Available online at www.sciencedirect.com

Journal of Theoretical Biology 231 (2004) 85–95

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

Cooperativity of the oxidization of cysteines in globular proteins

Song Jiang-Ning[a,b,*], Li Wei-Jiang[a,b], Xu Wen-Bo[c]

[a]The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, 170 Huihe Road, Wuxi 214036, China
[b]School of Biotechnology, Southern Yangtze University, 170 Huihe Road, Wuxi 214036, China
[c]School of Information Technology, Southern Yangtze University, 170 Huihe Road, Wuxi 214036, China

Received 25 November 2003; received in revised form 1 June 2004; accepted 7 June 2004

Abstract

Based on the 639 non-homologous proteins with 2910 cysteine-containing segments of well-resolved three-dimensional structures, a novel approach has been proposed to predict the disulfide-bonding state of cysteines in proteins by constructing a two-stage classifier combining a first global linear discriminator based on their amino acid composition and a second local support vector machine classifier. The overall prediction accuracy of this hybrid classifier for the disulfide-bonding state of cysteines in proteins has scored 84.1% and 80.1%, when measured on cysteine and protein basis using the rigorous jack-knife procedure, respectively. It shows that whether cysteines should form disulfide bonds depends not only on the global structural features of proteins but also on the local sequence environment of proteins. The result demonstrates the applicability of this novel method and provides comparable prediction performance compared with existing methods for the prediction of the oxidation states of cysteines in proteins.

Keywords: Bonding state of cysteines; Two stage classifier; Amino acid composition; Support vector machine; Protein folding

1. Introduction

Disulfide bonds are primary covalent crosslinks between cysteine side chains that play very important roles in the native structures of globular proteins. Such bonds can stabilize protein spatial conformation and ensure that protein will perform its biochemical function (Wittrup, 1995). The correct formation of disulfide bonds is the crucial step in the folding pathway (Creighton, 1993, 1995). Many theoretical and experimental studies indicated that disulfide bridges can increase the conformational stability of proteins mainly by reducing the conformational entropy of the unfolded state and constraining the unfolded conformation (Betz, 1993; Skolnick et al., 1997; Abkevich and Shakhnovich, 2000; Clarke et al., 2000; Wedemeyer et al., 2000; Welker et al., 2001). Several analyses of the characteristics of disulfide bonds and detailed conformational analysis of cysteines as well as amino acid neighbors in proteins have been performed (Harrison and Sternberg, 1994; Petersen et al., 1999). But information of such important bonds cannot be derived directly from amino acid sequences. Numerous researches on disulfide bridges were reported, most of which were mainly time-consuming experimental works (Morris and Pucci, 1985; Matsumura and Matthews, 1989, 1991; Eder and Wilmanns, 1992; Zhou et al., 1993; Kremser and Rasched, 1994; Xue et al., 1994).

Disulfide-bonding pattern information can help understand structural properties of proteins and identify which family a protein belongs to, giving important insights into its biological functions. More recently, Chuang et al. found that there exists a very close relationship between the disulfide-bonding patterns and protein structures, based on which it is feasible to discriminate structure similarities and identify protein

*Corresponding author. Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, 170 Huihe Road, Wuxi 214036, China. Tel.: +86-510-5867519; fax: +86-510-5806493

E-mail address: sjnbeckham@yahoo.com.cn (S. Jiang-Ning).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.06.002

homologs (Chuang et al., 2004). van Vlijmen and his co-workers constructed a comprehensive database of disulfide-bonding patterns and developed search method to find related protein homologs with similar disulfide patterns (Van Vlijmen et al., 2004). In protein folding prediction, the localization of disulfide bridges can strongly reduce the search in the conformational space (Huang et al., 1999; Fariselli and Casadio, 2001). Thus the accurate predictions of disulfide connectivity in proteins would have potentially important applications, such as in introducing engineered disulfide bonds to increase the conformational stability of proteins and helping locate disulfide bridges to aid three-dimensional structure predictions.

Methodologies related to the prediction of disulfide bridges can be decomposed into two steps. First, the disulfide-bonding state of each cysteine is predicted from protein amino acid sequence, a typical binary classification problem. Subsequently, the second step is to locate the actual disulfide connectivity from candidate oxidized cysteines, which has received relatively scarce attention in the published literature. Fariselli and Casadio presented a method based on the weighted graph representation of disulfide bridges and achieved 17 times accuracy higher than that of a random predictor in the case of proteins with four disulfide bonds (Fariselli and Casadio, 2001). Afterwards another approach based on neural network was utilized to solve the pairing problem and received satisfactory results for the simplest cases (two or three disulfide bonds in one protein) (see example Fariselli et al., 2002). More recently, Vullo and Passerini proposed a novel machine learning method based on extended recursive neural networks (RNN) to predict the disulfide-connectivity patterns in cysteine-rich proteins (Vullo and Frasconi, 2004). They further improved the prediction performance by incorporating evolutionary information in the form of multiple alignment profiles.

This paper focused on the first task of the prediction of the disulfide-bonding state of cysteines in proteins, i.e. to predict which cysteines in protein sequence are oxidized. Concerning this topic, theoretical investigations emerged recent years. Muskal and his co-workers predicted the disulfide-bonding states of cysteines by means of neural networks (Muskal et al., 1990). They used local sequences, i.e., the flanking amino acid sequences of cysteines as input and achieved an overall accuracy of 80%. By adding evolutionary information, higher success rate can be obtained (Fariselli et al., 1999). Fiser et al. also used local sequence information but they employed statistical method. Their method performed at 71% prediction accuracy (Fiser et al., 1992). Since disulfide bridges are crucial to maintain proper structures of proteins, oxidized cysteines that take part in disulfide bonds should be more conserved than free cysteines. Based on this idea, multiple sequence alignment was used to predict the oxidation state of cysteines, the success rate of which was about 80% (Fiser and Simon, 2000). Mucchielli-Giorgi and his co-workers used logistic functions learned with subsets of proteins with similar amino acid compositions to predict the disulfide-bonding state and reached success rates close to 84% (Mucchielli-Giorgi et al., 2002).

