Improving the Accuracy of Predicting Disulfide Connectivity by Feature Selection

LIN ZHU,1 JIE YANG,1 JIANG-NING SONG,2 KUO-CHEN CHOU,1,3 HONG-BIN SHEN1
1Department of Bioinformatics, Institute of Image Processing & Pattern Recognition, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
2Department of Bioinformatics, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
3Department of Bioinformatics, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130

Received 8 February 2009; Revised 26 August 2009; Accepted 31 August 2009
DOI 10.1002/jcc.21433
Published online 1 February 2010 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Disulfide bonds are primary covalent cross-links formed between two cysteine residues in the same or different protein polypeptide chains, which play important roles in the folding and stability of proteins. However, computational prediction of disulfide connectivity directly from protein primary sequences is challenging due to the nonlocal nature of disulfide bonds in the context of sequences, and the number of possible disulfide patterns grows exponentially when the number of cysteine residues increases. In the previous studies, disulfide connectivity prediction was usually performed in high-dimensional feature space, which can cause a variety of problems in statistical learning, such as the dimension disaster, overfitting, and feature redundancy. In this study, we propose an efficient feature selection technique for analyzing the importance of each feature component. On the basis of this approach, we selected the most important features for predicting the connectivity pattern of intra-chain disulfide bonds. Our results have shown that the high-dimensional features contain redundant information, and the prediction performance can be further improved when these high-dimensional features are reduced to a lower but more compact dimensional space. Our results also indicate that the global protein features contribute little to the formation and prediction of disulfide bonds, while the local sequential and structural information play important roles. All these findings provide important insights for structural studies of disulfide-rich proteins.


Key words: protein structure prediction; disulfide connectivity prediction; support vector machine; feature selection; Fisher score

Introduction

The three-dimensional (3D) structures can provide vital knowledge when addressing the problems in protein folding and functions. The hypothesis that protein sequence uniquely determines its structure inspires us to develop methods for the prediction of protein 3D structure directly from the primary amino acid sequence.1 Although the direct prediction of the 3D structure of a protein from its sequence based on the least free energy principle is scientifically sound, and some encouraging results in elucidating the handedness problems and packing arrangements in proteins have been obtained,1,2 it is far from successful yet for predicting its 3D structure owing to the notorious local minimum problem, except for some very special cases or by utilizing some additional information from experiments.3,4 Actually, it is even not successful yet for simply predicting the folding pattern of a query protein based on its sequence alone. As it is very difficult to directly predict the whole 3D structure of a protein from its primary sequence, computational prediction of various structural characteristics have attracted much attention,5,6 such as residue–residue contact map,7 disordered regions,8,9 residue solvent accessibility,7,10 helix–helix contacts,11 protein secondary structures,12–17 etc. All these findings will provide valuable insights into protein 3D structural studies.

Additional Supporting Information may be found in the online version of this article.

Correspondence to: H. B. Shen; e-mail: hbshen@sjtu.edu.cn
Contract/grant sponsor: National Natural Science Foundation of China; contract/grant numbers: 60704047, 60805001
Contract/grant sponsor: Science and Technology Commission of Shanghai Municipality; contract/grant numbers: 08ZR141060, 08JC1410600
Contract/grant sponsor: Innovation Program of Shanghai Municipal Education Commission; contract/grant number: 10ZZ17

© 2010 Wiley Periodicals, Inc.
One of the most important structural characteristics in protein structures is the disulfide bridges formed by the cysteine–cysteine residue pairs, which play important roles in stabilizing protein 3D structure. For example, a disulfide bond can still stabilize the secondary structures in its vicinity even when the water molecules attack amide–amide hydrogen bonds and break up secondary structure elements. Although disulfide bridges in a protein can be determined by various experimental techniques, it is both very expensive and time consuming. Especially in the postgenomic era, with an avalanche of newly emerging protein sequences, it is highly desired to develop computationally efficient methods that are capable of accurately predicting disulfide connectivity of any protein, given its amino acid sequence only. Much work has been done in this regard, and some of them provided detailed reviews of computational approaches for disulfide-bond determination. In most of these existing statistic methods, disulfide connectivity patterns are generally encoded by the global and local features of protein sequences, where the former include protein length, amino acid composition, protein molecular weight, while the latter are represented by the local sequential and structural characteristics, which can be further extracted from the sequence alignment profile position-specific score matrix (PSSM) and the predicted secondary structure information by the sliding window technique. Although combining both the global and local feature information has been demonstrated to be useful for correctly predicting disulfide bridges, this brings up another challenging issue to tackle, that is, disulfide connectivity patterns are often represented by high-dimensional vectors.

