortes C, Vapnik V. Support-vector network. Machine Learning 1995;20:273-297.
51. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397-407.
52. Yu CS, Wang JY, Yang JM, Lyu PC, Lin CJ, Hwang JK. Fine-grained protein fold assignment by support vector machines using generalized npeptide coding schemes and jury voting from multiple-parameter sets. Proteins 2003;50:531-536.
53. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 2004;13:1402-1406.
54. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Software availavle from http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2001.
55. Duan K, Keerthi SS, Poo AN. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurcomputing 2003;51:41-59.
56. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-3402.

57. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr B 1983;39:480-490.
58. Gomar J, Petit MC, Sodano P, Sy D, Marion D, Kader JC, Vovelle F, Ptak M. Solution structure and lipid binding of a nonspecific lipid transfer protein extracted from maize seeds. Protein Sci 1996;5:565-577.
59. Breg JN, Sarda L, Cozzone PJ, Rugani N, Boelens R, Kaptein R. Solution structure of porcine pancreatic procolipase as determined from 1H homonuclear two-dimensional and three-dimensional NMR. Eur J Biochem 1995;227:663-672.
60. Prigge ST, Kolhekar AS, Eipper BA, Mains RE, Amzel LM. Amidation of bioactive peptides: the structure of peptidylglycine alpha-hydroxylating monooxygenase. Science 1997;278:1300-1305.
61. Akamatsu Y, Ohno T, Hirota K, Kagoshima H, Yodoi J, Shigesada K. Redox regulation of the DNA binding activity in transcription factor PEBP2. The roles of two conserved cysteine residues. J Biol Chem 1997;272:14497-14500.
62. Gladyshev VN. Thioredoxin and peptide methionine sulfoxide reductase: convergence of similar structure and function in distinct structural folds. Proteins 2002;46:149-152.
63. Balatri E, Banci L, Bertini I, Cantini F, Ciofi-Baffoni S. Solution structure of Sco1: a thioredoxin-like protein Involved in cytochrome c oxidase assembly. Structure 2003;11:1431-1443.
64. Chinenov YV. Cytochrome c oxidase assembly factors with a thioredoxin fold are conserved among prokaryotes and eukaryotes. J Mol Med 2000;78:239-242.
65. Fiser A, Cserzo M, Tudos E, Simon I. Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. FEBS Lett 1992;302:117-120.
66. Mucchielli-Giorgi MH, Hazout S, Tuffery P. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins 2002;46:243-249.
67. Muskal SM, Holbrook SR, Kim SH. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng 1990;3:667-672.
68. Papadimitrou CH S. Combinatorial Optimization: Algorithms and Complexity. In: Prentice-Hall EC, NJ, editor; 1982.

Table 1. Comparison number of disulfide patterns which are observed in CYS090506 and in statistics according to number of disulfide bond.

Disulfide Bonds	N_{po} [a]	N_{pp} [b]
$B = 2$	3	3
$B = 3$	15	15
$B = 4$	51	105
$B = 5$	69	945
$B = 6$	42	10395
$B = 7$	22	135135
$B = 8$	24	2027025

[a] N_{po} = number of observed disulfide patterns in CYS090506.

[b] N_{pp} =number of possible disulfide patterns in statistics.

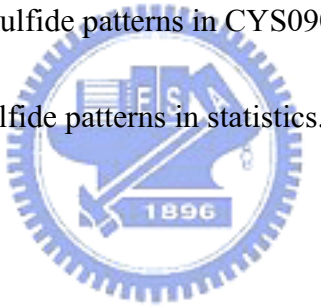


Table 2. The performance of PSI-BLAST method on different dataset.

	B[a]=2		B=3		B=4		B=5		Overall	
Dataset	Q_p[b]	Q_c[c]	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
SP39-ID30	0.52	0.52	0.49	0.51	0.42	0.49	0.29	0.42	0.47	0.49
CYS040307-NR	0.91	0.91	0.90	0.92	0.91	0.92	0.86	0.90	0.90	0.91
CYS090506	0.96	0.96	0.93	0.93	0.95	0.96	0.90	0.93	0.95	0.97

[a]B represents the number of disulfide bond in a protein.

[b]Q_p = accuracy in protein-based level.

[c]Q_c = accuracy in disulfide-bridged level.

Table 3. The performance of PSI-BLAST method according to different windowsize.

	B[a]=2		B=3		B=4		B=5		Overall	
Dataset	Q_p[b]	Q_c[c]	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
CYS090506(25)[d]	0.92	0.92	0.88	0.88	0.94	0.95	0.88	0.90	0.90	0.91
CYS090506	0.96	0.96	0.93	0.93	0.95	0.96	0.90	0.93	0.95	0.95

[a]B represents the number of disulfide bond in a protein.

[b]Q_p = accuracy in protein-based level.

[c]Q_c = accuracy in disulfide-bridged level.

[d]CYS090506(25) represents residues around disulfide cysteines with windowsize 25 are kept and others are replaced as X.

Table 4. Sub patterns of KLK_PIG.

Protein	Disulfide connectiviey			
KLK_PIG (query protein)	**26_42**	**121_190**	**155_169**	**180_205**
1hiaB		39_104	71_85	94_119
1hiaY		39_104	71_85	94_119
2kaiB		39_104	71_85	94_119
2pkaB		39_104	71_85	94_119
2pkaY		39_104	71_85	94_119
1tfxB		116_181	148_162	171_195
1zr0C	25_41	116_181	148_162	
1gvzA	26_42		149_163	
1hiaA	26_42			
1hiaX	26_42			
2kaiA	26_42			
2pkaA	26_42			
2pkaX	26_42			

Table 5. Sub patterns of HGF_HUMAN

Protein	Disulfide connectiviey				
HGF_HUMAN (query protein)	**70_96**	**74_84**	**128_206**	**149_189**	**177_201**
2hgfA	40_66	44_54			
1gmnA				112_154	140_164
1gmnB				53_93	81_105
1ki0A			4_82	25_65	53_77
1i5kA			2_79	23_62	51_74
1i5kB			3_80	24_63	52_75
1kiv_			1_78	22_61	50_73
3kiv_			2_79	23_62	51_74
1jfnA			25_102	46_85	74_87
1b2iA			1_78	22_61	50_73
1i71A			2_79	23_62	51_74

Table 6. The performance of the SVMs based on a single feature vector type

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2 \ldots 5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
R	0.33	0.33	0.06	0.20	0.01	0.14	0.001	0.11	0.14	0.20
A	0.61	0.61	0.38	0.51	0.13	0.20	0.07	0.27	0.39	0.42
S	0.67	0.67	0.47	0.60	0.17	0.24	0.12	0.32	0.45	0.48
D	0.67	0.67	0.54	0.64	0.28	0.39	0.12	0.30	0.50	0.54

Table 7. The performances of the SVMs based on multiple feature vectors

Method	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2 \ldots 5$	
	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
$D + A$	0.74	0.74	0.54	0.64	0.28	0.39	0.12	0.30	0.52	0.55
$D + S$	0.71	0.71	0.60	0.66	0.30	0.41	0.12	0.30	0.54	0.55
$D + S + A$	0.74	0.74	0.61	0.69	0.30	0.40	0.12	0.31	0.55	0.57

Table 8. Comparison of predictive performances of different approach to predict disulfide connectivity

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2…5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
MCGM[14]	0.56	0.56	0.21	0.36	0.17	0.37	0.02	0.21	0.29	0.38
NNGM[15]	0.68	0.68	0.22	0.37	0.20	0.37	0.02	0.26	0.34	0.42
RNN[16]	0.73	0.73	0.41	0.51	0.24	0.37	0.13	0.30	0.44	0.49
DiANNA[18]	0.62		0.40		0.55		0.26		0.49	
CSP[17]	0.74	0.74	0.44	0.53	0.26	0.44	0.18	0.31	0.49	0.52
This work[19]	0.74	0.74	0.61	0.69	0.30	0.40	0.12	0.31	0.55	0.57
pairSVM[20]	0.79	0.79	0.53	0.62	0.55	0.70	0.58	0.71	0.63	0.70
2-level SVM[21]	0.85		0.67		0.57		0.58		0.70	
SVM_GA[22]	0.86	0.86	0.75	0.80	0.63	0.77	0.48	0.71	0.74	0.79
SVR[43]	0.87	0.87	0.67	0.73	0.79	0.85	0.47	0.64	0.74	0.78

Table 9. The performance of PSI-BLAST method based on different method.

	B[a]=2		B=3		B=4		B=5		Overall	
Method	Q_p[b]	Q_c[c]	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
PSI-BLAST	0.52	0.52	0.49	0.51	0.42	0.49	0.29	0.42	0.47	0.49
SVM_GA	0.86	0.86	0.75	0.80	0.63	0.77	0.48	0.71	0.74	0.79
S-S predictor	0.90	0.90	0.80	0.84	0.75	0.84	0.60	0.76	0.81	0.84

[a]B represents the number of disulfide bond in a protein.