Support vector machine-based predictor that operated at two stages (a multi-class classifier at the protein level and a binary classifier at cysteine level) was suggested (Ceroni et al., 2003). They achieved 85% accuracy measured by five-fold cross-validation. Martelli et al. implemented a hybrid system (hidden neural network) that combined a hidden Markov model (HMM) and neural networks (NN). After 20-fold cross-validation procedure, the predictor accuracy scored as high as 88% and 84%, measured on cysteine and protein basis, respectively (Martelli et al., 2002a, b). Up to now, this is the best-of-all prediction accuracy which has been achieved for the prediction of disulfide-bonding states of cysteines.

If one for the moment does not consider the more difficult problem of disulfide-connectivity prediction, the results of predicting the oxidation state of cysteines are relatively satisfied. But in its own nature, disulfide bonding is not merely a local interaction. It must be affected by some global factors of proteins as well as its local sequence environment. The present study has successfully constructed a new hybrid prediction system with two-stage architecture by combining a global linear classifier based on the overall amino acid composition and a local binary SVM classifier using the flanking subsequence surrounding the centered cysteine resides as input, to reveal the hidden information conductive to disulfide formation and provide an efficient prediction performance for the disulfide-bonding state of cysteines in proteins.

2. Database

Six hundred and thirty nine cysteine-containing protein chain structures were used in this work, which were taken from the PISCES Culled PDB (Wang and Dunbrack, 2003), a protein sequence culling server, which is a representative dataset of accurately resolved non-homologous Protein Data Bank (PDB) (Berman et al., 2000) structures. All structures used have resolution better than 2.5 Å. Sequence identity between each pair of the sequences is less than 25%. Structures with sequence length shorter than 50 amino acids were excluded. Information about disulfide bonds was extracted directly from the SSBOND records of the PDB entries.

According to whether containing intra-chain disulfide bonds, the 639 cysteine-containing protein chains were

Table 1
The PDB codes of 218 proteins containing disulfide bonds in the dataset

153L	1A6WL	1A6WH	1AAZA	1ABRB	1AC5	1AGQA	1AHO	1AIR	1AISA
1ALKA	1ALU	1AMP	1AOCA	1AOZA	1APA	1APYA	1APYB	1AQB	1AQZA
1ARB	1ARU	1ATLA	1AU1B	1AUK	1AV4	1BEBA	1BEO	1BGC	1BGP
1BHP	1BNDA	1BNDB	1BOVA	1BPI	1BTL	1CELA	1CEX	1CFB	1CGT
1CNSA	1CNV	1CPO	1CTN	1CVL	1DANL	1DANH	1DANT	1DANU	1DDT
1DPE	1EAGA	1ECEA	1ECY	1EDMB	1EPTA	1EPTC	1ESC	1EXTA	1EZM
1FLEE	1FLEI	1FUS	1FVKA	1FXD	1G3P	1GAI	1GAL	1GEN	1GOF
1HCGA	1HCGB	1HIAA	1HIAB	1HIAI	1HOE	1HPLA	1HSBA	1HSBB	1HSSA
1HTRB	1HUCA	1HUCB	1HXN	1HYP	1IAE	1IDK	1ILR1	1IMBA	1IVYA
1JER	1JETA	1JFRA	1JMCA	1JPC	1KLO	1KPTA	1KSIA	1KTE	1KUH
1KVEA	1KVEB	1LBEA	1LBU	1LIT	1LKI	1LPBA	1LPBB	1LST	1LT5D
1LTSD	1MHLA	1MHLC	1MPP	1MUP	1MZM	1NEU	1NNC	1NOYA	1NSCA
1NWPA	1OBR	1ONC	1OVA	1PGS	1POA	1POC	1PPN	1PYTB	1PYTC
1PYTD	1RCB	1RFS	1RGEA	1RIE	1RMG	1SFP	1SMD	1SMNA	1SMPI
1SRA	1SVB	1TABE	1TABI	1TCA	1TDE	1TF4A	1TFE	1TGSZ	1TGSI
1TGXA	1THG	1THV	1TIID	1TML	1TN3	1TVDA	1UKZ	1UMAH	1VCAA
1VMOA	1WBA	1WHTA	1XJO	1XSOA	1YAIA	1ZXQ	2AAA	2ACK	2AMG
2AYH	2BBKH	2BBKL	2CBP	2CTC	2DNJA	2ENG	2ERL	2GMFA	2HLCA
2ILK	2LIV	2MCM	2MPRA	2MSBA	2MTAH	2MTAL	2OVO	2PKAA	2PKAB
2PSPA	2RHE	2SAS	2SGA	2SICI	2SIL	2TGI	2TRXA	2VPFA	2WEA
3CD4	3EBX	3FRUA	3FRUB	3GRS	3LADA	3LZT	3PTE	3SEB	3TGL
4AAHA	4AAHB	4HTCH	4HTCI	5PTP	7RSA	8FABA	8FABB		

divided into two classes, which are called OXICYS and REDCYS for convenience. Proteins in REDCYS class have no intra-chain disulfide bridges, all cysteines are in reduced form. Every protein in OXICYS class has at least one disulfide bond. Among the total 639 protein chains, there are 218 chains belonging to OXICYS and 421 chains belonging to REDCYS, with totally 1316 cysteine-containing segments in the disulfide-bonded state forming 584 disulfide bonds and 1594 in the non-disulfide-bonded state. Two hundred and Eighteen PDB codes containing SSBOND records are shown in Table 1.

3. Method

In this paper, we construct a prediction system for disulfide-bonding state of cysteines in proteins operating at two stages by combining a first-stage global linear predictor based on the protein basis and a second-stage local predictor based on the cysteine level. Both the global classifier and the local one are binary predictors to classify two states of proteins (OXICYS and REDCYS protein) or cysteines (oxidized and reduced cysteines).

3.1. The global classifier-linear discriminant classifier using amino acid composition based on protein level

The first binary classifier uses the global information—20 amino acid composition as input to discriminate the two protein classes (OXICYS and REDCYS protein).

The 218 OXICYS proteins have 1316 cysteines, of which 1168 take part in intra-chain disulfide bonds. That is to say, almost all (89%) cysteines in OXICYS proteins are oxidized. While 1594 cysteines in 421 REDCYS proteins are all in free form. This is an obvious cooperation phenomenon that cannot be elucidated by only local sequences near cysteines. We call this phenomenon as "the cooperativity of oxidation of cysteines in globular proteins". This cooperativity is a global characteristic that reflects properties concerning protein structure, and there must be some global sequence information to account for it.

