Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural Information and Random Forest Regression

Dong-Jun Yu, Yang Li, Jun Hu, Xibei Yang, Jing-Yu Yang, and Hong-Bin Shen

Abstract—Disulfide connectivity is an important protein structural characteristic. Accurately predicting disulfide connectivity solely from protein sequence helps to improve the intrinsic understanding of protein structure and function, especially in the post-genome era where large volume of sequenced proteins without being functional annotated is quickly accumulated. In this study, a new feature extracted from the predicted protein 3D structural information is proposed and integrated with traditional features to form discriminative features. Based on the extracted features, a random forest regression model is performed to predict protein disulfide connectivity. We compare the proposed method with popular existing predictors by performing both cross-validation and independent validation tests on benchmark datasets. The experimental results demonstrate the superiority of the proposed method over existing predictors. We believe the superiority of the proposed method benefits from both the good discriminative capability of the newly developed features and the powerful modelling capability of the random forest. The web server implementation, called TargetDisulfide, and the benchmark datasets are freely available at: http://csbio.njust.edu.cn/bioinf/TargetDisulfide for academic use.

Index Terms—Protein structure prediction, disulfide connectivity prediction, feature extraction, regression model, random forest

1 Introduction

It is common knowledge that the three-dimensional (3D) structure of a protein has a close relationship to its biological functions. In the post-genome era, a large volume of sequenced proteins has been accumulated without being structurally determined due to the rapid development of advanced sequencing technology and concerted genome projects [1]. Thus, accurately predicting protein structure from only the sequence is urgent to bridge the protein sequence-structure gap. During the preceding decades, considerable effort has been made to predict 3D structures solely from the protein sequence, and many encouraging results have been reported [2], [3], [4], [5], [6], [7]. However, until now, predicting a complete 3D structure directly from the sequence has been far from successful and is still a challenging and open problem. In this regard, researchers have resorted to decomposing the prediction of complete 3D structure into predictions of special structure segments or characteristics, such as disordered regions [8], [9], [10], [11], transmembrane helices [12], [13], beta-sheets [14], [15], [16], residue-residue contact maps [17], [18], [19], [20], disulfide connectivity [21], [22], [23], [24], [25], [26], solvent accessibility [27], [28], and so on. The knowledge derived from the protein structure segments or characteristics may provide valuable insights into protein 3D structures and may help to understand protein functions [23].

Disulfide connectivity is an important protein structure characteristic. Disulfide bonds are primary covalent cross-links formed between two cysteine residues in the same or different protein polypeptide chains, and these bonds play important roles in the folding and stability of proteins [29], [30], [31]. Predicting which cysteines in a protein sequence will form disulfide connections plays a relevant role in protein structural and functional annotation. Various effective methods have been developed in previous studies, such as DISULFIND [32], Pair-wise SVM [33], GASVM [21], SS_SVR [22] and FS_SVR [23], DBCP [24], DISLOCATE [25], DMC (DISLOCATE+MIp+iCOV) [26], and so on. All these existing methods can be grouped into three categories [25]: (I) disulfide-bonding state prediction; (II) connectivity pattern prediction using the prior knowledge of the disulfide-bonding states of cysteines; and (III) methods that predict both disulfide-bonding state and connectivity patterns.

Recently, considerable attention has been paid to developing machine learning-based methods for disulfide connectivity predictions, and experimental results have shown that applying advanced machine learning algorithms is a promising route to further improve the prediction performance [21], [22], [23], [24], [25], [26], [33]. Extracting discriminative features and utilizing powerful machine

- *D.J. Yu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing 210094, China, and Changshu Institute, Nanjing University of Science and Technology, Research Road 5, Changshu 215513, China. E-mail: njyudj@njustedu.cn.*
- *Y. Li, J. Hu, and J. Y. Yang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing 210094, China. E-mail: balllee2011@sina.com, junh_cs@126.com, yangjy@njust.edu.cn.*
- *X. Yang is with the School of Computer Science and Engineering, Jiangsu University of Science and Technology, 2 Huancheng Road, Zhenjiang 212003, China. E-mail: yangxibei@hotmail.com.*
- *H.B. Shen is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China. E-mail: hbshen@sjtu.edu.cn.*