[b]Q_p = accuracy in protein-based level.

[c]Q_c = accuracy in disulfide-bridged level.

Table 10. Comparison of predictive performances of different approach to predict disulfide connectiviry

	$B = 2$		$B = 3$		$B = 4$		$B = 5$		$B = 2…5$	
Method	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c	Q_p	Q_c
MCGM[14]	0.56	0.56	0.21	0.36	0.17	0.37	0.02	0.21	0.29	0.38
NNGM[15]	0.68	0.68	0.22	0.37	0.20	0.37	0.02	0.26	0.34	0.42
RNN[16]	0.73	0.73	0.41	0.51	0.24	0.37	0.13	0.30	0.44	0.49
DiANNA[18]	0.62		0.40		0.55		0.26		0.49	
CSP[17]	0.74	0.74	0.44	0.53	0.26	0.44	0.18	0.31	0.49	0.52
This work[19]	0.74	0.74	0.61	0.69	0.30	0.40	0.12	0.31	0.55	0.57
pairSVM[20]	0.79	0.79	0.53	0.62	0.55	0.70	0.58	0.71	0.63	0.70
2-level SVM[21]	0.85		0.67		0.57		0.58		0.70	
SVM_GA[22]	0.86	0.86	0.75	0.80	0.63	0.77	0.48	0.71	0.74	0.79
SVR[43]	0.87	0.87	0.67	0.73	0.79	0.85	0.47	0.64	0.74	0.78
S-S predictor	0.90	0.90	0.80	0.84	0.75	0.84	0.60	0.76	0.81	0.84

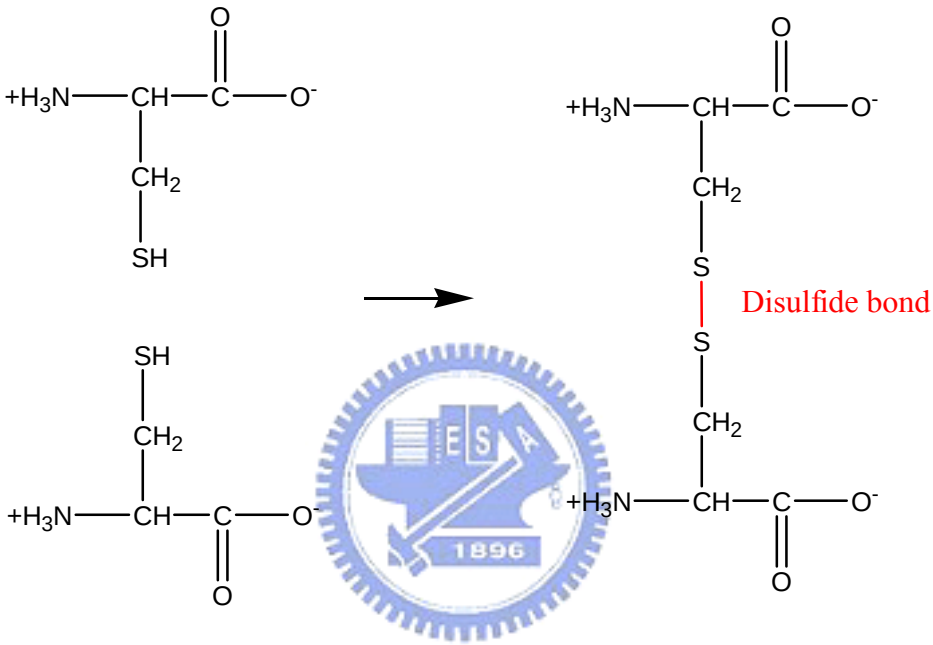


Figure 1. The formation of a disulfide bond.

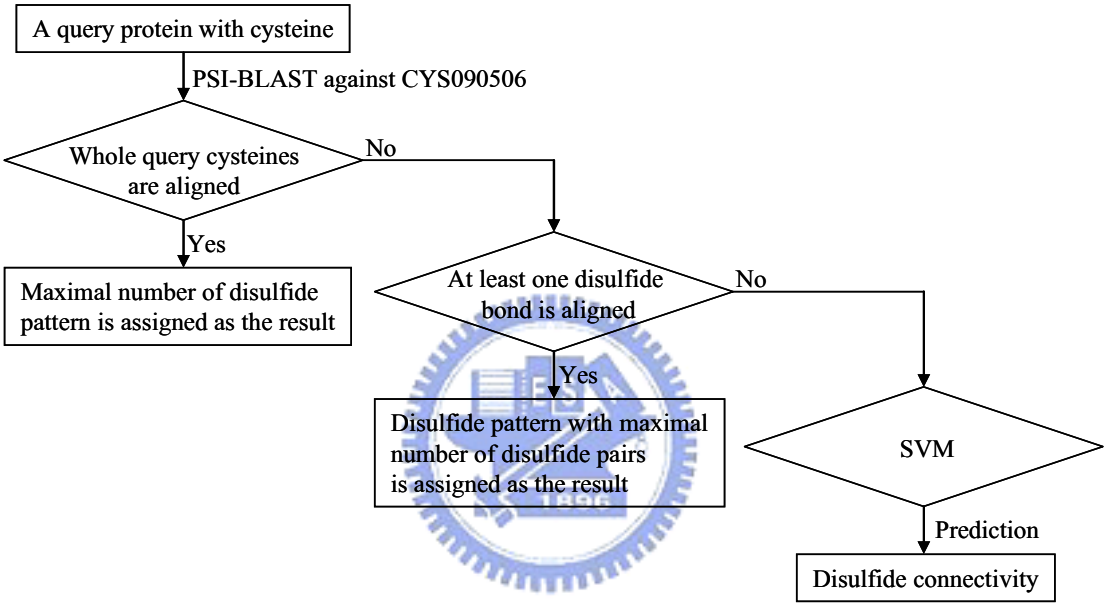
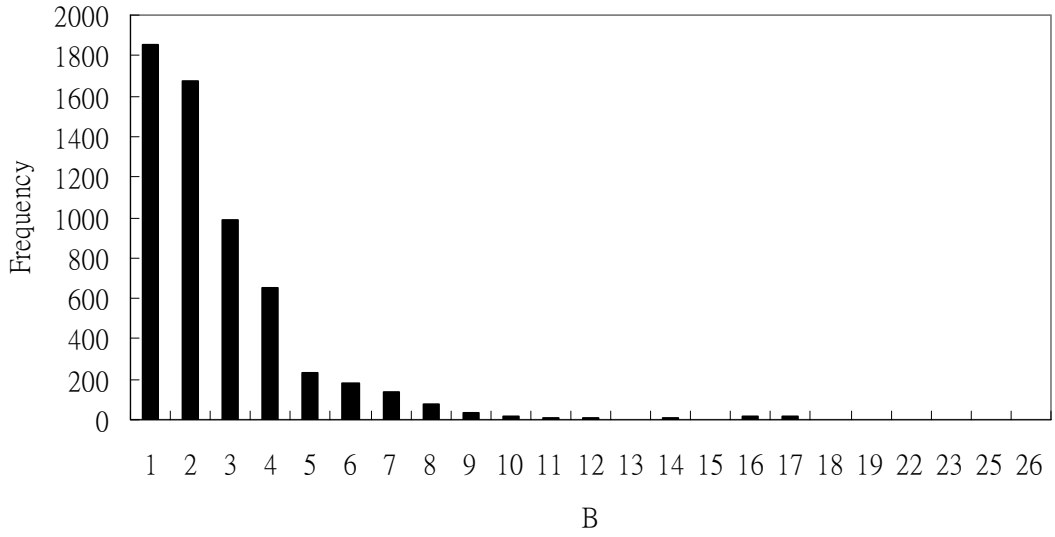


Figure 2. The flowchart of S-S Predictor.