In statistical learning theory, previous studies have shown that conducting machine learning in high-dimensional space will readily result to the overfitting problem. We are also interested in investigating whether all the components in the high-dimensional vectors are necessary for predicting disulfide connectivity, and which types of sequence features are the most important in prediction, with respect to the global or the local sequence environment. In a recent study, Lu et al. proposed a method using support vector machines (SVMs) in combination with the genetic algorithm optimized for predicting the disulfide bonding states of cysteine pairs. As a result, they obtained an accuracy of 70% for predicting disulfide connectivity pattern. The genetic algorithm they employed efficiently removed noise and the irrelevant features. However, their method cannot give us more domain-specific knowledge of the optimal features. On the other hand, feature selection techniques have attracted much attention in other bioinformatics problems, for example, Kernytsky and Rost developed a framework by using genetic algorithm to select the most predictive protein features and achieved better results by applying it to the prediction task of enzymatic activity.

In this study, we have proposed an efficient feature selection technique for predicting disulfide connectivity, based on which we find that the global protein features contribute little to the formation and prediction of the disulfide bridges, and that the high-dimensional feature vector contains much redundant information. It is observed through this study that the prediction accuracy can be further improved when the high-dimensional vector is reduced to a lower but more compact feature space.

### Materials and Methods

#### Dataset

The benchmark dataset constructed in ref. 24 was used as the testing dataset of this study, which was prepared according to the following: (1) the proteins were from the release of Swiss-Prot database version 39 at www.ebi.ac.uk/swissprot/; (2) only protein sequences containing intra-chain disulfide bonds that were experimentally verified were included, whereas the inter-chain disulfide bonds were not considered and discarded; (3) protein sequences containing at least 2 and at most 5 disulfide bonds were selected; (4) to avoid the bias, a redundancy cutoff was operated to exclude the sequences which have ≥30% pairwise sequence identity to any other in the dataset. Finally, there are 446 proteins in the current dataset, which can be formulated as follows:

\[
S = S_2 \cup S_3 \cup S_4 \cup S_5,
\]

where \(S\) represents the entire benchmark dataset; \(S_2\) represents the subset containing two disulfide bonds only; \(S_3\), three disulfide bonds only; \(S_4\), four disulfide bonds only; \(S_5\), five disulfide bonds only; and \(\cup\) is the symbol for union in the set theory. The number of protein sequences in each subset is given in Table 1. These 446 protein sequences are provided in the Supporting Information.

#### Disulfide Connectivity Pattern

Suppose a protein has \(M\) disulfide bridges, then the number of possible cysteine pairs \(N\) is:

\[
N = M \times (2M - 1).
\]

For instance, if a protein has two disulfide bridges, then it will have at least four disulfide-bonding cysteine residues, denoted as \(c_1\), \(c_2\), \(c_3\), and \(c_4\), respectively. This will yield six possible disulfide-bonding cysteine pairs: \(c_1-c_2\), \(c_1-c_3\), \(c_1-c_4\), \(c_2-c_3\), \(c_2-c_4\), and \(c_3-c_4\).