Manuscript received 11 June 2014; revised 3 Sept. 2014; accepted 15 Sept. 2014. Date of publication 25 Nov. 2014; date of current version 1 June 2015. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TCBB.2014.2359451

learning algorithms are two crucial aspects that will significantly affect the performance of machine learning-based disulfide connectivity prediction. Tremendous efforts have been made to uncover effective features and many feature sources, such as position specific scoring matrix (PSSM) [23], [34], predicted secondary structure (PSS) [34], correlated mutations (CM) [26], [35], cysteine separation distances [33], and subcellular localizations [25], were proven to be useful for disulfide connectivity prediction. As for classifiers, many advanced machine learning algorithms, e.g., support vector regression (SVR) [23], [34] and grammatical-restrained hidden conditional random fields (GRHCRFs) [25], have been widely applied to disulfide connectivity prediction.

In summary, there has been a number of good results in disulfide connectivity prediction [36]. Nevertheless, there is still room for further improving the prediction performance. This paper aims to accomplish this task by introducing new discriminative features derived from the predicted protein 3D structural information and the application of random forest (RF) regression, as proposed by Breiman [37]. The proposed method, called TargetDisulfide, falls into category II and mainly focuses on the connectivity pattern prediction of intra-chain disulfide bonds under the assumption that prior knowledge of the disulfide-bonding states of cysteines is available. We also integrated the disulfide-bonding state prediction module of DiANNA [38] into the proposed TargetDisulfide to improve its applicability under the scenario where knowledge of the disulfide-bonding states of cysteines is not available. Experimental results on three benchmark datasets demonstrate that the newly developed features derived from the predicted protein 3D structural information can significantly improve the prediction performance. Research in this study enriches the contents of disulfide connectivity prediction, and the implemented TargetDisulfide complements existing disulfide connectivity predictors.

2 Materials and Methods

2.1 Benchmark Datasets

To objectively evaluate the proposed method and fairly compare it with existing sequence-based disulfide connectivity prediction methods, two datasets that were used in previous studies were taken as benchmark datasets in this study.

The first dataset, denoted as SP39, was constructed by Fariselli and Casadio [39] as follows: first, sequences containing at least two ($B = 2$) and at most five ($B = 5$) intra-chain disulfide bonds that have been experimentally verified were extracted from the Swiss-Prot 39 release; then, the maximal pairwise sequence identity of the extracted protein sequences was culled to 30 percent; the remaining 446 sequences constitute SP39, among which there are 156, 146, 99, and 45 sequences that have two, three, four, and five disulfide bonds, respectively.

The second dataset, PDBCYS-R, was constructed from a recently released dataset, i.e., PDBCYS [40], for disulfide bond predictions. The original PDBCYS dataset consists of 1,797 protein sequences, the maximal pairwise sequence identity of which was reduced to 25 percent. We removed the sequences that had less than two or more than five disulfide bonds from the original PDBCYS dataset, and the remaining 263 protein sequences constitute the reduced dataset PDBCYS-R.

Note that we will perform cross-validation comparisons between the proposed method and other existing methods on both benchmark datasets, and the final online web-server, i.e., TargetDisulfide will be implemented based on the SP39 dataset.

To objectively evaluate the generalization capability of the proposed TargetDisulfide, we compared it with several popular predictors, including DiANNA [38], DISULFIND [32], and DBCP [24], with an independent validation dataset consisting of 54 sequences, denoted as IVD-54. We constructed the independent validation dataset from UniProtKB/Swiss-Prot [41] as follows:

Step 1. Sequences that are annotated as "reviewed" and contain at least two and at most five intra-chain disulfide bonds are selected from UniProtKB/Swiss-Prot [41] to constitute the initial dataset; In addition, all of the sequences in the initial dataset that have corresponding experimentally determined structures in the PDB are also removed.

Step 2. Removal of the sequences that are annotated with uncertain terms, such as "by similarity", "probable", or "potential", from the initial dataset.