A



B

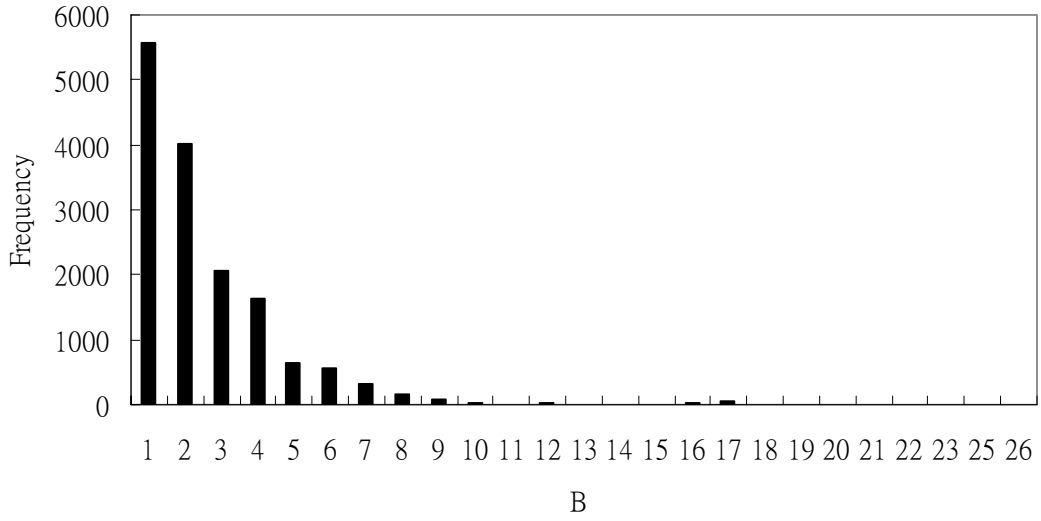
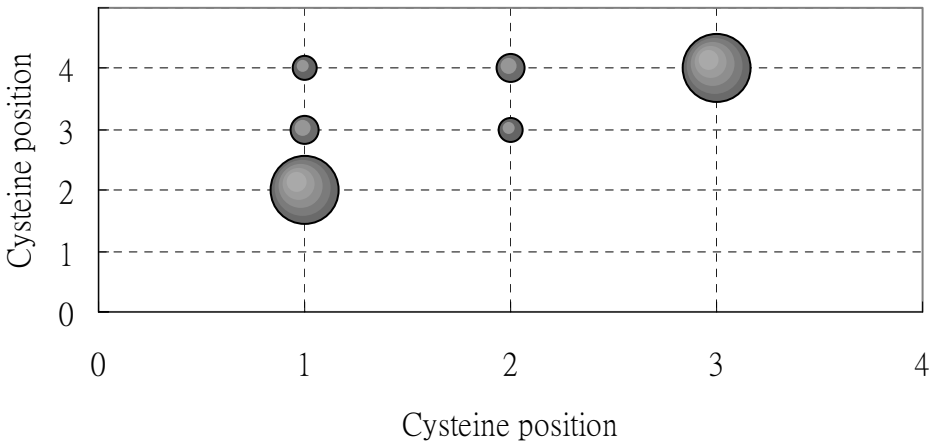
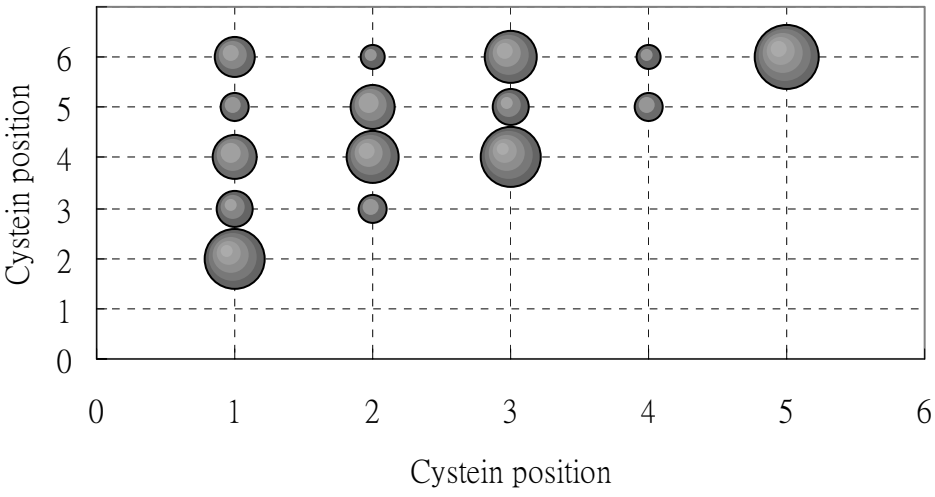


Figure 3. The number of disulfide proteins versus disulfide bond B in (A) CYS040307-NR and (B) CYS090506.