For the \(N\) possible cysteine pairs, the number of possible disulfide connectivity patterns in a protein can thus be calculated as:

\[
P = \frac{(2M)!}{M!2^M} = \prod_{i=1}^{M} (2i - 1).
\]

<table>
<thead>
<tr>
<th>Subset</th>
<th>Number of proteins</th>
<th>Number of cysteine pairs</th>
<th>Number of disulfide bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S_2)</td>
<td>156</td>
<td>936</td>
<td>312</td>
</tr>
<tr>
<td>(S_3)</td>
<td>146</td>
<td>2190</td>
<td>438</td>
</tr>
<tr>
<td>(S_4)</td>
<td>99</td>
<td>2772</td>
<td>396</td>
</tr>
<tr>
<td>(S_5)</td>
<td>45</td>
<td>2025</td>
<td>225</td>
</tr>
<tr>
<td>Total</td>
<td>446</td>
<td>7923</td>
<td>1371</td>
</tr>
</tbody>
</table>

Table 1. The Benchmark Dataset \(S\) with Detailed Information about the Numbers of Protein Chains According to the Categorization Based on the Disulfide Bridge Numbers.
From Eq. (3), we can see that we will have three possible disulfide connectivity patterns for a protein when \( M = 2 \), i.e., “c1–c2, c3–c4,” “c1–c3, c2–c4,” and “c1–c4, c2–c3.” As a result, the number of possible disulfide connectivity patterns \( P \) increases exponentially with the increasing of \( M \), which is 10,395 when \( M = 6 \) according to Eq. (3).

### Feature Representation

In previous studies, disulfide connectivity of cysteine–cysteine pair was usually represented by both global and local features of proteins.\(^{22,36} \) The global features include the amino acid composition, protein molecular weight, and protein sequence length, while the local features comprise of the cysteine–cysteine coupling effect, cysteine–cysteine local secondary structural environment, cysteine separation distance, and cysteine ordering. The amino acid composition is represented by a 20-D vector, while the protein molecular weight and the protein sequence length are two scalars. The cysteine–cysteine coupling effect is generally extracted from evolutionary profile PSSM (generated by PSI-BLAST program) with a local sliding window, where the PSSM can be presented by a matrix of \( L \times 20 \) for the protein of length \( L \). When the local window length is 13, we will get a 520-D vector. The cysteine–cysteine local secondary structural information can be extracted from the predicted secondary structure by the PSIPRED program, which can be represented by a matrix of \( L \times 3 \) and the three columns represent the three secondary structural types, i.e., \( \alpha \)-helix, \( \beta \)-sheet, and the coiled-coil. Then, with a sliding window of length 13, we will have a 78-D vector for this kind of feature. Cysteine separation distance is used to measure the sequential distance between two cysteine residues that form a disulfide bond, which is a scalar. Cysteine ordering represents the sequential order difference between each cysteine–cysteine pair, which is a 2-D vector. By incorporating all the global and local features together, we can obtain a 623-D vector to represent a cysteine–cysteine pair. For more details about how to generate all these features refer refs. 22, 36.

### Feature Selection

When dealing with high-dimensional bioinformatics problems, it is particularly important to investigate and compare the contributions of different features. By doing so, we can get more useful insights into the biological mechanisms, which will allow us to better design the corresponding prediction tools. Hence, to know how many and which features are closely correlated with the variance score, Laplacian score and the Fisher score methods are used: variance score, Laplacian score and the Fisher score. Among them, the variance score and the Laplacian score methods are unsupervised, whereas the Fisher score is supervised, which means the class labels are taken into account. More specifically, in the case of the filter-based models, we computed the feature scores and subsequently ranked them in two steps. First, the scores of all features were computed for each fold in the benchmark dataset. Second, the average of the feature scores over the four folds was calculated, which was further used to rank the original set of 623 features. This cross-validation procedure allows us to maintain the consistency with the experimental evaluation of disulfide connectivity prediction.