Step 3. Further, we reduce the maximal pairwise sequence identity of the initial dataset to 30 percent by applying the CD-HIT program [42], and the reduced dataset is obtained.

Step 4. Moreover, if a given sequence in the reduced dataset shares >30 percent identity to a sequence in the SP39, we remove the sequence from the reduced dataset; the purpose for doing this is to guarantee that the sequences in the obtained dataset are actually independent from the sequences in the training dataset, i.e., SP39, which is used by TargetDisulfide.

Step 5. Among the predictors used for independent validation comparison, DBCP [24] is one of the state-of-the-art predictors for disulfide connectivity prediction. Importantly, DBCP continues to evolve by improving its underlying learning algorithm and training dataset. Currently, DBCP has been updated to EDBCP. To fairly compare the proposed TargetDisulfide with EDBCP, any sequence in the dataset obtained in Step (4) that shares > 30 percent identity and has an E-value < 10 to a template sequence used by EDBCP will also be removed.

Step 6. The final independent validation dataset, denoted as IVD-54, consists of 54 sequences, among which there are 29, 15, 7, and 3 sequences that have two, three, four, and five disulfide bonds, respectively.

Table 1 summarizes the detailed compositions of the three benchmark datasets, i.e., SP39, PDBCYS-R, and IVD-54.

2.2 Feature Representation

Feature representation is a critical step in designing a machine learning-based predictor. As stated in the introduction section, several feature sources, e.g., position specific scoring matrix [23], [35], predicted secondary structure [34], correlated mutations [26], [35], and cysteine separation

TABLE 1
Compositions of SP39, PDBCYS-R, and IVD-54 Datasets

No. of Disulfide Bonds	Dataset					
	SP39		PDBCYS-R		IVD-54	
	No. of Sequences	No. of disulfide bonds	No. of Sequences	No. of disulfide bonds	No. of Sequences	No. of disulfide bonds
B = 2	156	312	100	200	29	58
B = 3	146	438	85	255	15	45
B = 4	99	396	41	164	7	28
B = 5	45	225	37	185	3	15
Total	446	1,371	263	804	54	146

distance [33], have been demonstrated to be especially useful for extracting effective discriminative features in previous studies and thus will also be utilized for the disulfide connectivity prediction in this study. In addition, a new feature derived from a protein 3D structure modelled by homology modelling software MODELLER [6] is introduced to further improve the discriminative capability of the extracted features. Next, we briefly describe how to extract these features as follows:

2.2.1 Position Specific Scoring Matrix Feature

Disulfide connectivity is a stereo-specific secondary structural element [34] that will be well conserved during the process of protein evolution. As the position specific scoring matrix encodes the evolutionary information of a protein, the feature derived from PSSM has been widely and successfully applied to disulfide connectivity predictions [23], [34], [43]. In this study, we extract the PSSM feature as follows: the original PSSM of a given protein sequence is obtained by executing PSI-BLAST [44] to search the Swiss-Prot database through three iterations with a default *E*-value cutoff; then, we transform the original PSSM to a normalized one by applying the logistic function $f(x) = 1/(1 + \mathrm{e}^{-x})$ to each element x contained in the original PSSM; finally, each cysteine residue is encoded into a $13 \times 20 = 260$-D feature vector that consists of the normalized PSSM elements corresponding to a sequence segment of length 13 centered on the cysteine residue [23], [34].

2.2.2 Predicted Secondary Structure Feature

The predicted secondary structure is another feature source that was proven useful for not only disulfide connectivity prediction [34] but also many other bioinformatics problems, such as protein-ligand binding [45], [46], proline *cis*/*trans* isomerization prediction [47], residue-wise contact order prediction [48], and so on. In view of this, the protein secondary structure predicted from protein sequence by performing PSIPRED [49] is also used as another feature source.

PSIPRED [49] software predicts the propensities of belonging to three secondary structure classes (coil, helix, and strand) for each residue in a protein sequence. More specifically, for a protein sequence with L residues, the PSIPRED outputs an $L \times 3$ probability matrix, which represents the predicted secondary structure information of the protein. Again, a sliding window of size 13 was used to extract the predicted protein secondary structure feature of each cysteine residue and the dimensionality of the extracted feature vector is $13 \times 3 = 39$-D.