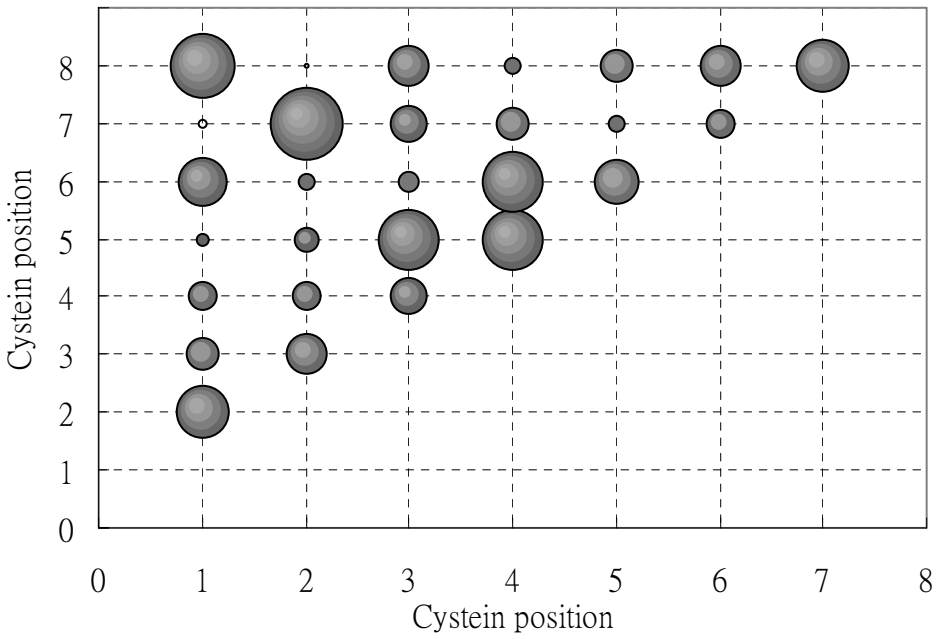
A


Cysteine position
0
1
2
3
4
Cysteine position

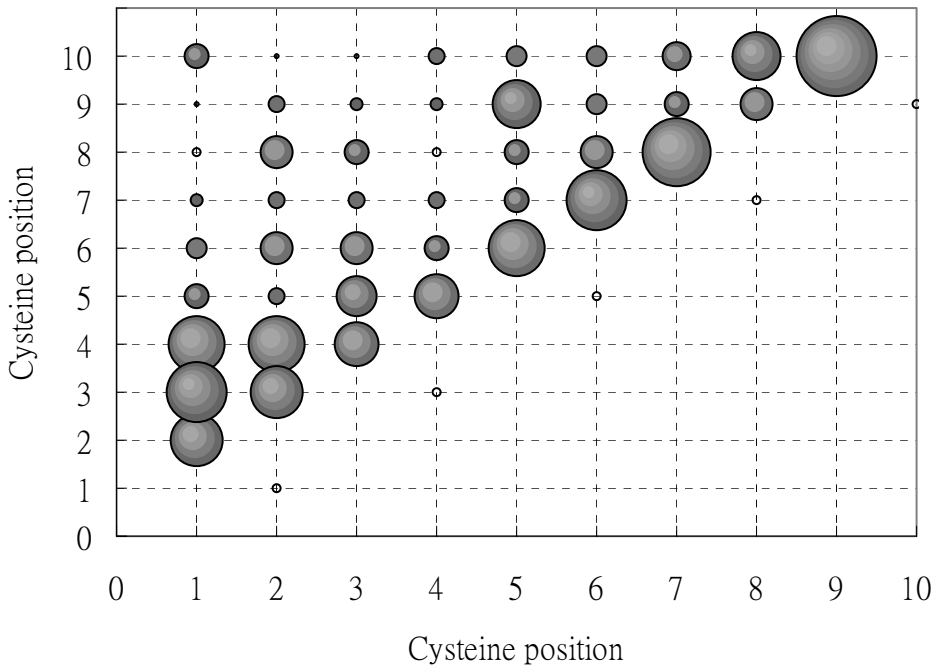
B


Cystein position
0
1
2
3
4
5
6
Cystein position

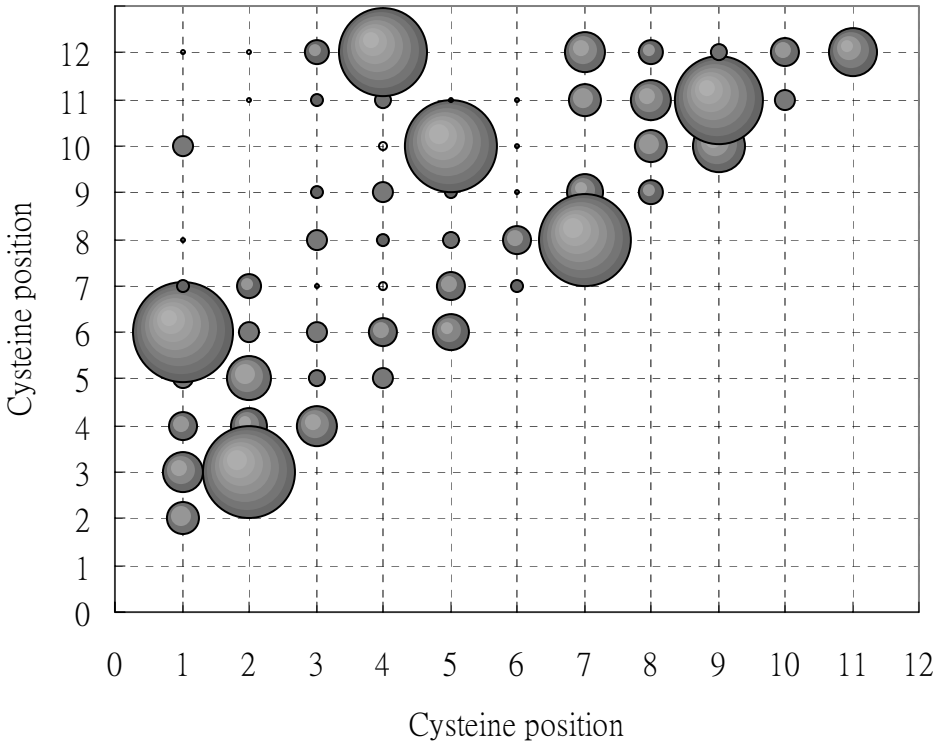
C


Cystein position
0
1
2
3
4
5
6
7
8
Cystein position

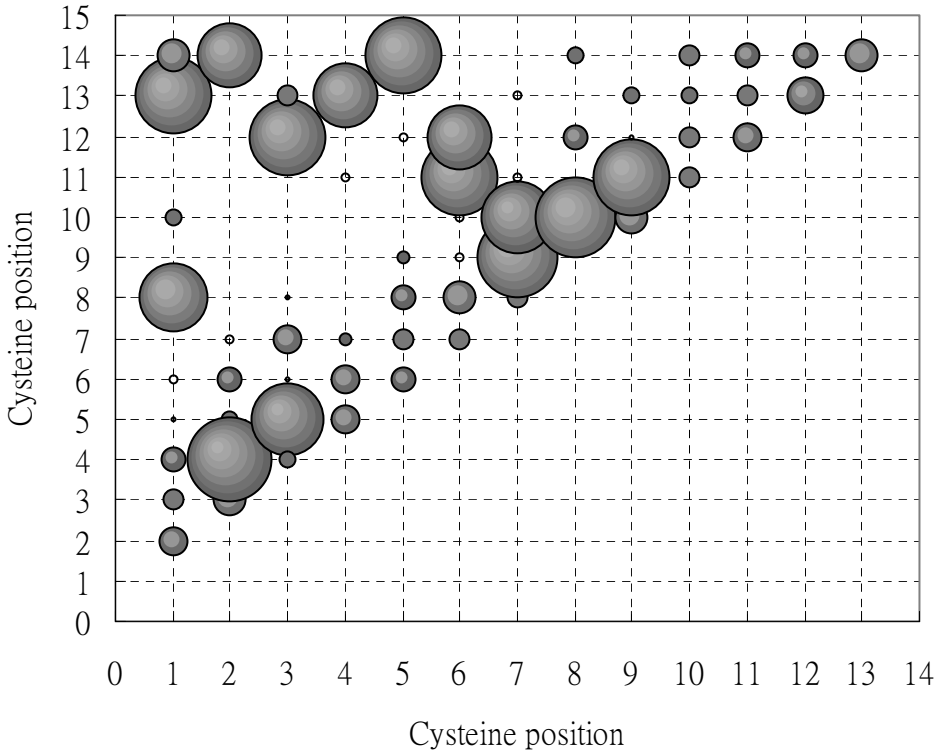
D


Cysteine position
0
1
2
3
4
5
6
7
8
9
10
Cysteine position

E


Cysteine position
0 1 2 3 4 5 6 7 8 9 10 11 12
Cysteine position
0 1 2 3 4 5 6 7 8 9 10 11 12

F


Cysteine position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Cysteine position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

G

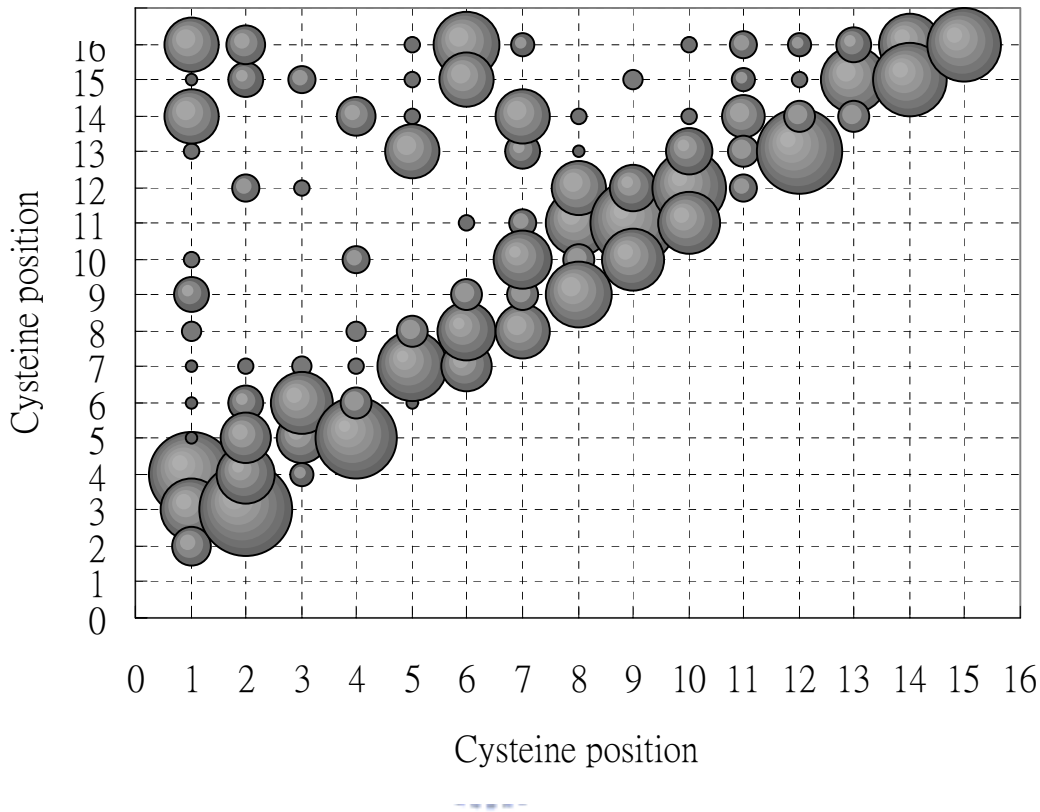


Figure 4. The distribution of disulfide connectivity based on disulfide pair for (A) B=2, (B) B=3, (C) B=4, (D) B=5, (E) B=6, (F) B=7, (G) B=8 in dataset CYS090506.

