Inter- and Intra-Chain Disulfide Bond Prediction Based on Optimal Feature Selection

Shen Niu[2,3,§], Tao Huang[2,3,§], Kai-Yan Feng[3], Zhisong He[5], Weiren Cui[5], Lei Gu[6], Haipeng Li[5,*], Yu-Dong Cai[1,4,*] and Yixue Li[2,3,*]

[1]Institute of Systems Biology, Shanghai University, Shanghai, P. R. China; [2]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China; [3]Shanghai Center for Bioinformation Technology, Shanghai, P. R. China; [4]Centre for Computational Systems Biology, Fudan University, Shanghai, P. R. China; [5]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China; [6]Division of Theoretical Bioinformatics (BO80), German Cancer Research Center, Heidelberg, Germany

Abstract: Protein disulfide bond is formed during post-translational modifications, and has been implicated in various physiological and pathological processes. Proper localization of disulfide bonds also facilitates the prediction of protein three-dimensional (3D) structure. However, it is both time-consuming and labor-intensive using conventional experimental approaches to determine disulfide bonds, especially for large-scale data sets. Since there are also some limitations for disulfide bond prediction based on 3D structure features, developing sequence-based, convenient and fast-speed computational methods for both inter- and intra-chain disulfide bond prediction is necessary. In this study, we developed a computational method for both types of disulfide bond prediction based on maximum relevance and minimum redundancy (mRMR) method followed by incremental feature selection (IFS), with nearest neighbor algorithm as its prediction model. Features of sequence conservation, residual disorder, and amino acid factor are used for inter-chain disulfide bond prediction. And in addition to these features, sequential distance between a pair of cysteines is also used for intra-chain disulfide bond prediction. Our approach achieves a prediction accuracy of 0.8702 for inter-chain disulfide bond prediction using 128 features and 0.9219 for intra-chain disulfide bond prediction using 261 features. Analysis of optimal feature set indicated key features and key sites for the disulfide bond formation. Interestingly, comparison of top features between inter- and intra-chain disulfide bonds revealed the similarities and differences of the mechanisms of forming these two types of disulfide bonds, which might help understand more of the mechanisms and provide clues to further experimental studies in this research field.

Keywords: Disulfide bond, inter-chain, intra-chain, incremental feature selection, maximum relevance minimum redundancy, nearest neighbor algorithm.

1. INTRODUCTION

Protein disulfide bond is formed by the oxidation of thiol (-SH) groups between inter- or intra-chain cysteine residues, during post-translational modifications. Disulfide bonds are common to many proteins and relate closely to protein structures since they can impose geometrical constraints on the protein backbones [1-2]. Correct localization of disulfide bonds can greatly limit the search space of possible protein conformations [3-4] and thus facilitate the prediction of protein 3D structure. Disulfide bonds have been demonstrated to be involved in various physiological functions, such as hemostasis [5], cell death [6], G-protein-receptors [7] and growth factors [8]. Disulfide bonds have also been implicated in various pathological processes, such as tumor immunity [9] and neurodegenerative diseases [6].

However, determining cysteine disulfide bonds by conventional experimental approaches such as mass spectrometry method [10-11], NMR method [12] and radiation experiment [13], may be time-consuming and labor-intensive, especially for large scale data sets. Yet, it is much more convenient and efficient to predict cysteine disulfide bonds using in-silico algorithms at the proteome level. Some computational methods exist in the literature for the prediction of disulfide bonds. For instances, Lin Zhu *et al.* [14] applied both global and local features of proteins to predict disulfide bonds, using support vector regression model and based on some newly developed feature selection methods. The disulfide bond prediction accuracy of their method achieved 80.3% [14]. Rotem Rubinstein *et al.* [15] analyzed correlated mutation patterns based on multiple sequence alignments to predict disulfide bonds. The prediction accuracies of their method for proteins with two, three and four disulfide bonds are 73, 69 and 61% respectively [15]. The limitation of the method is that it cannot unambiguously predict all disulfide bonds of a protein if more than one fully conserved disulfide bond exists. Hsuan-Hung Lin *et al.* developed a web server

*Address correspondence to these authors at the CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P.R. China; Tel: 86-21-54920460; Fax: 86-21-54920451; E-mail: lihaipeng@picb.ac.cn

[§]These authors contributed equally to this work

1875-5305/13 $58.00+.00

for disulfide bond prediction using the coordination of the Cα of each amino acid in a protein as the feature [16]. Their method performed better than former methods, but it is not suitable for protein sequences containing cysteines located in the metal binding sites. F. Ferre *et al.* developed DiANNA for classifying cysteines into reduced, half-cysteine or ligand-bound state using a support vector machine with spectrum kernel [17]. Marc Vincent *et al.* developed methods for predicting disulfide bridges using two decomposition kernels to measure the similarity between protein sequences according to the amino acid environments around cysteines [18]. Jiangning Song *et al.* developed a method to predict disulfide connectivity patterns from protein primary sequence, based on a support vector regression (SVR) approach using multiple sequence features and secondary structures [19]. The above-mentioned methods primarily predict the intra-chain disulfide bonds [14].

In this work, we developed a computational method based on nearest neighbor algorithm (NNA) by integrating it with a feature selection method (IFS coupled with mRMR) for the prediction of both inter- and intra-chain disulfide bonds. Sequence conservation, residual disorder and amino acid factor features were used for inter-chain disulfide bond prediction. And in addition to these features, sequence distance between a pair of cysteines was used for intra-chain disulfide bond prediction. Our approach achieved a prediction accuracy of 0.8702 for inter-chain disulfide bonds using 128 features and 0.9219 for intra-chain disulfide bonds using 261 features. Further analysis and comparison of the optimal feature sets (especially the top features) for inter- and intra-chain disulfide bonds may provide clues to understand the disulfide bond formation mechanisms and future studies in this research field.

2. MATERIALS AND METHODS

2.1. Training and Independent Test Data Sets

2.1.1. Training Data Sets

We downloaded 2930 protein sequences containing disulfide bonds from SysPTM (version 1.1) [20]. A segment of 9 consecutive residues (including cysteine itself in the center, 4 residues upstream and 4 residues downstream) is considered as the mini-environment of each cysteine. For inter-chain disulfide bonds, we extracted all 9-residue peptides resulting in totally 26858 cysteine segments. These 26858 segments consist of 374 segments with the center cysteines forming inter-chain disulfide bonds and 26484 with no inter-chain disulfide bonds. We took all 374 segments with the center cysteines forming inter-chain disulfide bonds as positive samples and took 1870 segments (5 folds 374*5=1870 of the positive samples) from the 26484 negative segments as negative samples. From which we excluded 13 samples whose features cannot be calculated in the study, resulting in totally 2227 samples consisting of 370 positive samples and 1857 negative samples. The training data set of inter-chain disulfide bond prediction was given in DataSet S1.

For intra-chain disulfide bonds, we calculated all cysteine pairs within each sequence, resulting in totally 770702 cysteine pairs. We then took 8457 cysteine pairs with known intra-chain disulfide bonds as positive samples and the remaining 762245 cysteine pairs as the candidates of negative samples. Because the sequence distance between paired cysteines in 94.05% of the positive samples is less than 100 residues, we selected 42285 (5 folds 8457*5=42285 of the positive samples) from the remaining cysteine pairs with distances less than 100 residues as negative samples. By excluding cysteine pairs, either of which cannot form 9 consecutive residues, we totally got 46988 samples including 7089 positive samples and 39899 negative samples. The training data set for intra-chain disulfide bond prediction was given in DataSet S2.