2.2.3 Sequence Distance between Oxidized Cysteines Feature (DOC)

The sequence distance between oxidized cysteines was first proposed by Tsai et al. [33]. DOC encodes each cysteine pair into a real value as follows:

$$\mathrm{DOC}(i,\ j) = |i - j|, \tag{1}$$

where i and j are the positions of the two cysteine residues within a protein sequence.

Three scaling strategies for DOC, i.e., DOC_L, DOC_{max}, and DOC_{log}, which normalize DOC by the protein sequence length L, the maximum value of the protein sequence length in the whole dataset, and the logarithm function, respectively, have been tested, and the last one, i.e., DOC_{log}, has been demonstrated to be the most effective [33] and thus will also be used in this study

$$\mathrm{DOC}(i,\ j) = \frac{1}{1 + \log(|i - j|)}. \tag{2}$$

2.2.4 Correlated Mutations Feature

Two cysteine residues in a disulfide bond form a strong interaction, which is expected to lead to interdependency between the two positions and can be traced through evolution [35]. In light of this, Rubinstein and Fiser [35] first performed disulfide bond connectivity prediction by correlated mutation analysis. In this study, the same scoring scheme is used, which encodes each cysteine pair into a real value in the range of [0,1], as developed by Rubinstein and Fiser [35]. Details on the correlated mutations scoring scheme can be found in reference [35].

2.2.5 Predicted Distance between Two Cysteine Residues Feature (PDTCR)

Theoretically speaking, the disulfide connectivity prediction problem will be completely solved if we can develop a prediction model that can perfectly predict the accurate 3D structure of a protein from its sequence. Unfortunately, such an ideal model is still unavailable. Nevertheless,

several promising methods, e.g., MODELLER [6] and I-TASSER [7], have been developed and demonstrate the feasibility of modeling a 3D structure from the protein sequence. The predicted structural distance information between two cysteine residues contained in the modeled 3D structure may potentially help with disulfide connectivity predictions. In this regard, we extract the predicted distance between the two cysteine residues for each cysteine pair from the modeled 3D structure as follows: for a protein sequence, we model its 3D structure with the homology modelling algorithm MODELLER [6]; then, the PDTCR feature of a cysteine pair is obtained by calculating the distance between the two residues in the cysteine pair according to their coordinates in the modeled 3D structure. The smaller the distance is, the higher the possibility that the cysteine pair forms a disulfide bond.

Clearly, the reliability of the modelled structure for a query protein sequence will be improved by using the templates that have higher homology identity to the query sequence. Accordingly, the PDTCR feature extracted from a modeled 3D structure with higher reliability will be more accurate. However, to objectively evaluate the effectiveness of the PDTCR feature, we restricted the homology identity of the templates to be 40 percent in this study. More specifically, when modelling the 3D structure for a query sequence, all the known structures of those sequences, which share >40 percent identity to the query sequence, are removed from the underlying template library used by MODELLER. For benchmark comparison purpose, we also conducted experiments on different homology identity cutoffs, e.g. from 30 to 70 percent.

Finally, the feature vector of a cysteine pair is formed by combining all the abovementioned five types of features, i.e., PSSM feature (520-D, $520 = 2 \times 260$), PSS feature (78-D, $78 = 2 \times 39$), DOC (1-D), CM (1-D), and PDTCR (1-D), and the dimensionality of the obtained feature vector is $520 + 78 + 1 + 1 + 1 = 601$-D.

2.3 Regression Models

Many machine learning methods, such as Bayesian regression analysis [50], hidden Markov model [51], support vector regression [52], [53], random forest [37], and so on, can be applied to construct a prediction model. In this study, two powerful regression models, i.e., SVR and RF were tested, and the better one, i.e., RF, was chosen for constructing the online web server.