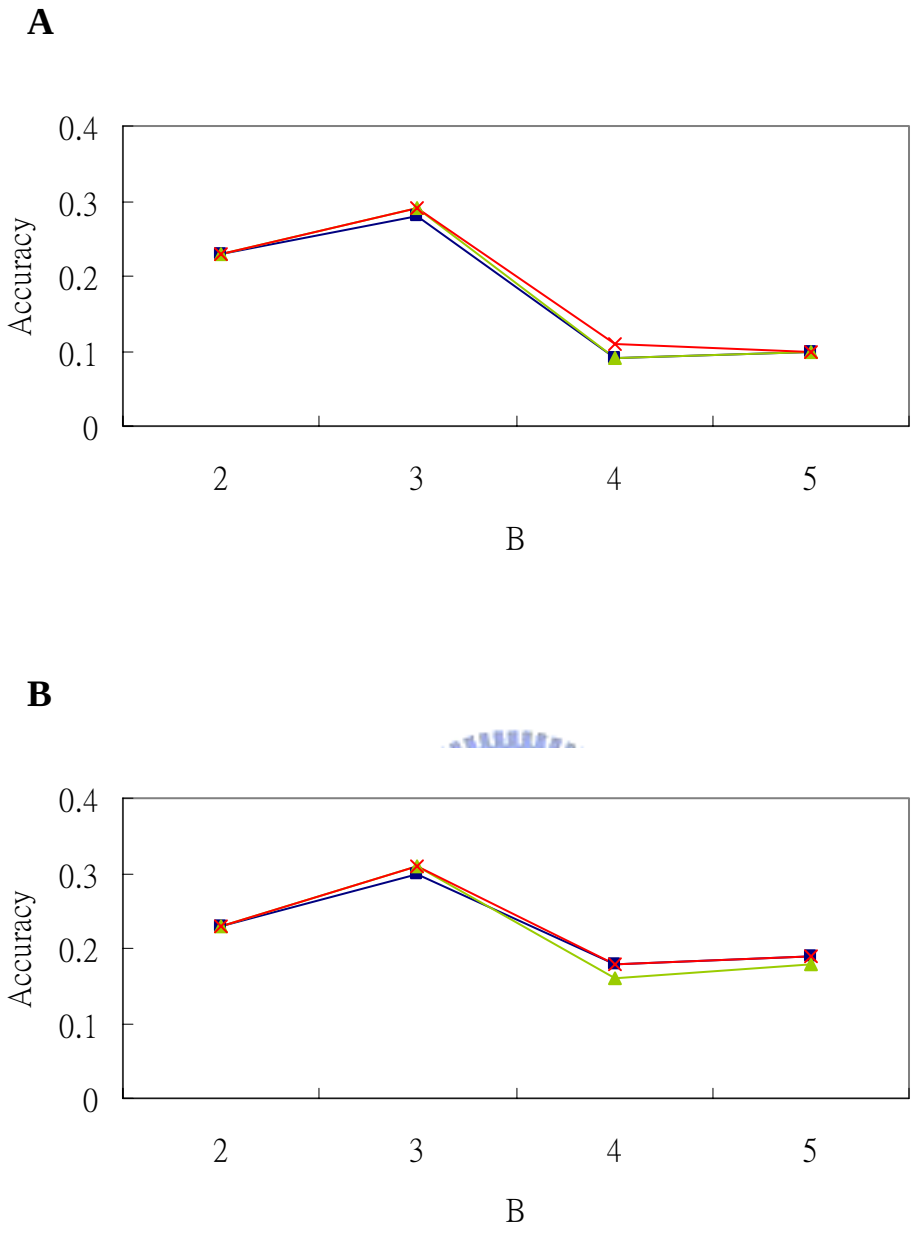


Figure 5. Comparison the performance based on different methods in dataset SP39-ID30 with (A) protein-based assessment index Q_p, and (B) cysteine pair-based assessment index Q_c, and the disulfide bond number in searching dataset is the same with query protein. Square (■) represents disulfide pair method, triangle (▲) represents disulfide pattern method, and cross (×) represents hybrid method.

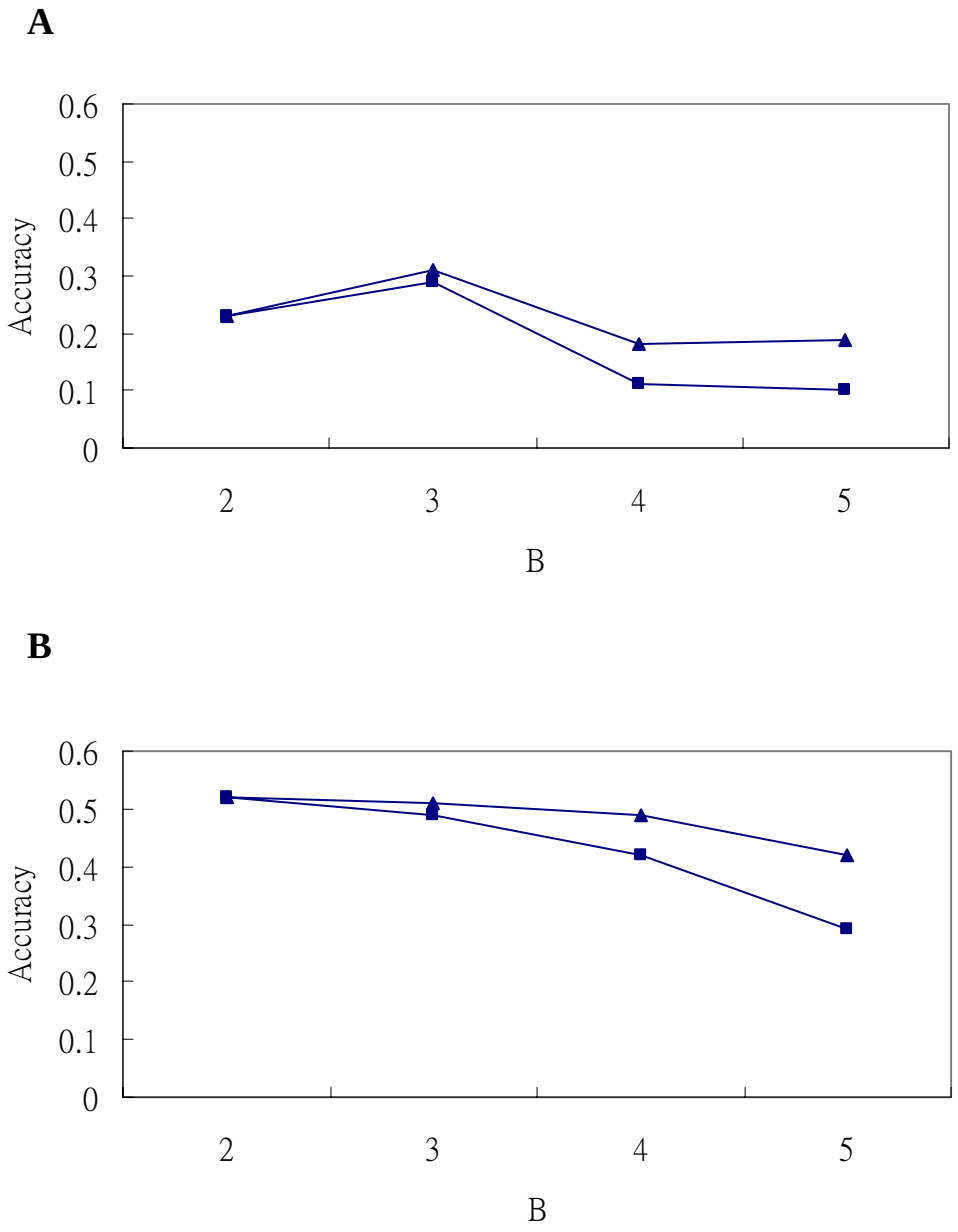
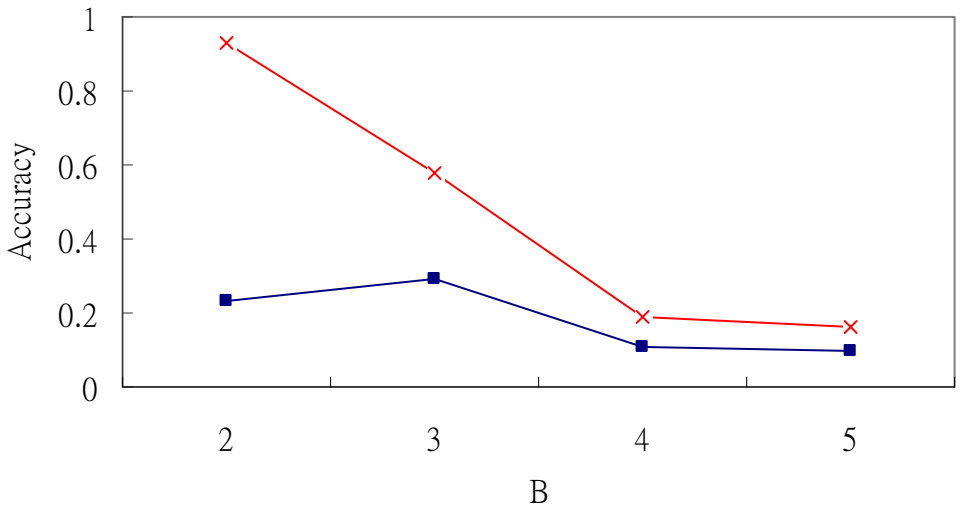


Figure 6. Comparison the performances based on (A) the number of disulfide bond in searching dataset is the same with query protein B, and (B) the number of disulfide bond in searching dataset include $B = 2,3,4,...5$, and the searching dataset is SP39-ID30. Protein-based assessment index Q_p (■), and cysteine pair-based index Q_c (▲) based on different criteria.