2.1.2. Independent Test Data Sets

We downloaded 3217 protein sequences containing experimentally validated disulfide bonds from UniProt (version 2010_06) [21-22]. We removed 2898 protein sequences that were already used in our training data set and protein sequences with less than 50 residues, resulting in 260 protein sequences.

For inter-chain disulfide bonds, we extracted all 9-residue peptide segment including cysteine itself and 4 residues at both the directions of C- and N-terminals, resulting in totally 2750 sample peptides, including 54 positive sample peptides and 2696 negative sample peptides. The independent test data set for inter-chain disulfide bond prediction was given in DataSet S3.

For intra-chain disulfide bonds, there are totally 37948 possible cysteine pairs, including 747 intra-chain disulfide bond pairs and 37201 non-disulfide bond pairs. Within the 37201 non-disulfide bond pairs, there are 10911 cysteine pairs with distances less than 100 residues. We then excluded cysteine pairs, either of which having peptide segment less than 9 residues, from the 747 disulfide bond pairs and 10911 non-disulfide bond pairs, resulting in totally 11213 sample cysteine pairs including 695 positive pairs and 10518 negative pairs. The independent test data set for intra-chain disulfide bond prediction was given in DataSet S4.

2.2. Feature Construction

2.2.1. PSSM Conservation Score Features

Evolutionary conservation is an important aspect in biological functions and plays important roles in post-translational modifications, such as tyrosine sulfation [23] and disulfide bond formation [14]. In our study, we used position specific iterative BLAST (PSI BLAST) [24] to quantify the conservation probabilities of each amino acid against 20 amino acids, yielding a 20-dimensional vector. The 20-dimensional vectors for all residues in a given protein sequence formed a matrix called the position specific scoring matrix (PSSM). Residues that are more important to biological functions are usually more conserved through the cycles of PSI BLAST. In this study, PSSM conservation score was used as the conservation features of each amino acid in a given protein sequence.

2.2.2. Amino Acid Factor Features

The diversity and specificity of protein structures and functions are largely attributed to the different compositions of amino acids, which have their own intrinsic physicochemical properties. The effect of amino acid properties on post-translational modification has been demonstrated by previous studies [14, 23].

AAIndex [25] is a database maintaining various amino acid physicochemical and biochemical properties. Atchley *et al.* [26] performed multivariate statistical analyses on AAIndex. They summarized and transformed AAIndex to five highly compact numeric patterns reflecting polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. We used these five numerical pattern scores (denoted as "amino acid factors") to represent the respective properties of each amino acid in our research.

2.2.3. Disorder Score Features

Protein disordered region is a protein segment that lacks 3-D structures under physiological conditions. Previous studies showed that these regions always contain PTM sites and sorting signals, and play important roles in regulating protein structures and functions [27-29].

In our study, VSL2 [30], which can accurately predict both long and short disordered regions in proteins, was used to quantify each of the amino acid disorder status in the protein sequence by calculating the disorder score. The disorder scores of cysteine site and 8 flanking sites at both C- and N-terminal are calculated as features in the study.

2.2.4. The Feature Space

2.2.4.1. For Inter-chain Disulfide Bonds

For cysteine site, 20 PSSM conservation scores and 1 disorder score, totally 21 features were used. For each of the 8 surrounding residues, 20 PSSM conservation scores, 5 amino acid factors and 1 disorder score, totally 26 features were used. Overall, each sample was encoded by $26\times 8+21=229$ features.

2.2.4.2. For intra-chain Disulfide Bonds

We calculated the absolute values of the sum and difference of the PSSM conservations, amino acid factors and disorder scores between each pair of cysteine sites, resulting in totally 458 features. The sequence distance between the paired cysteine sites was also included as a feature. So the overall feature space contains 458+1=459 features.

2.3. mRMR Method

To rank the features according to their importance, we used maximum relevance, minimum redundancy (mRMR) method [31], which could rank features based on the trade-off between maximum relevance to target and minimum redundancy to the already selected features. Features having a smaller index mean that they are more important features.

We used mutual information (MI) to quantify the relation between two vectors, which was defined as following:

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (1)$$

In equation (1), x and y denote vectors. $p(x)$ and $p(y)$ denote the marginal probabilistic densities. $p(x,y)$ denotes joint probabilistic density.

To quantify both relevance and redundancy, we defined Ω as the whole feature set, Ω_s as the already-selected feature set containing m features and Ω_t as the to-be-selected feature set containing n features. The relevance D between feature f in Ω_t and the target c can be calculated by:

$$D=I(f,c) \qquad (2)$$

The redundancy R between the feature f in Ω_t and all the features in Ω_s can be calculated by:

$$R=\frac{1}{m}\sum_{f_i\in\Omega_s} I(f,f_i) \qquad (3)$$

The mRMR function, which combined equation (2) and equation (3) and can be used to obtain the feature f_j in Ω_t with maximum relevance and minimum redundancy, was defined as following:

$$\max_{f_j\in\Omega_t}\left[I(f_j,c)-\frac{1}{m}\sum_{f_i\in\Omega_s} I(f_{j,}f_i)\right](j=1,2,...,n) \qquad (4)$$

Given a feature set with $\mathrm{N}\left(\mathrm{N}=\mathrm{m}+\mathrm{n}\right)$ features, the feature evaluation will be performed N rounds. After these evaluations, mRMR method will generate a feature set S :

$$S=\left\{f_1^{'},f_2^{'},...,f_h^{'},...,f_N^{'}\right\} \qquad (5)$$

In this feature set S , each feature has an index h, indicating at which round the feature is selected. A more important feature will be selected earlier and have a smaller index h.

2.4. Nearest Neighbor Algorithm

Nearest neighbor algorithm (NNA) was used as the prediction model of our method. NNA calculates similarities between the test sample and all the training samples. In our study, the distance between vector p_x and p_y is defined as following [32-33]:

$$D(p_x,p_y)=1-\frac{p_x\cdot p_y}{\|p_x\|\cdot\|p_y\|} \qquad (6)$$

In equation (6), $p_x\cdot p_y$ denotes the inner product of p_x and p_y . $\|p\|$ denotes the module of vector p . The smaller $D(p_x,p_y)$ is, the more similar p_x to p_y is.

In NNA, given a vector p_t and training set $P=\{p_1,p_2,...,p_n,...,p_N\}$, p_t will be designated to the same class of its nearest neighbor p_n in P , i.e. the vector having the smallest $D(p_n,p_t)$:

$$D(p_n,p_t)=\min\{D(p_1,p_t),D(p_2,p_t),...,D(p_z,p_t),...,D(p_N,p_t)\}(z\neq t) \qquad (7)$$

2.5. Jackknife Test

There are several methods to examine the accuracy of a statistical prediction method, such as the independent dataset test, sub-sampling (e.g., 5 or 10-fold cross-validation) test, and jackknife test [34]. Within the above three methods, the jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, as elucidated in [35-36] and demonstrated by Eqs.28-32 of [37]. Therefore, investigators had increasingly recognized and widely adopted the jackknife test to examine the power of various prediction methods [38-47]. In view of this, we also used the jackknife test to examine the predictive power of our computational method. In jackknife test, every sample is tested by the predictor trained with all the other samples. The prediction accuracies for the positive samples, negative samples and the overall samples were defined as following:

$$\begin{cases} accuracy\phi\, positive\, dataset = \dfrac{correctly\ predicted\ positive samples}{positive samples} \\ accuracy\phi\, negative\, dataset = \dfrac{correctly\ predicted\ negative samples}{negative samples} \\ overall\, accuracy = \dfrac{correctly\ predicted\ positive samples + correctly\ predicted\ negative samples}{positive samples + negative samples} \end{cases} \tag{8}$$

2.6. Incremental Feature Selection (IFS)

After ranking features by mRMR method based on their importance, we used incremental feature selection (IFS) to determine the optimal number of features.