SVR is a regression model based on a support vector machine (SVM) by utilizing a kernel function, the ε-insensitive loss function and the regularization parameter ξ. When training a SVR model, two parameters, i.e., regularization parameter ξ and the kernel parameter γ need to be determined in advance. The selection of these two parameters is important to the performance of the trained SVR model. In this study, LIBSVM [53], which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/, was used. A radial basis function was chosen as the kernel function. The other two parameters, i.e., the regularization parameter ξ and the kernel parameter γ, were optimized based on cross-validation using a grid search strategy in the LIBSVM software.

RF proposed by Breiman [37] is an ensemble method that adds an additional layer of randomness to bagging. In a random forest, each node is split using the best among a subset of predictors *randomly* chosen at that node, which is different from standard trees, where each node is split using the best split among all variables [54]. The randomness strategy used in RF has been demonstrated to be very effective compared with many other classifiers and is robust against over-fitting [37], [54]. RF can be used to perform both classification and regression. In this study, a random forests regression model is taken. We utilized the RF code, which is freely accessible at http://scikit-learn.org/stable/modules/ensemble.html#random-forests, to evaluate and implement the proposed TargetDisulfide. Two parameters, i.e., the number of trees to grow (*nTree*) and the number of dimensions randomly sampled as candidates at each split (*mTry*), were set to be 500 and $24 \approx \sqrt{601}$, respectively, where 601 is the dimensionality of the feature used for disulfide connectivity prediction.

2.4 Workflow of the Proposed Method

Fig. 1 illustrates the workflow of the proposed TargetDisulfide. TargetDisulfide accomplishes disulfide connectivity pattern prediction for a query sequence with a two-stage scheme as follows:

Stage I: Predicting the propensity of disulfide bond information for each of the possible cysteine pairs. In the first stage, TargetDisulfide predicts the propensity of being a disulfide bond for each of the possible cysteine pairs contained in the query sequence. Taking the cysteine pair illustrated in Fig. 1 as an example, the PSSM, PSS, CM, DOC, and PDTCR features for the cysteine pair, which is composed of the two cysteine resides located at positions *3* and *n-1* in the sequence and is highlighted in Red and Blue, respectively, are extracted. Then, the extracted features are combined and fed into the trained RF regression model to give the *p*redict *p*ropensity of the *c*ysteine *p*air, denoted as *ppcp*, of being a disulfide bond.

Stage II: Predicting disulfide connectivity pattern. Let P be the number of possible disulfide connectivity patterns; the score of the *i*th ($1 \le i \le P$) possible disulfide connectivity pattern is obtained by taking the scoring scheme that has been previously used by other studies [23], [34] as follows:

$$S_i = \sum_{j=1}^{B} ppcp_j,\ 1 \le i \le P, \tag{3}$$

where B and $ppcp_j$ are the number of pairs and the propensity of the *j*th cysteine pair, respectively, contained in the *i*th disulfide connectivity pattern.

Then, the i^*th possible disulfide connectivity pattern, which has the maximal score, is predicted as the final result:

$$i^* = \arg\max_{1 \le i \le P} S_i. \tag{4}$$

2.5 Evaluation Indexes

In this study, two routinely used evaluation indexes [23], [25], [26], [34], i.e., Q_C and Q_P were also taken to fairly compare the proposed method with existing methods.