A



B

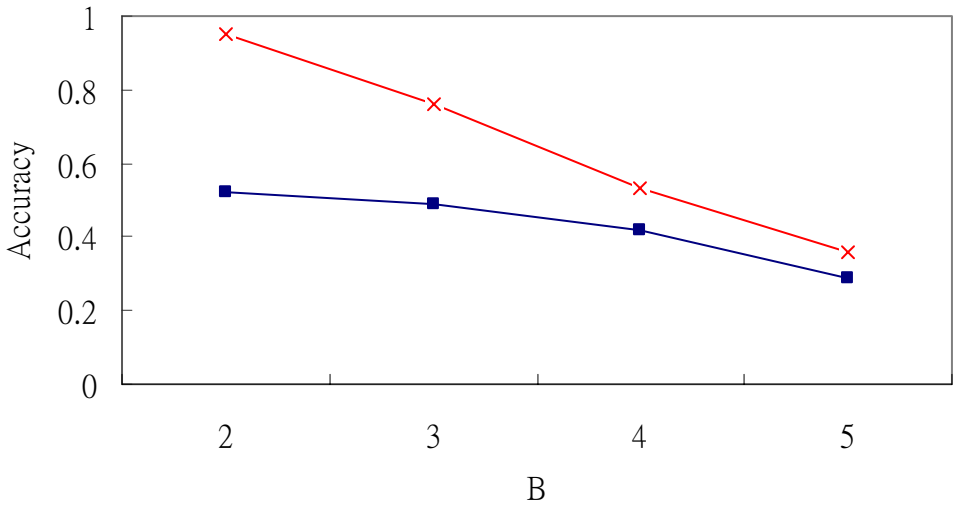
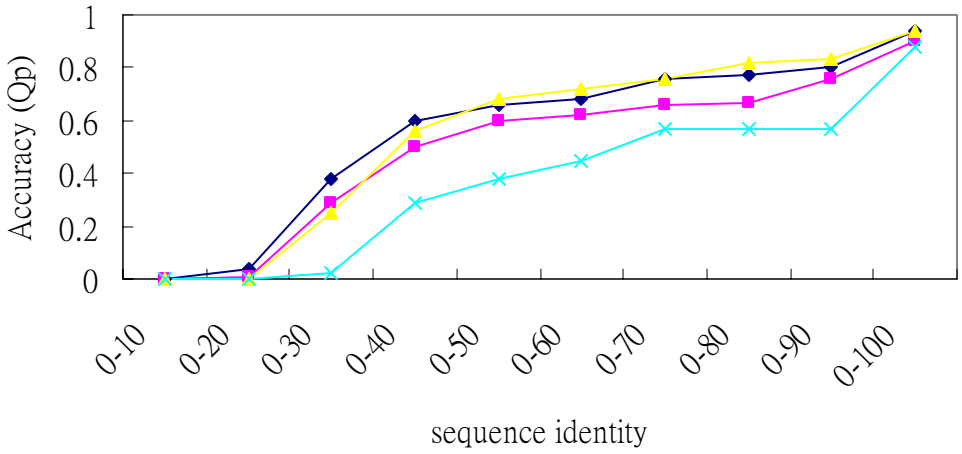


Figure 7. Comparison the performances based on (A) the number of disulfide bond in searching dataset is the same with query protein B, and (B) the number of disulfide bond in searching dataset include $B = 2,3,4,...5$, and the searching dataset is SP39-ID30. Protein-based assessment index Q_p (■), and the accuracy Q_s (×) based on the proteins whose disulfide connectivity have been predicted.

A



B

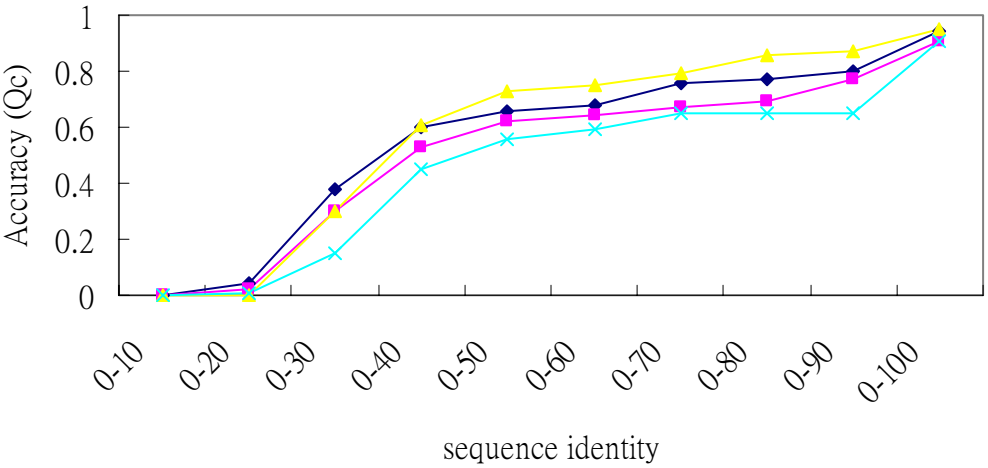


Figure 8. Comparison (A) protein-based assessment index Q_p, and (B) cystein pair-based index Q_c based on different sequence identity in searching dataset CYS090506. Diamond (◆) represents B=2, square (■) represents B=3, triangle (▲) represents B=4, and cross (×) represents B=5.

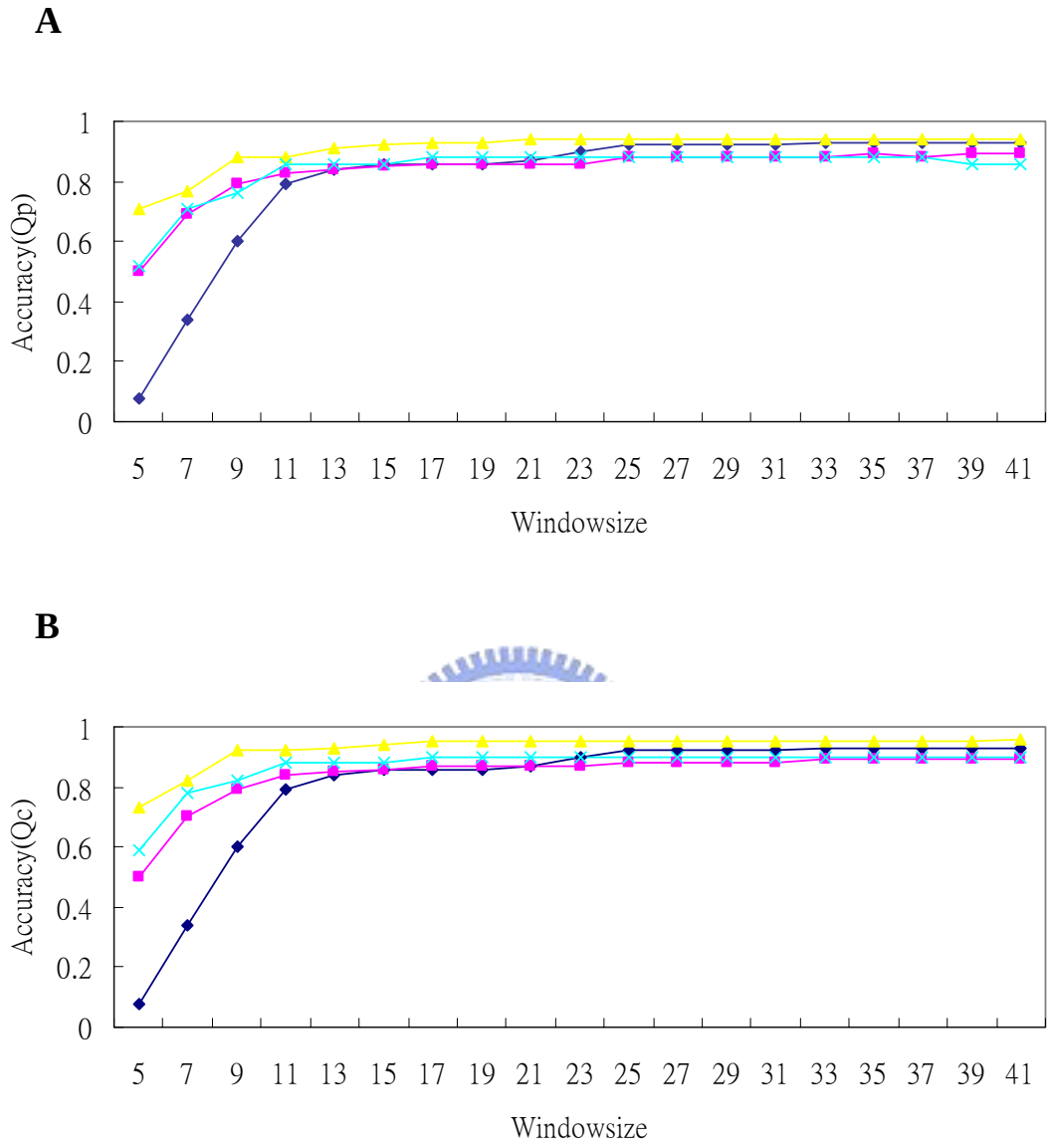
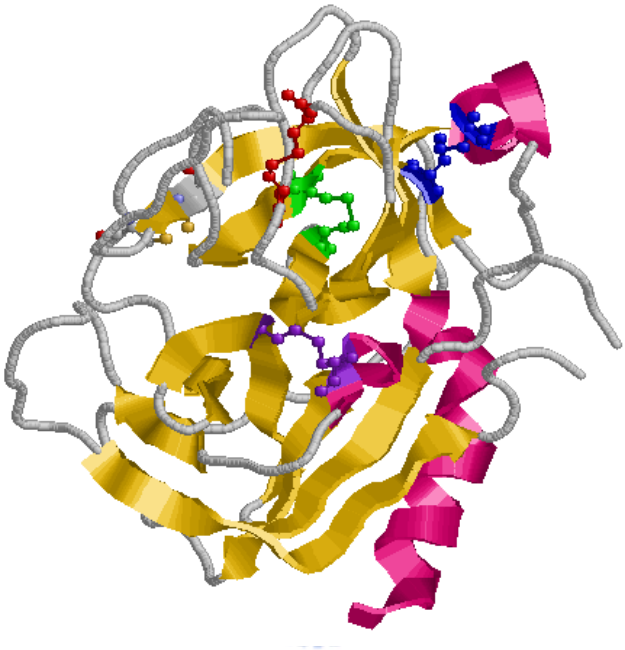


Figure 9. Comparison (A) protein-based assessment index Q_p, and (B) cystein pair-based index Q_c based on different windowsize in searching dataset CYS090506(*w*); *w* represents windowsize around cysteine. Diamond (◆) represents B=2, square (■) represents B=3, triangle (▲) represents B=4, and cross (×) represents B=5.