This key fact that cysteines (REDCYS) and half cysteines (OXICYS) rarely co-occur was also noticed by other researchers (Mucchielli-Giorgi et al., 2002; Ceroni et al., 2003) before. In the present paper, we proposed a new two-class predictor for predicting the oxidation state of cysteines in proteins by means of a linear discriminator, which explores the overall 20 amino acid composition of protein sequence.

For a protein k in the dataset, we define a characteristic index Q_k,

$$Q_k = \begin{cases} +1, \text{if protein } k \text{ belongs to OXICYS class,} \\ -1, \text{if protein } k \text{ belongs to REDCYS class.} \end{cases} \tag{1}$$

We try predicting the characteristic index Q_k of protein k by means of its amino acid composition $p_a^{(k)}$. We use the simplest linear function of $p_a^{(k)}$ to approximate Q_k, namely,

$$Q_k = \sum_a v_a p_a^{(k)}, \tag{2}$$

where a stands for an amino acid, and the summation runs over all the 20 types of amino acids. The parameters v_a are constants for all proteins. To choose the parameters v_a that best fit the dataset, we minimize

$$Z = \sum_k \left(Q_k - \sum_a v_a p_a^{(k)} \right)^2, \tag{3}$$

by letting $\partial Z/\partial v_b = 0$ for all amino acids b, which lead to

$$\sum_a \left(\sum_k p_a^{(k)} p_b^{(k)} \right) v_a = \sum_k Q_k p_b^{(k)}, \tag{4}$$

where the summations on k run over all protein sequences in the dataset. By solving Eq. (4), the fitted parameters v_a could be obtained.

With these parameters one can calculate the quantity Q for a given protein with amino acid composition p_a as follows:

$$Q = \sum_a v_a p_a, \tag{5}$$

which is designed to approach the characteristic index of the protein (+1 for OXICYS and −1 for REDCYS). To test the fitness, we computed the following cumulate distributions

$$F_{OXICYS}(Q_c) = \frac{\text{The number of OXICYS proteins with } Q \geqslant Q_c}{\text{The number of all OXICYS proteins}} \tag{6}$$

and

$$F_{REDCYS}(Q_c) = \frac{\text{The number of REDCYS proteins with } Q < Q_c}{\text{The number of all REDCYS proteins}}. \tag{7}$$

where Q_c is a critical value for the classification of the two protein classes.

3.2. The local classifier-support vector machine classifier based on cysteine level

The second binary classifier is constructed using support vector machine (SVM) method based on the cysteine level, which utilizes the local sequence context information-sequence segments flanking the centered cysteine as input to differentiate from the disulfide-bonded cysteines and non-disulfide-bonded cysteines in the protein.

SVM is a kind of learning machine based on well-developed statistical learning theory and a very effective method for general purpose pattern recognition, which was initially proposed by Vapnik and his co-works (Vapnik, 1995, 1998). SVM approach has been successfully applied to deal with a wide range of problems including drug design (Burbidge et al., 2000), text classification (Joachims, 1999), microarray data analysis (Brown et al., 2000), membrane protein types prediction (Cai et al., 2004), peptidyl prolyl *cis*/*trans* isomerization prediction (Wang et al., 2003), protein secondary structure prediction (Hua and Sun, 2001a; Kim and Park, 2003), protein structural class prediction (Cai et al., 2002, 2003), protein subcellular location prediction (Cai et al., 2000; Hua and Sun, 2001b; Chou and Cai, 2002), etc. In most of these cases, the SVM approach provides comparable or superior performance to that of other machine learning approaches.

Here, we will briefly describe the basic idea of applying SVM method for pattern recognition, especially for the two-class classification problem in this paper. For a comprehensive description of SVM, readers could refer to Vapnik's books (Vapnik, 1995, 1998).

For the two-class classification problem (disulfide-bonded cysteines versus non-disulfide-bonded cysteines) in this study here, suppose that we have a set of samples, i.e. a series of input vectors $\vec{x_i} \in R^d (i = 1, 2, ..., N)$ with corresponding labels $y_i \in \{+1, -1\} (i = 1, ..., N)$, where +1 and −1 indicate the positive and negative samples of the two classes, respectively. In this research, the input vector dimension is 20, and each input unit is a sequence segment flanking on the centered cysteine residue with a sliding window length $l = 2k + 1 (k = \ldots, 7, 8, 9, 10, \ldots)$. The input sequence in SVM is coded by transforming the 20 amino acids into numerical forms composed of only 0 and 1 (Ala = 100000…000, Cys = 010000…000,…, Tyr = 000000…001). 1 and −1 denoted the disulfide-bonded and non-disulfide bonded cysteine, respectively.

As shown in Fig. 1, the basic idea of SVM can be illustrated as follows: First, map the input vectors into a possible higher-dimensional feature space, associated with the selection of proper kernel function. Second, seek an optimal separating hyperplane (OSH) in this space which maximizes the distance from the dataset, separating the two classes (See Fig. 1). The mapping is typically achieved by the kernel function $K(\vec{x_i}, \vec{x_j})$ that defines the inner product in the feature space. There are two typical kernel functions:

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d, \tag{8}$$

$$K(\vec{x_i}, \vec{x_j}) = \exp(-r||\vec{x_i} - \vec{x_j}||^2), \tag{9}$$

where Eq. (8) is the *polynomial kernel function* of degree d which will revert to the linear function when $d = 1$, and Eq. (9) is the *radial basic function* (RBF) kernel with one parameter r.