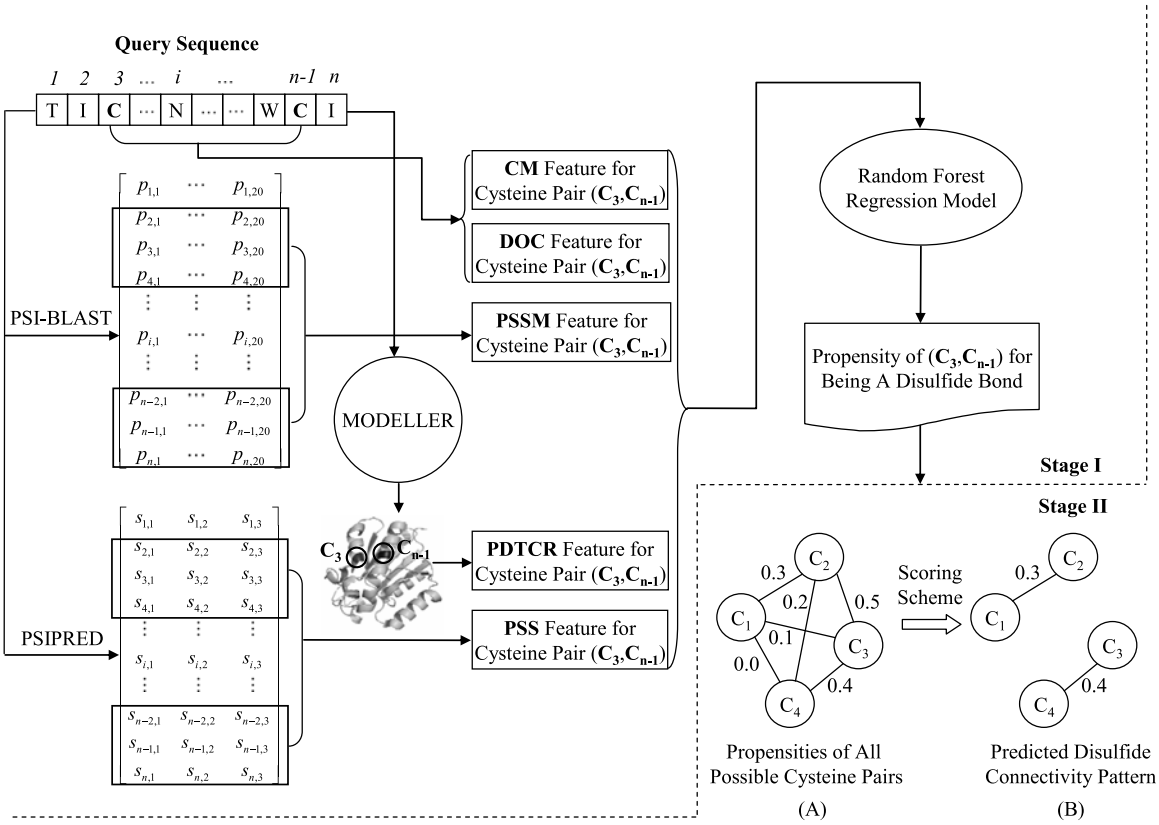


Fig. 1. Workflow of the proposed TargetDisulfide.

Q_C measures the percentage of correctly predicted disulfide bonds among all the observed disulfide bonds, which is defined as follows:

$$Q_C = \frac{N_C}{T_C} \times 100\%, \tag{5}$$

where N_C is the number of observed disulfide bonds that are correctly predicted and T_C is the total number of observed disulfide bonds throughout a test dataset.

Q_P measures the percentage of the proteins where the observed disulfide bonds are all correctly predicted. A protein is considered being correctly predicted if and only if all the observed disulfide bonds in this protein are correctly predicted and the number of predicted disulfide bonds is equal to the number of observed disulfide bonds in the protein. Q_P is defined as follows:

$$Q_P = \frac{N_P}{T_P} \times 100\%, \tag{6}$$

where N_P is the number of proteins that have been correctly predicted and T_P is the total number of proteins in a dataset.

3 Experimental Results and Analysis

3.1 PDTCR Feature Improves Prediction Performance

In this section, we will empirically demonstrate that the prediction performance can be significantly improved by integrating the new PDTCR feature (Predicted Distance between Two Cysteine Residues) with the traditional features, i.e., PSSM, PSS, CM, and DOC. Experiments on the SP39 and PDBCYS-R datasets were carried out as follows:

We obtained the baseline prediction performances by cross-validating each of the two benchmark datasets with only the four traditional features (i.e., without PDTCR feature); similarly, we obtain the prediction performances on each of the two benchmark datasets with the four traditional features plus the PDTCR feature.