An incremental feature selection is conducted for each of the independent predictor with the ranked features. Features in a set are added one by one from higher to lower rank. If one feature is added, a new feature set is obtained. Thus we get N feature sets where N is the number of features, and the i-th feature set is:

$$S_i = \{f_1, f_2, ..., f_i\} (1 \le i \le N)$$

Based on each of the N feature sets, an NNA predictor was constructed and tested using Jackknife cross-validation test. With N overall accuracy prediction rates calculated, we obtain an IFS table with one column being the index i and the other column being the overall accurate rate. S_{optimal} is the optimal feature set that achieves the highest overall accurate rate.

For intra-chain disulfide bond prediction, the ranked features were added ten by ten from higher to lower rank. So we get the feature sets containing 1, 11, 21, 31,…, 451 features respectively, producing totally 46 feature sets for the 459 features.

3. RESULTS AND DISCUSSION

3.1. mRMR Result

Using the mRMR program, we obtained the ranked mRMR list of 229 and 459 features for inter- and intra-chain disulfide bonds respectively. Within the lists, the smaller index of a feature indicates its more important roles in discriminating positive samples from negative ones. The mRMR lists were used in IFS procedure for further feature selection and analysis.

3.2. IFS Result

3.2.1. Inter-chain Disulfide Bonds

Based on the outputs of mRMR, we built 229 individual predictors for the 229 sub-feature sets to predict inter-chain disulfide bonds. We tested each of the 229 predictors and obtained the IFS result which can be found in Table **S1**. Fig. (**1A**) shows IFS curve plotted based on Table **S1**. The maximum accuracy is 0.8752 containing 207 features. To focus our analysis on a relatively smaller set of features, we selected the first feature set that achieves a predictive accuracy higher than 0.87 that is 0.8702 containing 128 features as the optimal feature set. The 128 optimal features were given in Table **S2**.

3.2.2. Intra-chain Disulfide Bonds

Based on the outputs of mRMR, we built 46 individual predictors for the 46 sub-feature sets to predict intra-chain disulfide bonds. We tested each of the 46 predictors and obtained the IFS result which can be found in Table **S3**. Fig. (**1B**) shows the IFS curve plotted based on the data in Table **S3**. The maximum accuracy is 0.9219 containing 261 features. These 261 features were considered as the optimal feature set of our classifier. The 261 optimal features were given in Table **S4**.

3.3. Optimal Feature Set Analysis

We investigated and compared the feature- and site-specific distribution of the 128 and 261 optimal features for inter- and intra-chain disulfide bond prediction respectively.

3.3.1. Inter-chain Disulfide Bonds

As shown in Fig. (**2A**), in the optimized 128 features, there were 26 amino acid factor features, 3 disorder score features and 99 PSSM conservation score features. This suggests that all three kinds of features contribute to the prediction of inter-protein disulfide bonds and conservation may play an irreplaceable role in inter-chain disulfide bond prediction.

Fig. (**2B**) demonstrates that the center site (site 5) and distal sites (site 1, 2 and 9) have the greatest effect on inter-chain disulfide bond prediction. Features of site 3, 4, 6 and 8 have the second greatest effect on disulfide bond prediction. Features of site 7 have the least effect on inter-chain disulfide bond prediction. The site-specific distribution of the optimal feature set reveals that the residues at the distal sites and the center are more important for inter-chain disulfide bond prediction than residues at the directly adjacent sites to the cysteine.

3.3.2. Intra-chain Disulfide Bonds

In the optimized 261 features, there were 47 amino acid factor features, 3 disorder score features, 210 PSSM conservation score features and one distance feature. Fig. (**2C**) shows that all three kinds of features contribute to the prediction of intra-chain disulfide bonds and conservation plays the most important role in disulfide bond prediction. The index

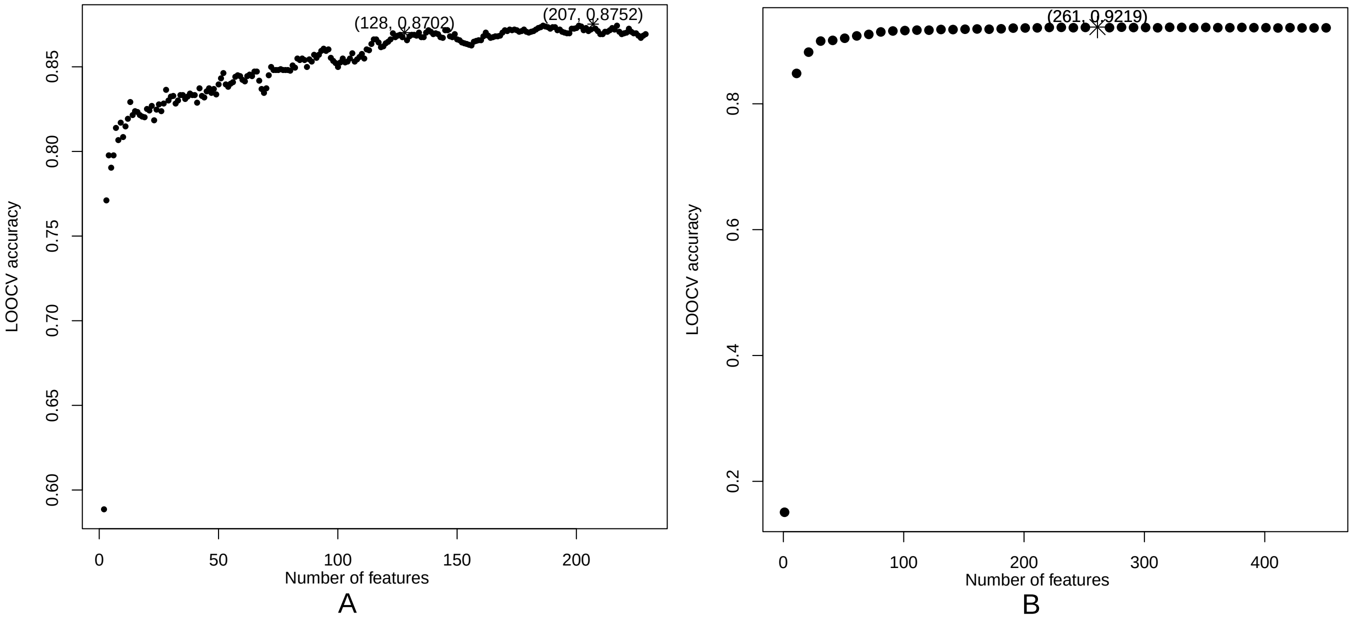

Figure 1. Distribution of prediction accuracy against feature numbers for inter- and intra-chain disulfide bond prediction
(**A**) Distribution of prediction accuracy against feature numbers for inter-chain disulfide bond prediction. The maximum accuracy is 0.8752 using 207 features. To focus our analysis on a relatively small set of features, we selected the first feature set that achieves prediction accuracy more than 0.87 as the optimal feature set, which contains 128 features with an accuracy of 0.8702. (**B**) Distribution of prediction accuracy against feature numbers for intra-chain disulfide bond prediction. The maximum accuracy is 0.9219 containing 261 features. These 261 features were considered as the optimal feature set of our classifier.

of sequence distance in the optimal feature set is 2, indicating its importance in intra-chain disulfide bond site prediction, which is consistent with a previous study [14].