For a given dataset, only the kernel function and the regularization parameter C should be selected to specify one SVM. In the present study, we finally selected the polynomial kernel function to train the SVM. The polynomial kernel function was defined as $K(\vec{x_i}, \vec{x_j}) =$

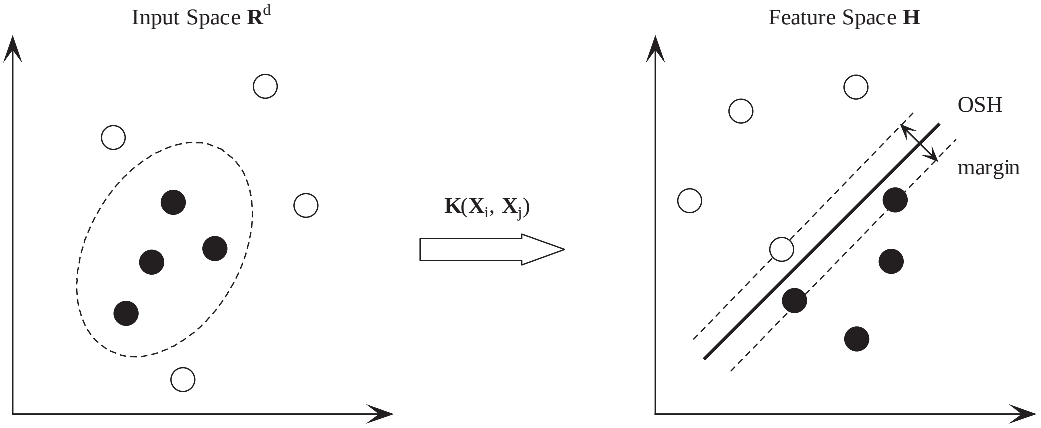
Input Space $\mathbf{R}^d$
Feature Space **H**
$\mathbf{K}(\mathbf{X}_i, \mathbf{X}_j)$
OSH
margin

Fig. 1. Basic idea of SVM application for pattern recognition. Two classes denoted by circles and disks, respectively, are linear non-separable in the input space $\mathbf{R}^d$. SVM constructs the optimal separating hyperplane (OSH) (continuous line) which maximizes the margin between two classes by mapping the input space into a high-dimensional space (the feature space **H**) by employing a mapping function $K(\vec{x_i}, \vec{x_j})$. Support vectors are identified with an extra circle.

$(\vec{x_i} \cdot \vec{x_j} + 1)^d$ with the parameter C, the default value in the implementation and $d = 7$, 8 and 9.

We downloaded the SVM$^{\text{light}}$ at http://download.joachims.org/svm_light/current/svm_light_windows.zip, which is an implementation (in C language) of Vapnik's SVM for the problem of pattern recognition, for the problem of regression, and for the problem of learning a ranking function. The optimization algorithms used in SVM$^{\text{light}}$ can be found in Joachims (1999, 2002).

3.3. *Construction of two-stage classifier prediction system*

In this paper, we have combined the first global classifier with the second local one described above to construct a two-stage binary classifier prediction system. The goal is to present a new method to provide the most accurate predictions for the disulfide-bonding state of cysteines in proteins. The architecture of this two-stage classifier prediction system could be depicted as the following Fig. 2.

If the tested protein is predicted as the REDCYS protein ($Q < Q_0$), then all the cysteines in this protein will be predicted as the reduced cysteines, whereas if the tested protein is predicted as the OXICYS one ($Q > Q_0$), we will apply the following prediction strategy: First, count the number of this protein's cysteines NC, for the case of NC = even, the tested protein will be classified as one of the two classes: protein with even free cysteine numbers and protein with all oxidized cysteines. For the case of NC = odd, the tested protein could be also regarded as the protein with odd free cysteine numbers. Once if the protein is assigned as the OXICYS protein, we will apply the second SVM classifier to predict the cysteines' disulfide-bonding states in this protein.

3.4. *Measurement accuracy*

The prediction quality was examined using the jack-knife test (leave-one-out procedure), an objective and rigorous testing procedure. In comparison with sub-sampling test or independent dataset test, the jack-knife test is thought to be more rigorous and reliable (Mardia et al., 1979). During the process of jack-knife test, each protein was singled out in turn as a test protein with the remaining proteins used as training set to calculate the test sample's v_a parameters and predict the class (OXICYS class or REDCYS class). The prediction quality was evaluated by the overall prediction accuracy and prediction accuracy for each cysteine and each protein chain.

Denote n_{xy} the number of proteins that are predicted as x class and in fact they belong to y class, where x, $y = \text{o}$ (OXICYS), or r (REDCYS). Therefore, the overall prediction accuracy is

$$Q2 = P/N = \frac{n_{oo} + n_{\text{rr}}}{n_{oo} + n_{or} + n_{rr} + n_{ro}}, \tag{10}$$

where P is the total number of correctly predicted cysteines, and N is the total number of cysteines.

The other measure of prediction accuracy is Matthew's correlation coefficient (MCC) (Matthews, 1975) between the observed and predicted cysteines, based on the cysteine basis or between the observed and predicted proteins, based on the protein basis, as given by

$$MCC(s) = \frac{p(s)n(s) - u(s)o(s)}{\sqrt{(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))}}. \tag{11}$$

Here, for each class s (OXICYS class or REDCYS class), $p(s)$ and $n(s)$ are the total number of correct