A



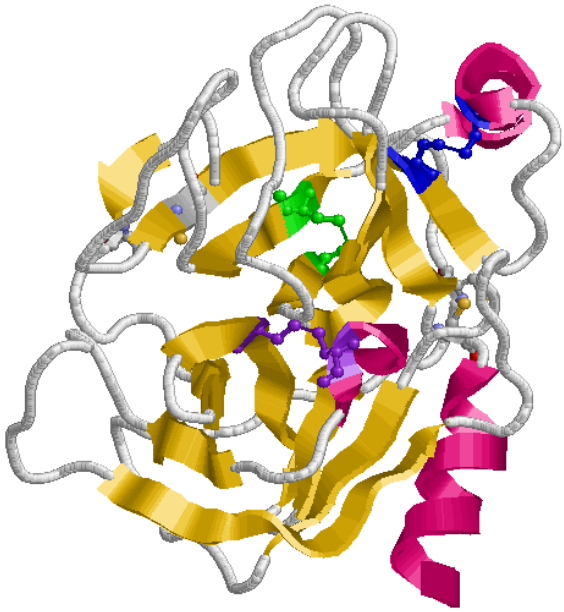
>2kaiA:26_42

IIGGRECEKN SHPWQVAIYH YSSFQCGGVL VNPKWVLTAA HCKNDNYEVWL
GRHNLFENE NTAQFFGVTA DFPHPGFN

>2kaiB:39_104 71_85 94_119

GKDYSHDLML LRLQSPAKIT DAVKVLELPT QEPELGSTCE ASGWGSIEPG
DFEFPDEIQC VQLTLLQNTF CADAHPDKVT ESMLCAGLPG GDTCMGDSGG
PLICNGMWQG ITSWGHTPCG ANKPSIYTKL IFYLDWIDDT ITENP

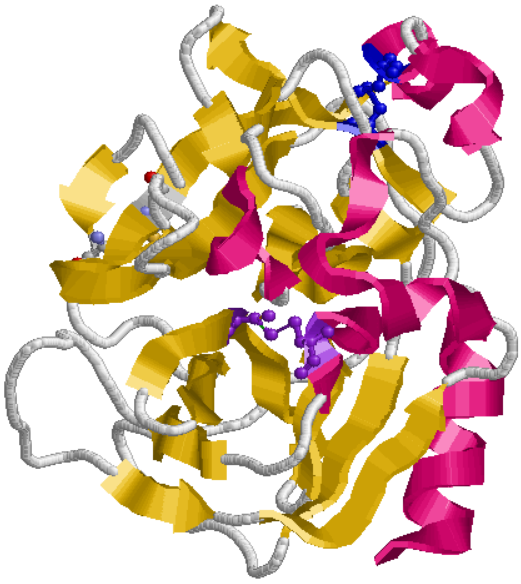
B



>1zr0C:7_137 25_41 109_207 116_181 148_162

IVGGYTCGAN TVPYQVSLNS GYHFCGGSLI NSQWVVSAAH CYKSGIQVRL
GEDNINVVEG NEQFISASKS IVHPSYNSNT LNNDIMLIKL KSAASLNSRV
ASISLPTSCA SAGTQCLISG WGNTKSSGTS YPDVLKCLKA PILSTSSCKS
AYPGQITSNM FCAGLEGGDS CQGDSGGPVV CSGKLQGIVS
WGSGCAKNKP GVYTKVCNYV SWIKQTIASN

C

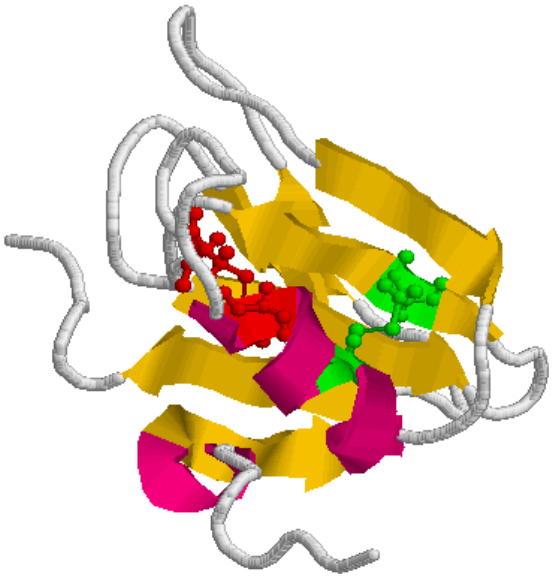


1gvzA:7_138 26_42 149_163

IIGGWECEKH SKPWQVAVYH QGHFQCGGVL VHPQWVLTAA HCMSDDYQIW
LGRHNLSKDE DTAQFHQVSD SFLDPQFDYD DISHDLMLLR LAQPARITDA
VKILDLPTQE PKLGSTCYTS GWGLISTFTN RGSGTLQCVE LRLQSNEKCA
RAYPEKMTEF VLCATDDSGS ICLGDSGGAL ICDGVFQGIT SWGYSECADF
NNFVFTKVMP HKKWIKETIE KNS

Figure 10. The cartoon models and sequences of (A) the KALLIKREIN A (2kai:A,B), (B) the cationic trypsin (1zr0:C), and (C) the KALLIKREIN (1gvz:A). The disulfide bonds are represented in the ball-and-stick model. Purple represents disulfide bond in proteins are aligned with disulfide bond 26_42, green are aligned with 121_190, blue are aligned with 155_169, and red are aligned with 180_205 in protein KLK_PIG.

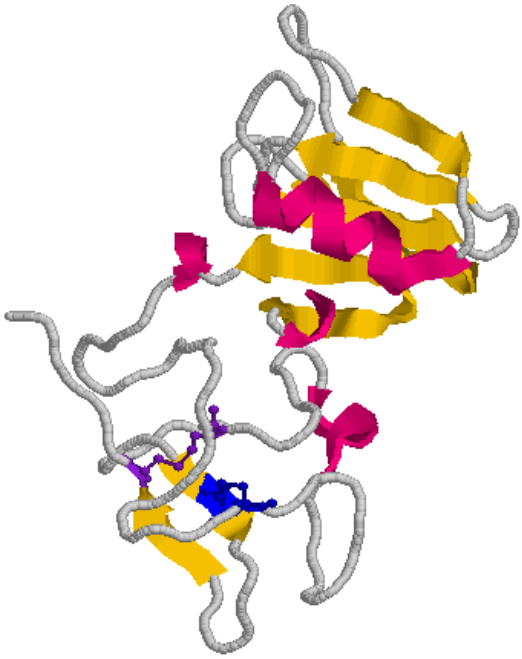
A



>2hgfA:40_66 44_54

GQRKRRNTIH EFKKSAKTTL IKIDPALKIK TKKVNTADQC ANRCTRNKGL
PFTCKAFVFD KARKQCLWFP FNSMSSGVKK EFGHEFDLYE NKDYIRN

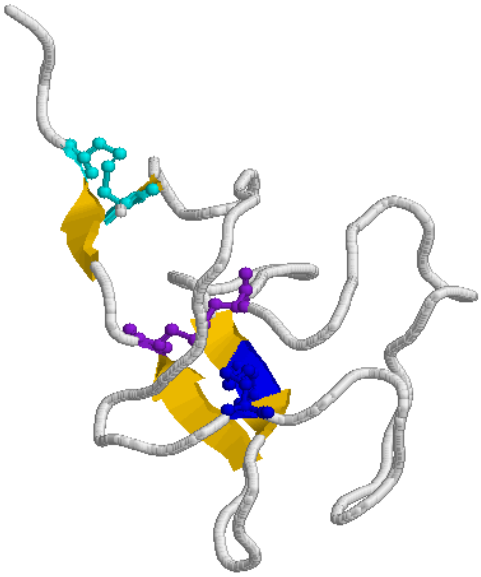
B



>1gmnA:112_152 140_164

TIHEFKKSAK TTLIKIDPAL KIKTKKVNTA DQCADRCTRN KGLPFTCKAF
VFDKARKQCL WFPFNSMSSG VKKEFGHEFD LYENKDYIRN CIIGKGRSYK
GTVSITKSGI KCQPWSSMIP HEHSFLPSSY RGKDLQENYC RNPRGEEGGP
WCFTSNPEVR YEVCDIPQCS

C



>1i71A:2_79 23_62 51_74

DCYHGDGQSY RGSFSTTVTG RTCQSWSSMT PHWHQRTTEY YPNGGLTRNY
CRNPDAEIRP WCYTMDPSVR WEYCNLTQCP VME

Figure 11. The cartoon models and sequences of (A) the HEPATOCYTE growth factor (2hgf:A), (B) the HEPATOCYTE growth factor (1gmn:A), and (C) the APOLIPOPROTEIN (1i71:A). The disulfide bonds are represented in the ball-and-stick model. Green represents disulfide bond in proteins are aligned with disulfide bond 70_96, red are aligned with 74_84, cyan are aligned with 128_206, blue are aligned with 149_189 and purple are aligned with 177_201 in protein HGF_MUNAN.