Variance score might be the simplest unsupervised approach for feature selection. It uses the variance along a dimension to reflect its representative power and those features with the maximum variance are selected. Given a set of \( N \) possible cysteine pairs,

\[
D = \{X_1, X_2, X_3, \ldots, X_i, \ldots, X_N\},
\]

where \( X_i \) is the \( i \)th cysteine pair [cf. Eq. (2)] and can be represented by a 623-D vector as

\[
X_i = [f_{i,1}, f_{i,2}, f_{i,3}, \ldots, f_{i,m}, \ldots, f_{i,623}]^T,
\]

where \( T \) is a transpose operator. Then, the variance of the \( n \)th feature can be computed as follows:

\[
V_m = \frac{1}{N} \sum_{i=1}^{N} (f_{i,m} - \bar{f}_m)^2,
\]

where \( \bar{f}_m \) is the mean value of the \( n \)th feature. The larger the \( V_m \) value, the more important the \( n \)th feature.

The Laplacian score is a more sophisticated criterion than the variance, which is computed to reflect the locality preserving power of a corresponding feature. It is based on the observation that two data points are more likely to be related to the same class if they are close to each other. The Laplacian score \( L_m \) of the \( m \)th feature can be computed as follows:

\[
L_m = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (f_{i,m} - f_{j,m})^2 S_{ij}}{\sum_{i=1}^{N} (f_{i,m} - \bar{f}_m)^2 D_{ii}},
\]

where \( S_{ij} \) is the similarity matrix and is defined by the neighborhood relationship between samples:

\[
S_{ij} = \begin{cases} 
\exp\left(-\frac{\|X_i - X_j\|}{t}\right) & \text{if } X_i \text{ and } X_j \text{ are neighbors,} \\
0 & \text{otherwise},
\end{cases}
\]

where \( \|\cdot\| \) is the norm operator, \( t \) is a constant and is set to one for simplicity, “\( X_i \) and \( X_j \) are neighbors” means that based on
the Euclidean distance measure, either \( X_i \) is one of the \( k \) nearest neighbors of \( X_j \), or vice versa. \( k \) is set to five in this study. \( D \) in Eq. (7) is a diagonal matrix defined based on \( S_{ii} \):

\[
D_{ij} = \begin{cases} 
\sum_{p=1}^{N} S_{ip}, & \text{if } i=j \\
0, & \text{otherwise}.
\end{cases}
\]

(9)

In contrast to the variance score and Laplacian score defined by eqs. (6) and (7), respectively, Fisher score is a supervised measure with class labels and can be used to seek features that are efficient for the discrimination of different classes. It assigns the highest score to the feature, which requires that the data points of different classes are far from each other on one hand, while data points of the same class need to be close to each other on the other hand. Suppose that there are \( C \) classes in the dataset, then the Fisher score of the \( m \)th feature is computed as:

\[
H_m = \frac{\sum_{i=1}^{C} N_i (\bar{f}_m^i - \bar{f}_m)^2}{\sum_{i=1}^{C} N_i (\sigma_m^i)^2},
\]

(10)

where \( N_i \) is the number of samples in the \( i \)th class, \( \bar{f}_m^i \) and \( \sigma_m^i \) are the mean value and the variance of the \( m \)th feature in the \( i \)th class, respectively, and \( \bar{f}_m \) is the same as defined in Eq. (6).

\section*{Support Vector Regression}

Support vector machine is a popular machine learning algorithm based on the structural risk minimization for pattern classification.\(^{45}\) Support vector regression (SVR) is a regression model based on SVM by using kernel function, the \( \varepsilon \)-insensitive loss function and the regularization parameter \( \xi \). Kernel functions can take different forms, such as linear kernel function, polynomial kernel function, radial basis kernel function, sigmoid kernel function, or even a user-defined kernel function.\(^{45}\) In this work, we used the radial basis function, which has been shown to be an optimal choice in many previous studies.\(^{46}\)

\[
K(X_i, X) = \exp(-\gamma \|X_i - X\|^2),
\]

(11)

where \( \gamma \) is a parameter needed to be optimized.