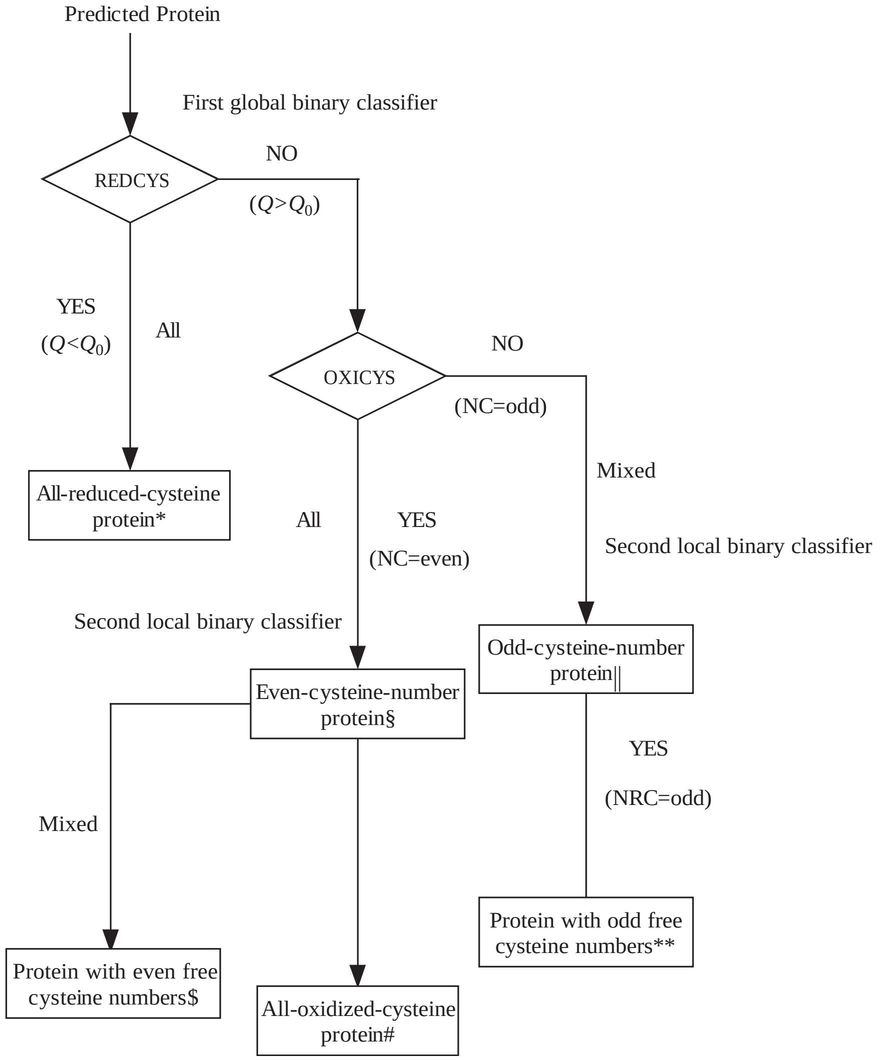
Predicted Protein
First global binary classifier
REDCYS
NO
(Q>Q0)
YES
(Q<Q0)
All
All-reduced-cysteine protein*
OXICYS
NO
(NC=odd)
Mixed
All
YES
(NC=even)
Second local binary classifier
Second local binary classifier
Odd-cysteine-number protein||
Even-cysteine-number protein§
YES
(NRC=odd)
Mixed
Protein with odd free cysteine numbers**
Protein with even free cysteine numbers$
All-oxidized-cysteine protein#

Fig. 2. The architecture of the two-stage classifier prediction system. It operates at two stages: the global protein stage and the local cysteine stage, where the former employs the linear discriminant classifier and the latter applies the binary SVM classifier.
*All cysteines in this protein are reduced ones.
$ **Concerning the mixed cysteines with different reduced–oxidized states in one protein, we classify these proteins into two categories: the protein with even free cysteine (considered as the reduced cysteine in this study) numbers and the protein with odd free cysteine numbers. NRC denotes the number of reduced cysteines.
§ The number of cysteines in this protein is even. It can be further categorized into two classes: the protein with even free cysteine numbers and the all-oxidized cysteine protein.
|| The number of cysteines in this protein is odd.
$ The number of free cysteines in this protein is even.
# All cysteines in this protein are oxidized ones.
** The number of free cysteines in this protein is odd. For this case, it indicates that there will be certainly at least one free cysteine in this protein.

predictions and correctly rejected assignments, respectively, and $u(s)$ and $o(s)$ are the number of under- and over-predictions. The more MCC is, usually the higher the prediction reliability is.

The accuracy for each discriminated class s is evaluated as

$$Q(s) = \frac{p(s)}{p(s) + u(s)}. \tag{12}$$

For sake of further explanation, when s refers to the OXICYS class and REDCYS, respectively, Eq. (12) equivalent to the following Eq. (13):

$$Q_{oxi} = \frac{n_{oo}}{n_{oo} + n_{or}}, Q_{red} = \frac{n_{rr}}{n_{rr} + n_{ro}}, \tag{13}$$

where Q_{oxi} and Q_{red} are the success rates for OXICYS and REDCYS class, respectively. $p(s)$ and $u(s)$ are the same as in Eq. (11).

Also, the probability of correct predictions $P(s)$ is calculated as

$$P(s) = \frac{p(s)}{p(s) + o(s)}, \tag{14}$$

where $n(s)$ and $o(s)$ are the same as in Eq. (11).

Finally, the prediction accuracy per protein is

$$Q2_{prot} = \frac{P_p}{N_p}, \tag{15}$$

where P_p is the number of the proteins whose cysteines are all correctly predicted and N_p is the total number of proteins.

We should point out that $Q2$ (prediction accuracy of reduced and oxidized CYS) and $Q2_{prot}$ (prediction accuracy of the type of proteins, OXICYS or REDCYS class proteins) should not give exactly the same prediction results, for the former and the latter are based on the cysteine level and protein level, respectively.

4. Result and discussion

4.1. Cumulative distribution of Q values

The cumulative distribution result of Q values was depicted in Fig. 3.

Fig. 3 shows clearly that the Q value is a good index to distinguish the two classes of proteins. Therefore, the classification of a protein can be predicted based on its Q value: If $Q > Q_c$ then the protein is predicted as OXICYS, otherwise REDCYS, where Q_c is a critical value. From Fig. 3, we could also observe that $Q_c = 0$ is usually not the best-fitted critical value, i.e. in most cases 0 and Q_c do not match each other. The highest prediction accuracy may be achieved at the value of Q_c less than zero.

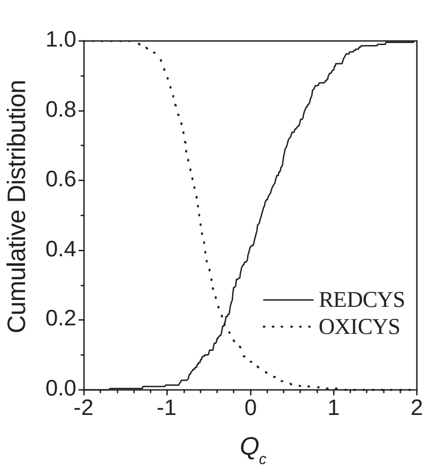

Fig. 3. Cumulative distribution of the Q values for OXICYS and REDCYS proteins. Solid line corresponds to the REDCYS-class proteins predicted, while dash line corresponds to the OXICYS-class proteins tested.