The site specific distribution of the optimal feature set shown in Fig. (**2D**) demonstrates that the center (site 4, 5, 6) and relatively distal sites (site 1, 2 and site 8, 9) have the greatest effect on cysteine disulfide bond determination. The remaining two sites (site 3 and site 7) have less effect on intra-chain disulfide bond determination. The site-specific distribution of the optimal feature set reveals that the residues at two distal and the center sites are more important for cysteine disulfide bond prediction than the remaining sites.

3.3.3. Comparison of Optimal Feature Set between Inter- and Intra-chain Disulfide Bonds

From Fig. (**2A**) and Fig. (**2C**), we can see that PSSM conservation, amino acid factor and disorder features all contribute to both inter- and intra-chain disulfide bond predictions. Site-specific distribution of the optimal feature set illustrated that sites at the center (site 4, 5 and 6) and two ends (site 1, 2 and site 8, 9) contribute more to the prediction of both inter- and intra-chain disulfide bonds. Site 7 contributes less to both inter- and intra-chain disulfide bond predictions. Features derived from site 3 contribute more to the inter-chain disulfide bond prediction than to the intra-chain disulfide bond prediction.

3.4. PSSM Conservation Feature Analysis

We investigated and compared the feature- and site-specific distribution of the 99 and 210 PSSM conservation features in the optimal feature sets for inter- and intra-chain disulfide bond prediction respectively.

3.4.1. Inter-chain Disulfide Bonds

As shown in Fig. (**3A**), the conservation status against Cysteine (C) play the most important role in disulfide bond prediction. The conservation status against A, M, W, H, P, Y, V plays the second most important role in disulfide bond prediction.

As shown in Fig. (**3B**), the conservation status of cysteine (site 5) is most important for disulfide bond prediction. Sites at both ends (site 1, 2, 3 and site 8, 9) play the second most important role in disulfide bond determination. The sites adjacent to the cysteine site play the least role in disulfide bond prediction.

3.4.2. Intra-chain Disulfide Bonds

As shown in Fig. (**3C**), the conservation status against C, S and A influences most on intra-chain disulfide bond determination than against other residues. The conservation status against R, H, K, and Y plays the second most important role in intra-chain disulfide bond determination.

As shown in Fig. (**3D**), the conservation status at the center (site 4, 5 and 6) and distal sites (site 1, 2 and site 8, 9) influence most on intra-chain disulfide bond determination.

3.4.3. Comparison of PSSM Conservation Feature between Inter- and Intra-chain Disulfide Bonds

From Fig. (**3A**) and Fig. (**3C**), we can see that the conservation status against C is the most important feature for both inter- and intra-chain disulfide bond predictions than conservation scores against other residues. Conservation status against A, H and Y play important roles in both inter- and intra-chain disulfide bond predictions. However, there