For the \( N \) possible cysteine pairs [cf. Eq. (2)] in a protein, we can calculate the possibility of a disulfide connectivity as:

\[
\text{SVR} \triangleright \bar{X}_i = \Lambda_i \quad (i = 1, 2, \ldots, N),
\]

(12)

where \( \triangleright \) represents an action operator, \( \bar{X}_i \) is the \( i \)th cysteine pair with reduced dimensions [cf. eqs. (4–10)] and \( \Lambda_i \) is the corresponding score of this cysteine pair to form disulfide connectivity obtained from the SVR models. Based on Eq. (12), we can compute the score of a disulfide connectivity pattern as defined in Eq. (3),

\[
Y_j = \Lambda_1 + \Lambda_2 + \cdots + \Lambda_R \quad (j = 1, 2, \ldots, P),
\]

(13)

where \( Y_j \) is the score of the \( j \)th disulfide connectivity pattern, \( \Lambda_j \) is the possibility of the \( j \)th cysteine pair contained in the \( j \)th disulfide connectivity pattern [cf. Eq. (3)]. Thus, the disulfide connectivity pattern that has the highest score in Eq. (13) will be predicted as the final result, i.e.,

\[
\mu = \arg \max_j \{Y_j\},
\]

(14)

where \( \mu \) is the argument of \( j \) that maximizes \( Y_j \).

Compared with the previous pattern-wise prediction methods, our approach of disulfide connectivity pattern prediction addressed above can significantly reduce the imbalance problem, which results from the high ratio of positive/negative training samples. The whole prediction processes can be generally classified into three steps: first, the global and local features were extracted from primary amino acid sequences; second, the variance, Laplacian and Fisher scores were computed for each of the 623 features; and third, all the features were ranked based on their scores, and the most important features were selected to make a further SVR prediction. To provide an intuitive picture, a flowchart is given in Figure 1 to show how to predict disulfide connectivity from the primary amino acid sequence based on this feature selection method.

\section*{Results and Discussions}

In this study, we used four-fold cross-validation test and two widely used criteria \( Q_c \) and \( Q_p \) to evaluate the prediction
The proposed feature selection method, \( Q_c \) and \( Q_p \) are defined as follows:

\[
Q_c = \frac{N_c}{N},
\]

\[
Q_p = \frac{N_p}{T},
\]

where \( N_c \) is the number of disulfide bridges which are correctly predicted, and \( N \) is the total number of disulfide bridges in the test dataset.

As described above, two parameters of SVR should be optimized, i.e., the kernel parameter \( \gamma \) and the regularization parameter \( \zeta \). The parameterization of SVR was performed through a grid search over \( \gamma \) and \( \zeta \) based on four-fold cross-validation on the benchmark dataset \( S \). We considered \( \gamma \) and \( \zeta \) values drawn from \( \{0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1\} \times \{1, 2, 3, 4, 5, 6, 7\} \) grid and then we tried \( 9 \times 7 = 63 \) different combinations. The optimized parameters are listed in Table 2, which are further used in the following experiments.

We compared the performances between the three different feature selection methods described above. The results are listed in Table 3. As can be seen, the Fisher score \( H_m \) outperforms the other two feature selection methods in most cases. This is because the variance and Laplacian score methods calculate the feature scores without using the knowledge of the class labels of the training data, while the Fisher score takes the advantage of this additional information. As shown in Table 3, when the most important 150 features from the original 623 features are selected using the Fisher score \( H_m \), the prediction performance is the best. In addition, we also performed calculations based on the support vector classification (SVC) probability estimation method and the results were given in Stable 4 of the Supporting Information. For comparison, the results from SVR were also given in Table 4. As can be seen from Stable 4, the prediction results based on SVR are better than those of SVC.

For the sake of comparison, success rates for predicting disulfide connectivity of our approach, as well as others are listed in Table 4. As can be seen, after the feature selection, our method achieved an accuracy of \( Q_p = 76.0\% \) and \( Q_c = 80.3\% \), which respectively increased by 1.6\% and 2.4\% in comparison with the state-of-art method SS_SVR.\(^{37}\) Moreover, our method also outperformed another exhaustive feature selection method GASVM (support vector machine with the genetic algorithm

<table>
<thead>
<tr>
<th>Feature number</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
<th>250</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel parameter ( \gamma )</td>
<td>0.7</td>
<td>0.5</td>
<td>0.3</td>
<td>0.07</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Regularization parameter ( \zeta )</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

\(^{a}\)The features are selected according to the importance scores obtained from the Fisher feature selection.