4.2. Cysteines conservation and sequence environment conducive to disulfide bond formation

Cysteines tend to be more conserved in proteins when they pair to form disulfide bridges, which may reflect their crucial and essential role in maintaining protein structure stability and biological functions. As shown in Fig. 4, the amino acid composition of proteins with OXICYS and REDCYS reveals clear difference. The analysis highlights that Cysteine (C), Asparagine (N), Serine (S), Threonine (T) and Tryptophan (W) are residues highly conducive to disulfide-bond formation. On the contrary, Glutamate (E), Histidine (H), Leucine (L), Methionine (M), Valine (V) and Arginine (R) are more frequently found in the case of reduced cysteines. These observations agree basically with previously reported results about the specific sequence environment of cysteines (Fiser et al., 1992; Fiser and Simon, 2000).

4.3. CATH structural classification for OXICYS and REDCYS proteins

Protein structures are determined by their amino acid sequence, which is a basically accepted hypothesis now. This outstanding work was first finished by Anfinsen, who successfully carried out an experiment to restore native structure of the pancreatic bovine ribonuclease in vitro (Anfinsen, 1973). A huge number of sequences have the same amino acid composition, so amino acid contents may contain very little sequence information. Though it may alter the local structures, in most cases shuffling sequence does not greatly change the global

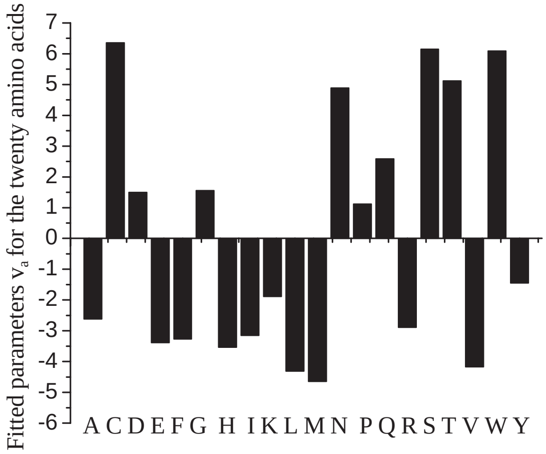

Fig. 4. Amino acid contribution to disulfide bond formation. Calculated v_a values can represent the propensity to form disulfide bond for the 20 amino acid residues. Bars above the midline indicate a propensity to disulfide-bond formation, and those below the midline are inclined to non-disulfide bond formation.

Table 2
Percentages of protein structures belonging to different CATH structural classifications

CATH classification[a]	Proteins used in this work			All CATH entries
	OXICYS	REDCYS	Entire dataset	
Mainly α	16.8	30.9	26.0	21.5
Mainly β	53.0	19.5	31.0	30.3
α and β	29.3	48.9	42.2	45.2
Few Secondary Structures	0.9	0.7	0.8	3.0

[a] CATH statistics are based on CATH Release 2.4. Several structures in our dataset have no CATH classifications and are not included in these statistics.

structural features of proteins, possibly because the protein structural classification can be well predicted by using amino acid contents (Zhang et al., 2001).

By counting the CATH structural classification (Orengo et al., 1997; Pearl et al., 2000) of the protein structures used in this work, it can be found that disulfide bond-containing proteins have some global structural features (Table 2): OXICYS proteins prefer β-structures and REDCYS proteins α-helices. Since β-sheets are less stable than helical structures, disulfide bonds may be necessary to maintain the native structures of the whole proteins that lack enough stable secondary structures.

4.4. Parameter optimization of the second SVM classifier

In the case of the second SVM classifier, we need to select the appropriate kernel function, regularization parameter C and the local input symmetrical sliding window size $l=2k+1$ (k denotes the number of flanking residues in positions from amino terminal to carboxyl terminal of each centered cysteine, and vice versa). The selection of the optimal kernel function parameters and the regularization parameter C plays an important role in improving the prediction accuracy. The optimization parameters are determined by the prediction performance.

We performed a preliminary test to determine the best-fitted kernel function type, and the optimal window size l by measuring the prediction accuracy of the various window sizes l from 13 to 21 (corresponding to k from 6 to 10). We also tried linear, polynomial, and radial basis function (RBF) kernel types. All the optimal parameters and functions used in the ultimate predictions are determined by choosing those leading to the best prediction performance. Eventually, we selected the polynomial kernel function of 8 degree ($d=8$) to perform the final SVM training and testing with the local subsequence window size $l=21$.

4.5. Prediction performance of the two-stage classifier prediction system

The jack-knife testing results are summarized in Table 3. As shown in Table 3, when we selected the "natural" value $Q_c=0$, the total prediction accuracy could be $Q2=83.3\%$, $Q_{oxi}=85.7\%$, $Q_{red}=78.9\%$, and $MCC=59.6\%$. When scored on a protein basis (we accept only those protein chains for which the predictions of all the disulfide- or non-disulfide-bonding states of the cysteines in the protein sequence are correctly predicted), the success rate $Q2_{prot}$ reaches up to 79.4%. However, fine-tuning of Q_c can slightly improve the prediction. In fact, if $Q_c=-0.1$ is chosen, the over prediction accuracy could be improved to $Q2=84.1\%$, $Q2_{prot}=80.1\%$, $Q_{oxi}=87.8\%$, $Q_{red}=77.8\%$, and $MCC=62.2\%$. Moreover, this prediction score can be further improved slightly by avoiding using those amino acids with the absolute v_a value less than 2 (data not shown), such as Proline (R) and Tyrosine (Y). In the case of these two amino acids, their absolute v_a values are both less than 1.5, for which the lower absolute v_a values may be due to the computationally statistical fluctuation.

The above prediction results suggest that this method could achieve relatively high prediction accuracy by taking into consideration the global characteristic of protein sequences based on the overall amino acid composition and the local sequence environment surrounding the target cysteines. It demonstrates that the overall amino acid contents do carry much information about disulfide bonding, as well as the flanking sequential context of cysteines, and it shows that determinant of whether the cysteines should form disulfide bridges owes not only to the global structural feature of a protein but also correlates with the striking local sequence context of cysteines in the protein. This finding is consistent with the observations drawn by Mucchielli-Giorgi et al. (2002) that predictor based on global descriptors is more accurate (77.7%) than that based on local descriptors alone (67.3%).