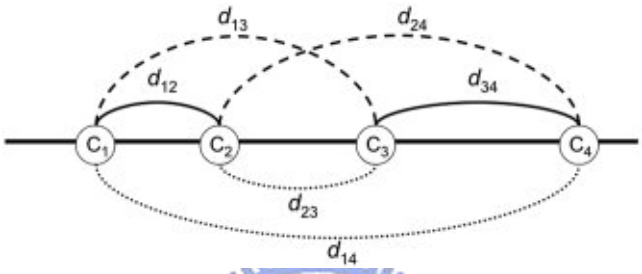


Figure 12. Example of disulfide patterns consisting of four cysteines $c_1c_2c_3c_4$, which form two disulfide bonds. Three possible disulfide patterns are (c_1c_2,c_3c_4), (c_1c_3,c_2c_4) and (c_1c_4,c_2c_4), where c_ic_j indicates a disulfide bridge between c_i and c_j. And the corresponding cysteine spacing patterns are given by (d_{12},d_{34}) (solid lines), (d_{13},d_{24}) (dashed line) and (d_{14},d_{23}) (dotted line).

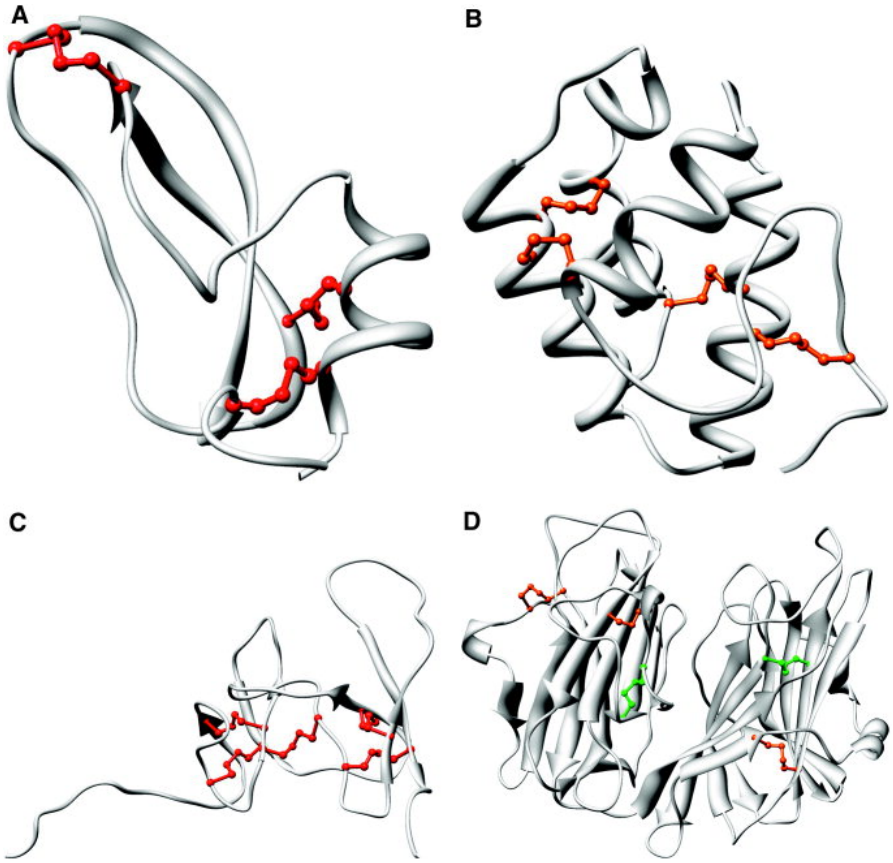


Figure 13. The ribbon models of (A) the bovine pancreatic trypsin inhibitor (1tpa:I), (B) the nonspecific lipid transfer protein (1afh), (C) porcine pancreatic procolipase (1pcn), and (D) peptidylglycine-hydroxylating monooxygenase (1phm). The disulfide bonds are represented in the ball-and-stick model. The correctly predicted disulfide bridges are in red, while the incorrectly predicted in green. The molecular images were generated by UCSF Chimera.[60]

Figure 14. Input form of S-S predictor.

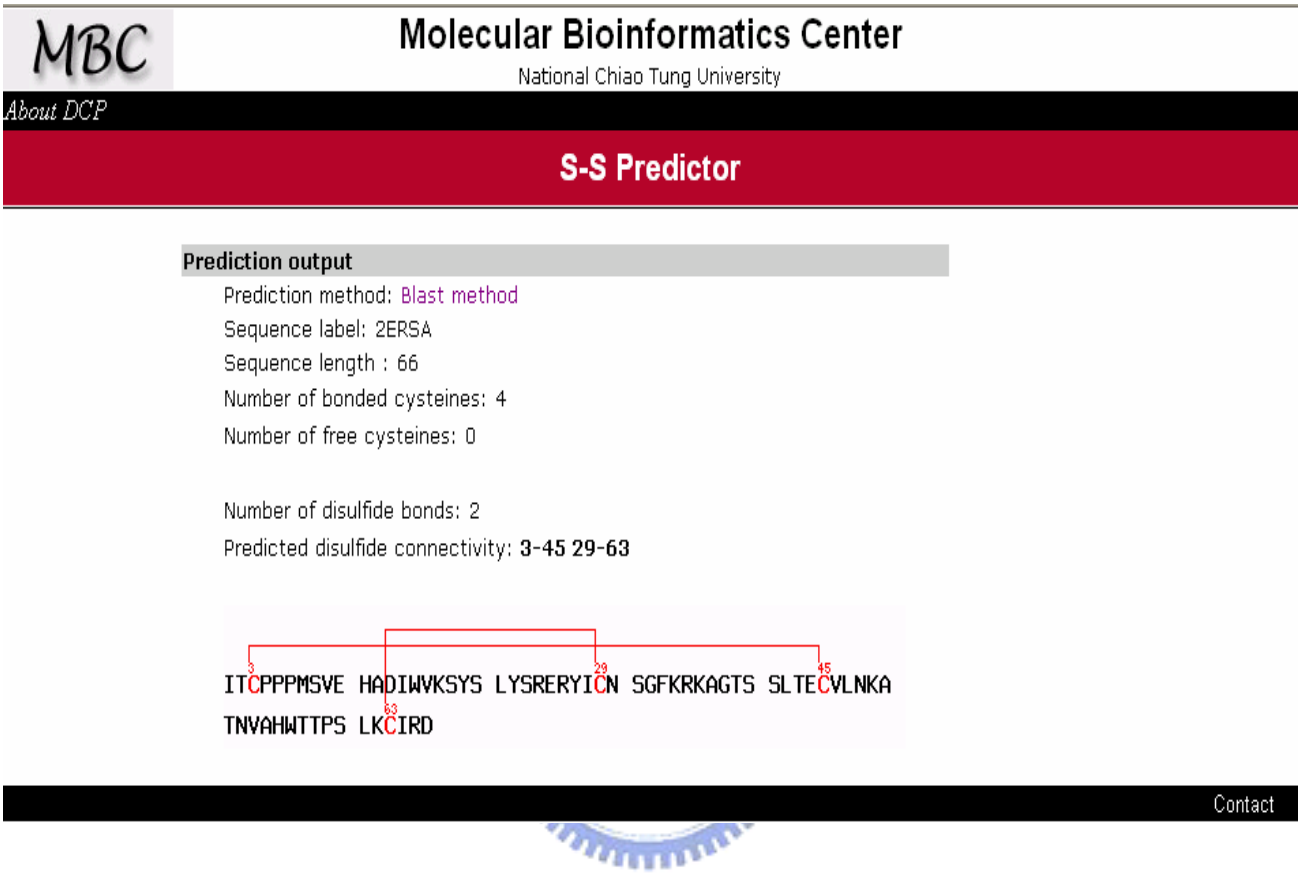


Figure 15. Output form of S-S predictor.