\(^{b}\)SVR, support vector regression.

<table>
<thead>
<tr>
<th>Feature number</th>
<th>( Q_p )</th>
<th>( Q_c )</th>
<th>( Q_p )</th>
<th>( Q_c )</th>
<th>( Q_p )</th>
<th>( Q_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.7129</td>
<td>0.7541</td>
<td>0.7130</td>
<td>0.7519</td>
<td>0.7131</td>
<td>0.7572</td>
</tr>
<tr>
<td>100</td>
<td>0.7264</td>
<td>0.7825</td>
<td>0.7398</td>
<td>0.7753</td>
<td>0.7580</td>
<td>0.7827</td>
</tr>
<tr>
<td>150</td>
<td>0.7377</td>
<td>0.7818</td>
<td>0.7510</td>
<td>0.7979</td>
<td>0.7400</td>
<td>0.7856</td>
</tr>
<tr>
<td>200</td>
<td>0.7378</td>
<td>0.7783</td>
<td>0.7555</td>
<td>0.7898</td>
<td>0.7400</td>
<td>0.7856</td>
</tr>
<tr>
<td>250</td>
<td>0.7444</td>
<td>0.7914</td>
<td>0.7510</td>
<td>0.7942</td>
<td>0.7511</td>
<td>0.8009</td>
</tr>
<tr>
<td>300</td>
<td>0.7355</td>
<td>0.7811</td>
<td>0.7444</td>
<td>0.7834</td>
<td>0.7557</td>
<td>0.8029</td>
</tr>
</tbody>
</table>

\(^{a}\)The features are selected according to the importance scores obtained from the feature selection.
optimization) which is optimized based on the genetic algorithm, where the $Q_p$ and $Q_c$ measures of our method improved by 2.1% and 1.1%, respectively. Taken together, our method has outperformed all the other previous methods in terms of both $Q_p$ and $Q_c$ measures and the improvement of prediction performance is remarkable, considering the fact that our approach only used 150 selected features out of the total 623 features. Another advantage of our approach is that it considerably reduced the computational cost by three folds.

In addition to the two rigorous measures $Q_p$ and $Q_c$, we also calculated the $p$-value by chi-square test, which is often used to illustrate the statistical significance of the prediction. The detailed results were given in Stables 1–3 of the Supporting Information. From Stable 2 and Stable 3, we can find that the significant predictions with $p < 0.001$ were observed for all tested cases before and after feature selection procedure, indicating feature selection process is indeed very useful when handling the complicated high-dimensional nonlinear biological prediction problem.

The feature selection methods of this study allow us to analyze the relative importance of different features and further select the most important ones for making final predictions of disulfide connectivity pattern. Now, let us go further into the 150 features selected. Table 5 gives the detailed information about the 150 features selected from the original 623 features by using the Fisher score $H_m$. We find that global protein features, i.e., amino acid composition, protein weight, and protein length, contribute little to the prediction performance and as a result most of them were not selected in the final feature set. More specifically, protein weight and protein length features are completely discarded. Table 5 also indicates that the formation of disulfide bridges is primarily dependent on the local features, such as the sequential distance between two cysteine residues, the cysteine orders, the local secondary structures, and the local evolutionary information. These findings are very helpful for our better understanding of the mechanisms of disulfide connectivity formation and provide an opportunity of developing more efficient computational tools in the future.