In conclusion, an efficient two-stage classifier prediction system composed of the first classifier based on the

Table 3
Prediction accuracy (%) of the two-stage classifier prediction system by the jack-knife test

Method	Prediction accuracy (%)					
	Q_c	MCC	Q_{oxi}	Q_{red}	$Q2$	$Q2_{prot}$
Global linear classifier + Local SVM classifier	−0.2	61.6	89.5	74.2	83.2	79.5
	−0.1	62.2	87.8	77.8	84.1	80.1
	0	59.6	85.7	78.9	83.3	79.4
	0.1	58.3	84.2	79.9	82.9	78.5
	0.2	55.7	81.7	80.7	81.9	77.9

overall amino acid composition using the simplest global descriptors of protein sequences together with the second SVM classifier using the local sequence context of cysteines in proteins as the input, is presented to reveal the hidden information of the disulfide-bonding states of cysteines. Even though it is difficult to compare all the existing methods tested on the different databases, it could be claimed that this new approach provides comparable prediction performance compared with the existing algorithms. Our studies support the phenomena that the oxidation of cysteines exhibits obvious cooperativity and demonstrate that amino acid contents carry much information about disulfide bonding. It is also shown that global structural feature of a protein as well as the local sequence environment of cysteines is the important determinant of whether the cysteines should form disulfide bonds.

4.6. Further improvemetns and other possible applications

There may be several directions for further improvement of the prediction performance. On the one hand, it should be pointed out that in the case of the first global classifier, the linear combination of amino acid contents may not be the best function for the purpose of prediction. Although the above results have demonstrated the capability of the simple linear discriminator to effectively discriminate the two cysteine classes (OXICYS and REDCYS), use of more complex functions (for example, the nonlinear polynomial functions) can possibly lead to better prediction results than the linear discriminator based classifiers. This aspect is worthy of a deeper investigation.

Moreover, the classification rule in the first global classifier to sort proteins into OXICYS and REDCYS proteins may be too simple. As suggested by Ceroni et al. (2003) and Frasconi et al. (2002), higher prediction accuracy is likely to be achieved by training and testing the homogeneous protein groups associated with their cellular compartments or domain structural classes. For example, it would be relevant to see whether amino acid composition of proteins belonging to different groups (OXICYS, REDCYS) correlates with their cellular position by iterating on a subset of the proteins that share the same cellular localization. There may be potentially interesting biological insights to be gained from the analysis of the cellular locations (Intracellular, extracellular, membrane, nuclear, etc.). However, detailed systematic analysis of these observations requires much more proteins sequence data derived from experimental studies.

On the other hand, single prediction methods do have limitations. A possible alternative strategy is to combine other complementary methods, such as neural networks (Fariselli, et al., 1999; Fiser and Simon, 2000), combinational logistic functions (Mucchielli-Giorgi et al., 2002), Hidden Markov models (Martelli et al., 2002), and fuzzy *k*-nearest-neighbor method (Huang and Li, 2004). Integration of other different methods incorporating more sequence-order information and evolutionary information together with global features and local sequence-order context may be likely to further improve prediction performance. Taking into consideration the conservation of disulfide bonds and the cysteines in proteins, it is anticipated to combine several methods to use protein primary sequence and three-dimensional structure information and construct the multistrategy approach to perfect the task of disulfide-bonding state of cysteines.

5. Conclusion

In the present study, a novel and efficient two-stage classifier prediction system combining a first global classifier based on the 20 amino acid composition with a second SVM classifier exploiting the local sequence context of cysteines in proteins, has been developed to discriminate the two protein classes (OXICYS and REDCYS proteins) and the two different redox state cysteines (disulfide- and non-disulfide-bonded cysteines). This novel approach provides at least comparable prediction performance compared with the existing methods and can be an efficient complimentary method to other existing methods for disulfide-bonding state prediction of cysteines in proteins. Our studies support the phenomena that the oxidation of cysteines exhibits obvious cooperativity and demonstrate that amino acid contents carry much information about disulfide bonding as well as the local sequence context of cysteines. The total prediction accuracy of this prediction system has achieved as high as 84.1% and 80.1%, when measured on cysteine and protein basis using the rigorous jack-knife procedure, respectively. The result indicates that global structural feature of the protein, as well as its local sequence environment of cysteines, is the important determinant of whether the cysteines should form disulfide bridges. The present studies demonstrate the applicability of this novel efficient method and provides at least comparative prediction performance compared with existing methods for the prediction of the oxidation states of cysteines in proteins.

Acknowledgements

This work was supported by Visiting Scholar Foundation of Key Laboratory in University, Ministry of Education of China (Grant No. 200102). The authors would like to thank Dr. Gexin Zhang for his kind support in this study. Thank Professor Thorsten

Joachims for making SVM$^{\text{light}}$ software available. The authors would also like to express their gratitude to an anonymous referee for carefully reviewing the manuscript, whose comments were very helpful in improving the presentation of this manuscript.

References

Abkevich, V.I., Shakhnovich, E.I., 2000. What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioinformatics analysis. J. Mol. Biol. 300, 975–985.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.

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

Betz, S.F., 1993. Disulfide bonds and the stability of globular proteins. Protein Sci 2, 1551–1558.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl Acad. Sci. 97, 262–267.

Burbidge, R., Trotter, M., Holden, S., Buxton, B., 2000. Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics, pp. 1–4.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2000. Support vector machines for prediction of protein subcellular location. Mol. Cell Biol. Res. Commun. 4, 230–233.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Prediction of protein structural classes by support vector machines. Comput. Chem. 26, 293–296.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2003. Support vector machines for prediction of protein subcellular location. J. Theor. Biol. 221, 115–120.

Cai, Y.D., Ricardo, P.W., Jen, C.H., Chou, K.C., 2004. Application of SVM to predict membrane protein types. J. Theor. Biol. 226, 373–376.

Ceroni, A., Frasconi, P., Passerini, A., Vullo, A., 2003. Predicting the disulfide bonding state of cysteines with combinations of kernel machines. J. VLSI Signal Process. 35, 287–295.

Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. J. Biol. Chem. 277, 45765–45769.