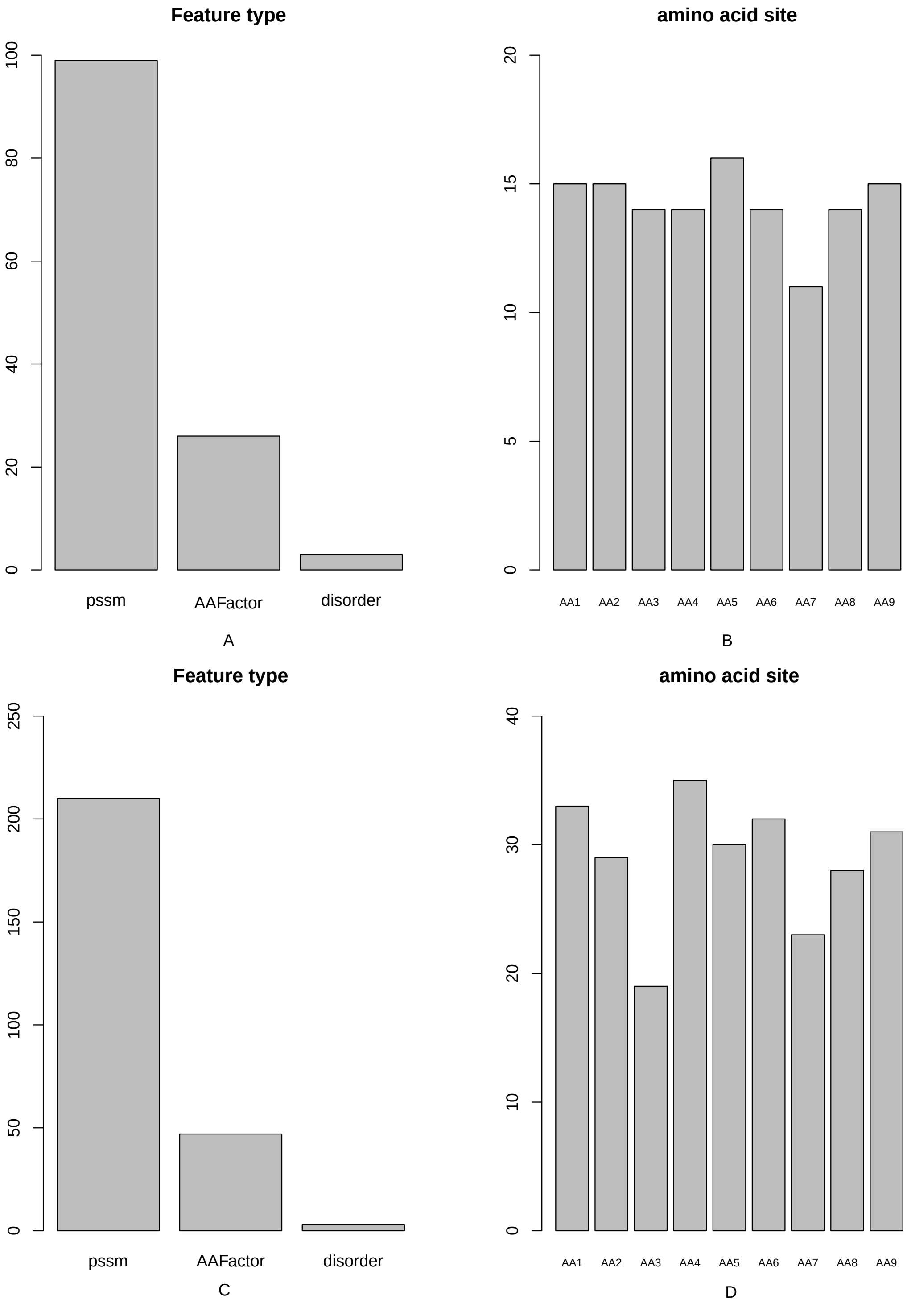

Figure 2. Distribution of the optimal feature set for inter- and intra-chain disulfide bond prediction
(**A**) Feature-specific distribution of the optimal feature set for inter-chain disulfide bond prediction. In the optimized 128 features, there were 26 amino acid factor features, 3 disorder score features and 99 PSSM conservation score features. This suggests that all the three kinds of features contribute to the prediction of protein cysteine disulfide bonds and conservation may play an irreplaceable role in disulfide bond prediction. (**B**) Site-specific distribution of the optimal feature set for inter-chain disulfide bond prediction. The center site (site 5) and distal sites (site 1, 2 and 9) play the most important role in inter-chain disulfide bond prediction. Site 3, 4, 6 and 8 play the secondary most important role in inter-chain disulfide bond prediction, and site 7 are relatively less important in inter-chain disulfide bond prediction. (**C**) Feature-specific distribution of the optimal feature set for intra-chain disulfide bond prediction. In the optimized 261 features, there were 47 amino acid factor features, 3 disorder score features and 210 PSSM conservation score features and one distance feature. (**D**) Site-specific distribution of the optimal feature set for intra-chain disulfide bond prediction. The center (site 4, 5, 6) and distal sites (site 1, 2 and site 8, 9) influence most on intra-chain disulfide bond determination. The remaining two sites (site 3 and site 7) influence less on intra-chain disulfide bond determination.

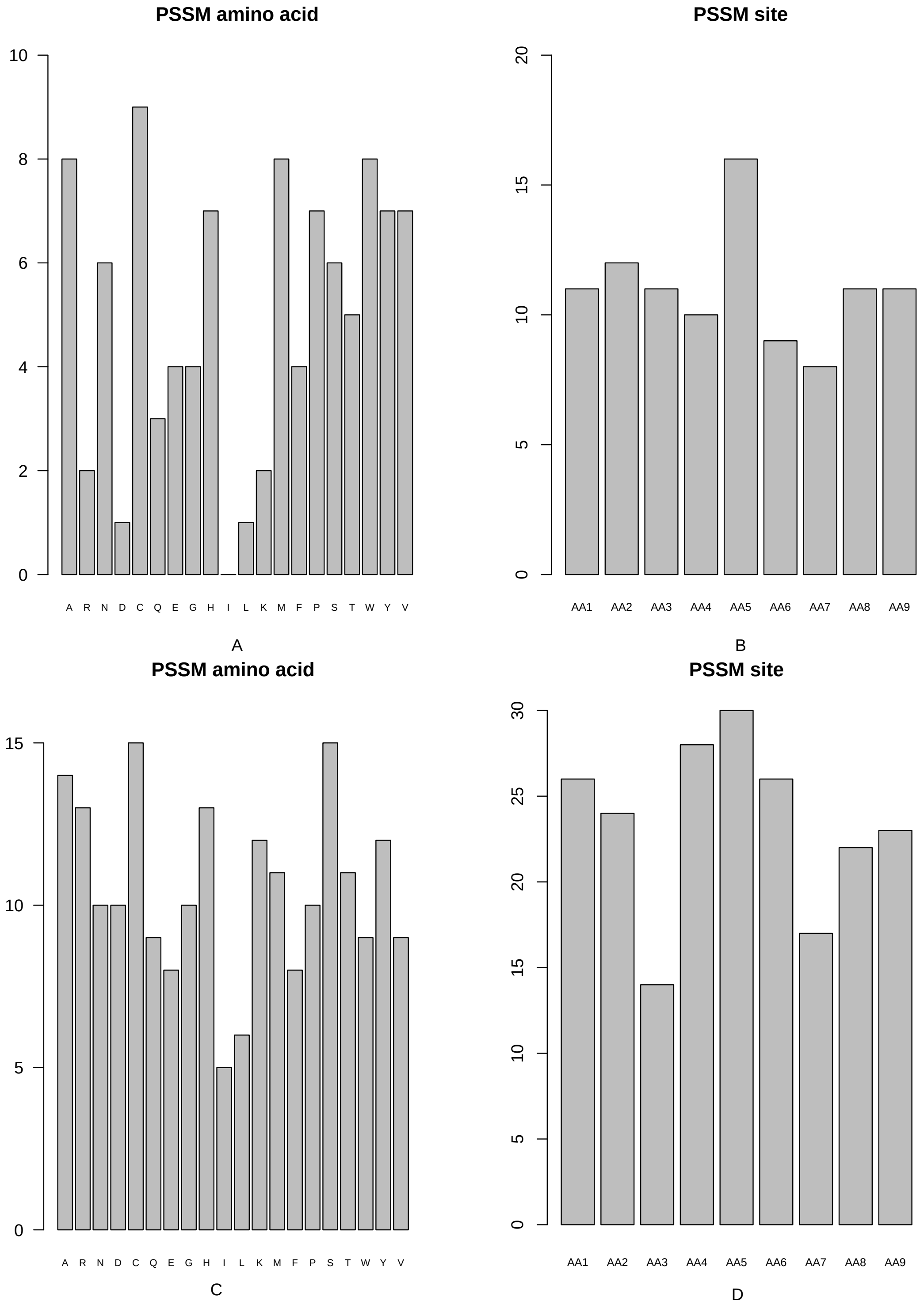

Figure 3. Distribution of the PSSM features in the optimal feature set

(**A**) Feature-specific distribution of the optimized PSSM features for inter-chain disulfide bond prediction. The conservation status against Cysteine (**C**) plays the most important role in disulfide bond determination. Otherwise, the conservation status against A, M, W, H, P, Y, V plays the second most important role in disulfide bond determination. (**B**) Site-specific distribution of the optimized PSSM features for inter-chain disulfide bond prediction. The conservation status of cysteine (site 5) is most important. Conservation status of sites at both sides (site 1, 2, 3 and site 8, 9) plays the second most important role in inter-chain disulfide bond determination. The conservation status of sites adjacent to the cysteine site play relatively less important role in inter-chain disulfide bond determination. (**C**)Feature-specific distribution of the optimized PSSM features for intra-chain disulfide bond prediction. The conservation status against C, S and A are the most influential features on disulfide bond determination than other residues, followed by the conservation status against R, H, K, and Y. (**D**) Site-specific distribution of the optimized PSSM features for intra-chain disulfide bond prediction. The conservation status at the center (site 4, 5 and 6) and relatively distal sites (site 1, 2 and site 8, 9) influence most on disulfide bond determination.

exist some differences between these two disulfide bond types: for inter-chain disulfide bond prediction, the conservation status against M, W, P and V play more important roles, and for intra-chain disulfide bond prediction, the conservation status against S, R and K play more important roles.

From Fig. (**3B**) and Fig. (**3D**), we can see that conservation status at site 5 plays the most important roles in both inter- and intra-chain disulfide bond prediction. Conservation statuses at two end sides (site 1, 2 and site 8, 9) play the second most important roles in both inter- and intra-chain disulfide bonds prediction. However, conservation status of site 4 and 6 are more important for intra-chain disulfide bond prediction than inter-chain disulfide bond prediction. Conservation status of site 3 is more important for inter-chain disulfide bond prediction than intra-chain disulfide bond prediction.

3.5. Amino Acid Factor Analysis

We investigated and compared the feature- and site-specific distribution of the 26 and 47 features of amino acid factors in the optimal feature sets for inter- and intra-chain disulfide bond prediction respectively.