The top 150 features out of the original 623-D vectors is provided in Supporting Information, where the 150 selected features by Fisher score are organized by feature types (see “The Feature Representation” section for more details). The overall importance order of the selected features is as follows: distance of cysteine > cysteine ordering > position-specific scoring matrix > predicted secondary structure > amino acid composition. As a well-conserved and stereo-specific secondary structural element of a protein, disulfide bridge has its preferred disulfide signatures with respect to its formation in the context of specific sequences. This preference can be encoded and reflected by two primary descriptors: the distance of disulfide-bonded cysteines and their ordering. On the other hand, the secondary structure conformation and sequence contexts assumed by disulfide-bonded cysteines and nondisulfide-bonded cysteines have remarkable preferences, which can be represented and encoded using predicted secondary structure, PSSM and amino acid compositions. Indeed, the improved prediction performance

---

**Table 5. The Number of Selected Features for Different Feature Types by Fisher Score.**

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Number of feature components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before feature selection</td>
</tr>
<tr>
<td>Amino acid composition$^a$</td>
<td>20</td>
</tr>
<tr>
<td>Protein weight$^a$</td>
<td>1</td>
</tr>
<tr>
<td>Protein length$^a$</td>
<td>1</td>
</tr>
<tr>
<td>Position-specific scoring matrix$^b$</td>
<td>520</td>
</tr>
<tr>
<td>Predicted secondary structure$^b$</td>
<td>78</td>
</tr>
<tr>
<td>Distance of cysteine$^b$</td>
<td>1</td>
</tr>
<tr>
<td>Cysteine order$^b$</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>623</td>
</tr>
</tbody>
</table>

$^a$Features represent protein global information.

$^b$Features represent local structural or sequential information.

---

**Figure 2.** Distributions of the selected PSSM features according to their corresponding Fisher scores $H_m$: (a) the PSSM profile for the first cysteine residue; (b) the PSSM profile for the second cysteine residue. The shadow of individual cells (features) corresponds to their absolute average (over four folds in SP39 dataset) Fisher score $H_m$: black for top 1–50, dark gray for top 51–100, light gray for top 101–150, and white for the features that were not selected.
achieved in this study demonstrates the significant contribution of the efficiently selected features to accurate prediction of disulfide connectivity. For these 150 selected features (c.f. Supporting Information), only 121 out of the total 520 original PSSM components were selected, which are displayed in Figure 2. These findings highlight that although the prediction performance could be improved by incorporating evolutionary information, the PSSM profile contains much redundant information useless for prediction. By applying the proposed feature selection techniques, we have successfully selected the most important features for the improvement of the prediction performance of disulfide connectivity.

Conclusions

Disulfide bonds, formed by the cysteine pairs, play important roles in stabilizing the protein structures. In this study, we have proposed three novel feature selection techniques for improving the prediction performance of disulfide connectivity patterns from the primary amino acid sequences. By analyzing the importance of the features, we successfully extracted the key information from the high-dimensional feature space and reduced the original high-dimensional vectors to a lower-dimensional but more representative subset. Our results have demonstrated that by applying these efficient feature selections, the prediction accuracy can be significantly improved. As a result, the Qm measure, for the first time, achieved the success rate of over 80% accuracy, with a much lower computation cost. We also find that it is the local features that dominate the formation and prediction of disulfide bridges, while global features contribute little. We expect that this study will provide valuable insights into structural studies of proteins. Our work provides a further opportunity to develop more efficient tools for solving the difficult problem of disulfide connectivity prediction. The feature selection techniques proposed in this study are specific to a given classifier: some classifiers are successful when being applied directly to high-dimensional data as they perform "internal" feature selection, while others require the removal of closely correlated features. Therefore, how to choose an effective feature selection method remains to be an open issue. We will endeavor to address this issue in our future studies, which will be expected to be valuable for handling other complicated life systems.

Acknowledgments

We gratefully thank three anonymous reviewers for their helpful and constructive comments that are very helpful for strengthening the presentation of this paper. This work was sponsored by Shanghai Pujiang Program and was also supported by the grants from the National Health and Medical Research Council of Australia (NHMRC), Australian Research Council (ARC), and Japan Society for the Promotion of Science (JSPS).

References

33. Wang, Y.; Miller, D. J.; Clarke, R. Br J Cancer 2008, 98, 1023.
35. Kernytsky, A.; Rost, B. Proteins Struct Funct Bioinf 2009, 75, 75.