Chuang, C.C., Chen, C.Y., Yang, J.M., Lyu, P.C., Hwang, J.K., 2004. Relationship between protein structures and disulfide-bonding patterns. Proteins 53, 1–5.

Clarke, J., Hounslow, A.M., Bond, C.J., Fersht, A.R., Daggett, V., 2000. The effects of disulfide bonds on the denatured state of barnase. Protein Sci. 9, 2394–2404.

Creighton, T., 1993. In: Freeman, W.H. (Eds.), Proteins: Structures and Molecular Properties. 2nd Edition. New York.

Creighton, T., 1995. Disulfide-coupled protein folding pathways. Philos. Trans. R. Soc. London B 348, 5–10.

Eder, J., Wilmanns, M., 1992. Protein engineering of a disulfide bond in a beta/alpha-barrel protein. Biochemistry 31, 4437–4444.

Fariselli, P., Casadio, R., 2001. Prediction of disulfide connectivity in proteins. Bioinformatics 17, 957–964.

Fariselli, P., Riccobelli, P., Casadio, R., 1999. Role of evolutionary information in predicting the disulfide bonding state of cysteine in proteins. Proteins 36, 340–346.

Fariselli, P., Martelli, P.L., Casadio, R., 2002. A neural network-based method for predicting the disulfide connectivity in proteins. In: Damiani, E. et al. (Eds.), Knowledge Based Intelligent Information Engineering Systems and Allied Technologies (KES 2002), vol. 1, IOS Press, pp. 464–468.

Fiser, A., Simon, I., 2000. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 16, 251–256.

Fiser, A., Cserzo, M., Tudos, E., Simon, I., 1992. Different sequence environment of cysteines and half cystines in proteins: application to predict disulfide forming residues. FEBS Lett. 302, 117–120.

Frasconi, P., Passerini, A., Vullo, A., 2002. A two stage SVM architecture for predicting the disulfide bonding state of cysteines. In: Proceedings of IEEE Neural Network for signal processing conference. IEEE Press, New York.

Harrison, P.M., Sternberg, M.J.E., 1994. Analysis and classification of disulphide connectivity in proteins: the entropic effect of cross-linkage. J. Mol. Biol. 244, 448–463.

Hua, S.J., Sun, Z.R., 2001a. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J. Mol. Biol. 308, 397–407.

Hua, S.J., Sun, Z.R., 2001b. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728.

Huang, E.S., Samudrala, R., Ponder, J.W., 1999. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J. Mol. Biol. 290, 267–281.

Huang, Y., Li, Y.D., 2004. Prediction of protein subcellular locations using fuzzy *k*-NN method. Bioinformatics 20, 21–28.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA.

Joachims, T., 2002. Learning to classify text using support vector machine. Dissertation, Kluwer.

Kim, H., Park, H., 2003. Protein secondary prediction based on an improved support vector machines approach. Protein Eng. 16, 553–560.

Kremser, A., Rasched, I., 1994. The adsorption protein of filamentous phage fd: assignment of its disulphide bridges and identification of the domain incorporated in the coat. Biochemistry 33, 13954–13958.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London, pp. 322, 381.

Martelli, P.L., Fariselli, P., Malaguti, L., Casadio, R., 2002a. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng. 15, 951–953.

Martelli, P.L., Fariselli, P., Malaguti, L., Casadio, R., 2002b. Prediction of the disulfide bonding state of cysteines in proteins at 88% accuracy. Protein Sci. 11, 2735–2739.

Matsumura, M., Matthews, B., 1989. Control of enzyme activity by an engineered disulfide bond. Science 243, 792–794.

Matsumura, M., Matthews, B., 1991. Stabilization of functional proteins by introduction of multiple disulfide bonds. Method Enzymot. 202, 336–355.

Matthews, B.W., 1975. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biophys. Acta 405, 442–451.

Morris, H., Pucci, P., 1985. A new method for rapid assignment of s–s bridges in proteins. Biochem. Biophys. Res. Commun. 126, 1122–1128.

Mucchielli-Giorgi, M.H., Hazout, S., Tuffery, P., 2002. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins 46, 243–249.

Muskal, S.M., Holbrook, S.R., Kim, S.H., 1990. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng. 3, 667–672.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH—a hierarchic classification of protein domain structures. Structure 8, 1093–1108.

Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., Orengo, C.A., 2000. Assigning genomic sequences to CATH. Nucleic Acids Res. 1, 277–282.

Petersen, M.T., Jonson, P.H., Petersen, S.B., 1999. Amino acid neighbours and detailed conformational analysis of cysteines in proteins. Protein Eng. 12, 535–548.

Skolnick, J., Kolinski, A., Ortiz, A.R., 1997. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. 265, 217–241.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Van Vlijmen, H.W., Gupta, A., Narasimhan, L.S., Singh, J., 2004. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. J. Mol. Biol. 335, 1083–1092.

Vullo, A., Frasconi, P., 2004. Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 20, 653–659.

Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.

Wang, M.L., Li, W.J., Xu, W.B., 2003. Support vector machines for prediction of peptidyl prolyl cis/trans isomerization. J. Peptide Res. 63, 23–28.

Wedemeyer, W.J., Welkler, E., Narayan, M., Scheraga, H.A., 2000. Disulfide bonds and protein folding, Biochemistry 39, 4207–4216.

Welker, E., Narayan, M., Wedemeyer, W.J., Scheraga, H.A., 2001. Structural determinants of oxidative folding in proteins. Proc. Natl Acad. Sci. 98, 2312–2316.

Wittrup, K.D., 1995. Disulfide bond formation and eukaryotic secretory productivity. Curr. Opin. in Biotechnol. 6, 203–208.

Xue, J., Kalafatis, M., Silveira, J.R., Kung, C., Mann, K.G., 1994. Determination of the disulfide bridges in factor va heavy rain. Biochemistry 33, 13019–13116.

Zhang, Z., Sun, Z.R., Zhang, C.T., 2001. A new approach to predict the helix/strand content of globular Proteins. J. Theor. Biol. 208, 65–78.

Zhou, N.E., Kay, C.M., Hodges, R.S., 1993. Disulfide bond contribution to protein stability: positional effects of substitution in the hydrophobic core of the two-stranded alpha-helical coiled-coil. Biochemistry 32, 178–187.