3.5.1. Inter-chain Disulfide Bonds

We investigated the feature- and site-specific distribution of the 26 features of amino acid factors in the optimal feature set.

As shown in Fig. (**4A**), secondary structure and molecular volume play the most important role in inter-chain disulfide bond determination. Codon diversity and polarity play the second most important role in inter-chain disulfide bond determination. Electrostatic charge contributes least in inter-chain disulfide bond determination.

As shown in Fig. (**4B**), amino acid factor features at site 4 and site 6 play the most important role in inter-chain disulfide bond determination. The remaining 6 sites play less important role in inter-chain disulfide bond determination.

3.5.2. Intra-chain Disulfide Bonds

As shown in Fig. (**4C**), secondary structure contributes most to intra-chain disulfide bond determination. Electrostatic charge and codon diversity contribute the second most to intra-chain disulfide bond determination. Polarity and molecular volume contribute relatively less to intra-chain disulfide bond determination.

As shown in Fig. (**4D**), amino acid factor features at site 4 contribute most to intra-chain disulfide bond determination. The amino acid factor features at the remaining sites contribute less to intra-chain disulfide bond determination.

3.5.3. Comparison of Amino Acid Factor Features between Inter- and Intra-chain Disulfide Bonds

From Fig. (**4A**) and Fig. (**4C**), we can see that secondary structure plays the most important role in both inter- and intra-chain disulfide bond predictions. The molecular volume is more important for inter-chain disulfide bond prediction while electrostatic charge feature is more important for intra-chain disulfide bond prediction.

3.6. Feature Analysis of Disorder Score

We investigated and compared the site-specific distribution of the 3 disorder features in the optimal feature sets for both the inter- and intra-chain disulfide bond prediction.

3.6.1. Inter-chain Disulfide Bonds

There were 3 disorder features in the optimal feature set. They located at site 1, 6 and 9, indicating that the disorder status at directly adjacent sites and distal sites is important for inter-chain disulfide bond determination.

3.6.2. Intra-chain Disulfide Bonds

There were 3 disorder features in the optimal feature set, 2 located at site 9 and 1 located at site 1. This indicates the importance of disorder status at site 9 and site 1 on intra-chain disulfide bond determination.

3.6.3. Disorder Score Comparison between Inter- and Intra-chain Disulfide Bonds

For inter-chain disulfide bond prediction, disorder features at site 1, 6 and 9 were selected. For intra-chain disulfide bond prediction, one disorder feature at site 1 and two disorder features at site 9 were selected. The results demonstrated that disorder scores at site 1 and 9 are important for both inter- and intra-chain disulfide bond determinations. Disorder status at site 6 is important for inter-chain disulfide bond prediction, but it was not selected in the optimal feature set for intra-chain disulfide bond prediction.

3.7. Distance Feature

The optimal feature set for intra-chain disulfide bond prediction included the sequence distance with an index of 2, indicating that the distance between a pair of Cysteines plays important role in intra-chain disulfide bond determination, which is consistent with a previous study [14].

3.8. Directions for Experimental Validation

We investigated the top 10 and 20 features (as shown in Table **1** and Table **2**) in the optimal feature sets for inter- and intra-chain disulfide bond prediction respectively. The detailed analysis of the top features may provide clues for understanding the mechanism of disulfide bond formation and for further experimental studies.

3.8.1. Inter-chain Disulfide Bonds

In inter-chain disulfide bond prediction, there are 5 amino acid factor features, 4 PSSM conservation features and 1 disorder feature among the top 10 features. This indicates that amino acid factor features play the most important role for inter-chain disulfide bond prediction. Previous study has demonstrated that inter-chain disulfide bond is more susceptible to reduction than intra-chain disulfide bond [48], which may be mediated by the physicochemical properties of the residues surrounding the disulfide bonding sites. The disorder score at site 6, which was not in the optimal feature set of intra-chain disulfide bond, has an index of 4 in the inter-chain disulfide bond optimal feature set, indicating that it functions differently in inter- and intra-chain disulfide bond determination.

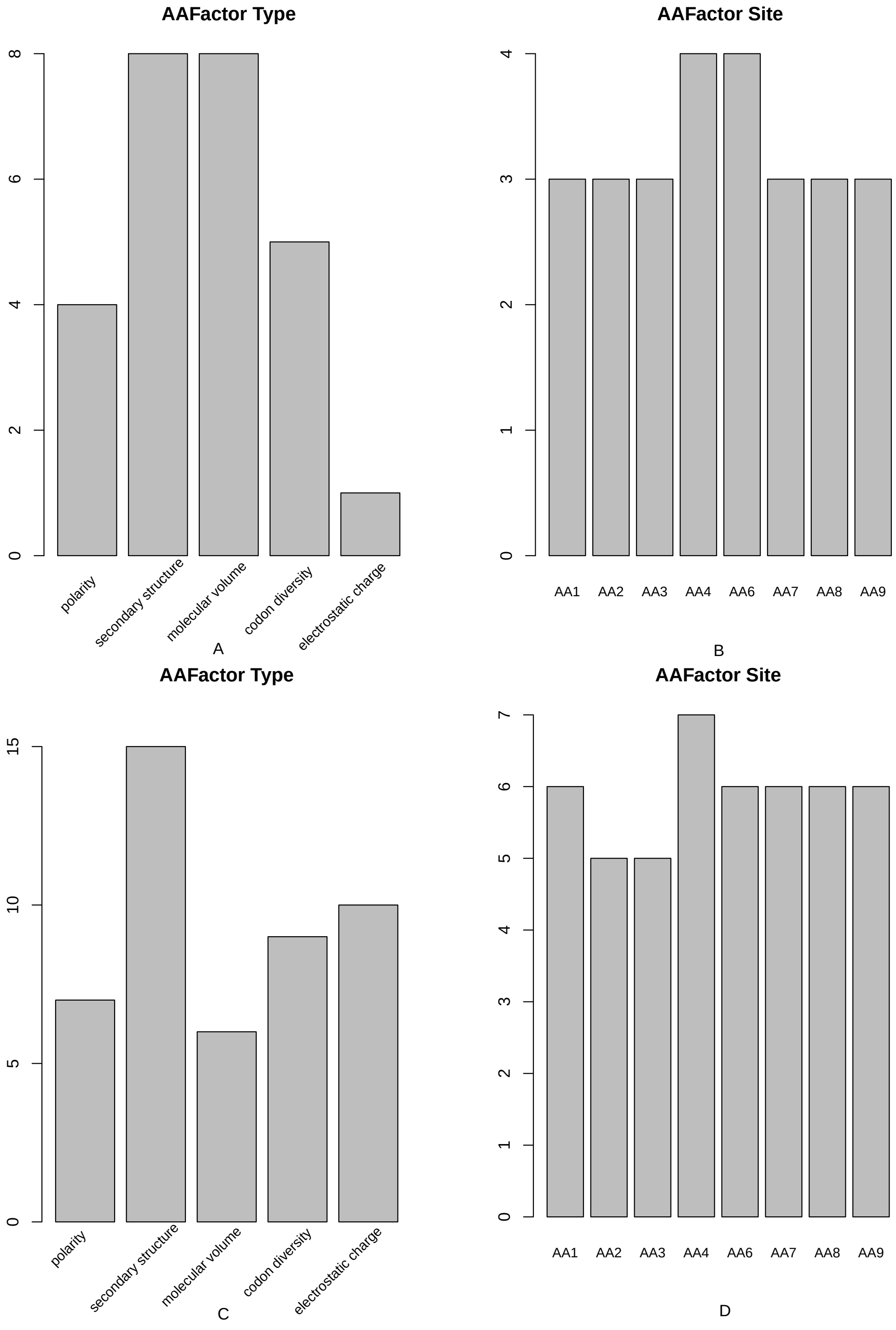

Figure 4. Distribution of the amino acid factor features in the optimal feature set
(**A**) Feature-specific distribution of the features of the optimized amino acid factors for inter-chain disulfide bond prediction. Secondary structure and molecular volume play the most important role in disulfide bond determination. Codon diversity and polarity play the second most important role in disulfide bond determination. Electrostatic charge plays relatively less important role. (**B**) Site-specific distribution of the optimized amino acid factor features for inter-chain disulfide bond prediction. Amino acid factor features at site 4 and site 6 play the most important role in disulfide bond determination. The remaining 6 sites play less important role in disulfide bond determination. (**C**) Feature-specific distribution of the optimized amino acid factor features for intra-chain disulfide bond prediction. Secondary structure features play the most important role in intra-chain disulfide bond determination. Electrostatic charge and codon diversity features play the second most important role in intra-chain disulfide bond determination. Polarity and molecular volume features play relatively less important role in intra-chain disulfide bond determination. (**D**) Site-specific distribution of the optimized amino acid factor features for intra-chain disulfide bond prediction. Amino acid factor features at site 4 play the most important role in intra-chain disulfide bond determination. The amino acid factor features at the remaining sites play relatively less important role in intra-chain disulfide bond determination.

3.8.2. Intra-chain Disulfide Bonds

For intra-chain disulfide bond prediction, we can see that within the top 20 features, there are 19 PSSM conservation features and 1 sequence distance feature. This indicates that the conservation status is the most important feature for intra-chain disulfide bond prediction. Previous study had shown that bound cysteines are significantly more conserved than unbound one [49]. The correlated mutation patterns of cysteine pairs forming disulfide bond had also been illustrated by [15]. The sequence distance between paired cysteine has an index of 2, indicating that it is an important feature for intra-chain disulfide bond prediction, which is consistent with a previous study [14].

3.8.3. Comparison of the Top Features between Inter- and Intra-chain Disulfide Bonds

Comparing the top features between inter- and intra-chain disulfide bonds, we can see that differences exist in the prediction mechanisms between inter- and intra-chain disulfide bond predictions. Within the top 10 features (shown in Table **1**) for inter-chain disulfide bond prediction, there were 5 amino acid factor, 4 PSSM conservation, and 1 disorder features. While within the top 20 features (shown in Table **2**) for intra-chain disulfide bond prediction, there were 19 PSSM conservation and 1 distance features. This may suggest that PSSM conservation and sequence distance played the most important roles in intra-chain disulfide bond determination, while amino acid factor and disorder features played more important roles in inter-chain disulfide bond determination than in intra-chain disulfide bond determination.

4. COMPARISONS WITH EXISTING METHODS USING INDEPENDENT TEST DATASET

We input the independent test data set into our prediction method. For inter-chain disulfide bonds, the prediction accuracies for positive, negative and total samples are 0.5556, 0.9065 and 0.8996 respectively. For intra-chain disulfide bonds, the prediction accuracies for positive, negative and total samples are 0.7151, 0.9028 and 0.8912 respectively.

We input the independent test data set for intra-chain disulfide prediction into DiANNA [17], a computational method to classify cysteines into reduced, half-cysteine or ligand-bound state using a support vector machine with spectrum kernel. By excluding 13 positive samples and 3175 negative samples in two protein sequences P78504 and Q14766, which cannot be predicted by DiANNA, we got totally 682 positive samples and 7343 negative samples. The prediction accuracies for positive, negative and total samples are 32.26%, 93.35% and 88.16% respectively. The results demonstrated that the prediction accuracy of our method is much better than DiANNA for positive samples, and slightly better than DiANNA for the overall samples.

5. CONCLUSION

In this study, we developed a new computational method for inter- and intra-chain disulfide bond prediction based on maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS), with nearest neighbor algorithm as its prediction model. We used sequence conservation, residual disorder, and amino acid factor as features for inter-chain disulfide bond prediction, and besides these features, the sequence distance between each pair of cysteines is also used for intra-chain disulfide bond prediction. Our approach achieved a prediction accuracy of 0.8702 for inter-chain disulfide bond prediction using 128 features and 0.9219 for intra-chain disulfide bond prediction using 261 features. Optimal feature set analysis demonstrated that different types of features contributed differently to disulfide bond formation and features at the center and two distal sites contributed more to the disulfide bond formation than other sites. Comparison of optimal feature sets and top features revealed the similarities and differences between inter- and intra-chain disulfide bonds, which might help understand more of the mechanism of forming the disulfide bonds and provide clues for researches in this research field.

Table 1. Top 10 Features in the Optimal Feature Set for Inter-chain Disulfide Bond Prediction

Order	Name	Feature Type	Site	Sub Feature Type
1	pssm5.12	PSSM	5	K
2	aai2.3	Amino acid factor	2	Molecular volume
3	aai4.2	Amino acid factor	4	Secondary structure
4	dis6	disorder	6	disorder
5	aai3.4	Amino acid factor	3	Codon diversity
6	pssm4.14	PSSM	4	F
7	aai1.5	Amino acid factor	1	Electrostatic charge
8	pssm5.13	PSSM	5	M
9	aai6.1	Amino acid factor	6	polarity
10	pssm8.3	PSSM	8	N

Table 2. Top 20 Features in the Optimal Feature Set for Intra-chain Disulfide Bond Prediction

Order	Name	Feature Type	Site	Sub Feature Type
1	pssm5.5+	PSSM	5	C
2	DISTANCE	Distance		distance
3	pssm2.12-	PSSM	2	K
4	pssm5.7-	PSSM	5	E
5	pssm6.2+	PSSM	6	R
6	pssm7.12-	PSSM	7	K
7	pssm5.5-	PSSM	5	C
8	pssm1.1+	PSSM	1	A
9	pssm4.14-	PSSM	4	F
10	pssm5.17-	PSSM	5	T
11	pssm5.11-	PSSM	5	L
12	pssm8.14-	PSSM	8	F
13	pssm5.9-	PSSM	5	H
14	pssm6.20+	PSSM	6	V
15	pssm5.6+	PSSM	5	Q
16	pssm1.20-	PSSM	1	V
17	pssm4.1-	PSSM	4	A
18	pssm4.20+	PSSM	4	V
19	pssm5.8-	PSSM	5	G
20	pssm1.15-	PSSM	1	P

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China (2011CB510102, 2011CB510101) and the Innovation Program of Shanghai Municipal Education Commission (12ZZ087).

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Chuang, C.C.; Chen, C.Y.; Yang, J.M.; Lyu, P.C.; Hwang, J.K. Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **2003**, *53*(1), 1-5.

[2] van Vlijmen, H.W.; Gupta, A.; Narasimhan, L.S.; Singh, J. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. *J. Mol. Biol*., **2004**, *335*(4), 1083-92.

[3] Huang, E.S.; Samudrala, R.; Ponder, J.W. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. *J. Mol. Biol*., **1999**, *290*(1), 267-281.

[4] Skolnick, J.; Kolinski, A.; Ortiz, A.R. MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol*., **1997**, *265*(2), 217-41.

[5] Hogg, P.J. Contribution of allosteric disulfide bonds to regulation of hemostasis. *J. Thromb. Haemost*., **2009**, *7 Suppl 1*, 13-6.

[6] Nakamura, T.; Lipton, S.A. Cell death: protein misfolding and neurodegenerative diseases. *Apoptosis*, **2009**, *14*(4), 455-68.

[7] Wess, J.; Han, S.J.; Kim, S.K.; Jacobson, K.A.; Li, J.H. Conformational changes involved in G-protein-coupled-receptor activation. *Trends Pharmacol. Sci*., **2008**, *29*(12), 616-25.

[8] Guo, Q.; Manolopoulou, M.; Bian, Y.; Schilling, A.B.; Tang, W.J. Molecular basis for the recognition and cleavages of IGF-II, TGF-alpha, and amylin by human insulin-degrading enzyme. *J. Mol. Biol*., **2010**, *395*(2), 430-43.

[9] Dranoff, G. Targets of protective tumor immunity. *Ann. N.Y. Acad. Sci*., **2009**, *1174*, 74-80.

[10] Sun, Y.; Smith, D.L. Identification of disulfide-containing peptides by performic acid oxidation and mass spectrometry. *Anal. Biochem*., **1988**, *172*(1), 130-8.

[11] Morris, H.R.; Pucci, P. A new method for rapid assignment of S-S bridges in proteins. *Biochem. Biophys. Res. Commun*., **1985**, *126*(3), 1122-8.

[12] Mobli, M.; King, G.F. NMR methods for determining disulfide-bond connectivities. *Toxicon*, **2010**, *56*(6), 849-54.

[13] Chaudhuri, A.R.; Khan, I.A.; Luduena, R.F. Detection of disulfide bonds in bovine brain tubulin and their role in protein folding and microtubule assembly *in vitro*: a novel disulfide detection approach. *Biochemistry*, **2001**, *40*(30), 8834-41.

[14] Zhu, L.; Yang, J.; Song, J.N.; Chou, K.C.; Shen, H.B. Improving the accuracy of predicting disulfide connectivity by feature selection. *J. Comput. Chem*., **2010**, *31*(7), 1478-85.

[15] Rubinstein, R.; Fiser, A. Predicting disulfide bond connectivity in proteins by correlated mutations analysis. *Bioinformatics*, **2008**, *24*(4), 498-504.

[16] Lin, H.H.; Tseng, L.Y. DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Res*., **2010**, *38 Suppl*, W503-7.

[17] Ferre, F.; Clote, P. DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Res*., **2006**, *34* (Web Server issue), W182-5.

[18] Vincent, M.; Passerini, A.; Labbe, M.; Frasconi, P. A simplified approach to disulfide connectivity prediction from protein sequences. *BMC Bioinformatics*, **2008**, *9*, 20.

[19] 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*(23), 3147-54.

[20] Li, H.; Xing, X.; Ding, G.; Li, Q.; Wang, C.; Xie, L.; Zeng, R.; Li, Y. SysPTM: a systematic resource for proteomic research on post-translational modifications. *Mol. Cell Proteomics*, **2009**, *8*(8), 1839-49.

[21] The universal protein resource (UniProt) in 2010. *Nucleic Acids Res*., **2010**, *38* (Database issue), D142-8.

[22] Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B.E.; Martin, M.J.; McGarvey, P.; Gasteiger, E. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **2009**, *10*, 136.

[23] Niu, S.; Huang, T.; Feng, K.; Cai, Y.; Li, Y. Prediction of tyrosine sulfation with mRMR feature selection and analysis. *J. Proteome Res*., **2010**, *9*(12), 6490-7

[24] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **1997**, *25*(17), 3389-402.

[25] Kawashima, S.; Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res*., **2000**, *28* (1), 374.

[26] Atchley, W.R.; Zhao, J.; Fernandes, A.D.; Druke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*(18), 6395-400.

[27] Wright, P.E.; Dyson, H.J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*., **1999**, *293*(2), 321-31.

[28] Liu, J.; Tan, H.; Rost, B. Loopy proteins appear conserved in evolution. *J. Mol. Biol*., **2002**, *322*(1), 53-64.

[29] Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci*., **2002**, *27*(10), 527-33.

[30] Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A.K.; Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **2006**, *7*, 208.

[31] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern. Anal. Mach. Intell*., **2005**, *27*(8), 1226-38.

[32] Qian, Z.; Cai, Y.D.; Li, Y. A novel computational method to predict transcription factor DNA binding preference. *Biochem. Biophys. Res. Commun*., **2006**, *348*(3), 1034-7.

[33] Huang, T.; Cui, W.; Hu, L.; Feng, K.; Li, Y.X.; Cai, Y.D. Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. *PLoS ONE*, **2009**, *4*(12), e8126.

[34] Chou, K.C.; Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*., **1995**, *30*, 275-349.

[35] Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem*, **2007**, *370*, 1-16.

[36] Chou, K.C.; Shen, H.B. Cell-PLoc: A package of web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, Natural Science, 2010, 2, 1090-1103). *Nat. Protocs*, **2008**, *3*, 153-162.

[37] Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol*., **2011**, *273*(1), 236-247

[38] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett*., **2009**, *16*(1), 27-31.

[39] Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett*., **2009**, *16*, 351-355.

[40] Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *J. Theor. Biol*., **2010**, *263*(2), 203-209.

[41] Gu, Q.; Ding, Y.S.; Zhang, T.L. Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. *Protein Pept. Lett*., **2010**, *17*(5), 559-567.

[42] Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett*., **2010**, *17*(10), 1207-1214.

[43] Ding, H.; Liu, L.; Guo, F.B.; Huang, J.; Lin, H. Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. *Protein Pept. Lett*., **2011**, *18*(1), 58-63.

[44] Hu, L.; Zheng, L.; Wang, Z.; Li, B.; Liu, L. Using pseudo amino acid composition to predict protease families by incorporating a series of protein biological features. *Protein Pept. Lett*., **2011**, *18*(6), 552-558.

[45] Wang, D.; Yang, L.; Fu, Z.; Xia, J. Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction. *Protein Peptide Lett*., **2011**, *18*(7), 684-689.

[46] Wang, J.; Wang, X.Y.; Shu, M.; Wang, Y.Q.; Lin, Y.; Wang, L.; Cheng, X.M.; Lin, Z.H. QSAR study on MHC class I a alleles based on the novel parameters of amino acids. *Protein Peptide Lett*., **2011**, *18*(9), 956-963.

[47] Zeng, Y.H.; Guo, Y.Z.; Xiao, R.Q.; Yang, L.; Yu, L.Z.; Li, M.L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. *J. Theor. Biol*., **2009**, *259*, 366-372.

[48] Liu, H.; Chumsae, C.; Gaza-Bulseco, G.; Hurkmans, K.; Radziejewski, C.H. Ranking the susceptibility of disulfide bonds in human IgG1 antibodies by reduction, differential alkylation, and LC-MS analysis. *Anal. Chem*., **2010**, *82*(12), 5219-26.

[49] Fiser, A.; Simon, I. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics*, **2000**, *16*(3), 251-6.

Received: September 30, 2011 Revised: October 23, 2011 Accepted: October 25, 2011