For the purpose of a fair comparison with existing predictors, this study performed a four-fold cross-validation on SP39 and a 20-fold cross-validation on PDBCYS-R, respectively, because many existing predictors developed on these two datasets were also evaluated with the corresponding cross-validations [21], [22], [23], [25], [26], [32], [33]. Table 2 summarizes the prediction performances between without- and with-PDTCR feature on SP39 and PDBCYS-R datasets using SVR and RF over cross-validation.

From Table 2 it can be seen that the prediction performance is significantly improved by incorporating the newly developed PDTCR feature. Throughout the different number of bonds (i.e., $\mathrm{B} = 2, 3, 4, 5$), the values of Q_P and Q_C with-PDTCR feature are almost consistently higher than those of the without-PDTCR feature, taking SVR and RF as classifiers on both the SP39 and PDBCYS-R datasets. Taking the overall Q_P and Q_C as examples, improvements of 3.6 and 2.8 percent for SVR and 1.6 and 1.9 percent for RF were observed after incorporating the PDTCR feature on the SP39 dataset; whereas the improvements of 8.0 and 7.3 percent for SVR, and 6.1 and 4.4 percent for RF were observed on PDBCYS-R. The results listed in Table 2 demonstrate the efficacy of the newly developed PDTCR feature. The

TABLE 2
Performance Comparisons between without- and with-PDTCR Feature on SP39 and PDBCYS-R Datasets Using SVR and RF Classifiers over Cross-Validation

Dataset	Classifier	Feature	B = 2		B = 3		B = 4		B = 5		Overall	
			Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
SP39[c]	SVR	without-PDTCR [a]	87.2	87.2	65.1	75.1	82.8	87.1	55.6	70.7	75.8	80.6
		with-PDTCR [b]	**90.4**	**90.4**	**71.9**	**79.5**	**82.8**	**88.4**	**57.8**	**72.9**	**79.4**	**83.4**
	RF	without-PDTCR [a]	90.4	90.4	75.3	80.4	**84.8**	**89.6**	57.8	71.1	80.9	83.8
		with-PDTCR [b]	**92.3**	**92.3**	**78.1**	**83.6**	82.8	87.9	**62.2**	**76.9**	**82.5**	**85.7**
PDBCYS-R[d]	SVR	without-PDTCR [a]	79.0	79.0	51.8	63.5	29.3	53.0	13.5	43.8	53.2	60.7
		with-PDTCR [b]	**84.0**	**84.0**	**60.0**	**70.6**	**36.6**	**59.1**	**29.7**	**55.1**	**61.2**	**68.0**
	RF	without-PDTCR [a]	82.0	82.0	64.7	73.3	46.3	65.2	16.2	67.9	61.6	67.9
		with-PDTCR [b]	**83.0**	**83.0**	**76.4**	**82.4**	**53.7**	**65.9**	**21.6**	**52.4**	**67.7**	**72.3**

[a] *Without-PDTCR: four traditional features, i.e., PSSM, PSS, CM, and DOC.*
[b] *With-PDTCR: four traditional features plus PDTCR feature.*
[c] *Results were obtained by performing four-fold cross-validation.*
[d] *Results were obtained by performing 20-fold cross-validation.*

underlying reason for this improvement is that the PDTCR feature measures the spatial distance between two cysteine residues, which has close relationship to the formation of a disulfide bond.

3.2 Can Feature Selection Further Improve Prediction Performance?

Feature selection, the purpose of which is to select the most discriminative feature subset from the original feature space to further improve prediction performance, has been widely used in related bioinformatics fields [55], [56], [57], [58], [59]. For example, our recent work [23] on disulfide connectivity prediction demonstrated that the PSSM feature contains redundant information, and the prediction performance can be further improved when the original PSSM feature is reduced to a lower but more compact feature subspace by performing feature selection. Motivated by this observation, in this study, we also performed feature selection on the original 601-D feature space, as described in the feature representation section, to determine whether the disulfide connectivity prediction performance could be further improved.

The next problem is how to perform feature selection. Existing traditional feature selection methods, such as data variance [23], Fisher score [60], and Laplacian score [61], that we used previously in [23] are faced with two problems: first, the importance of features are calculated individually; thus the correlation and dependence of different feature components are neglected. Second, the dimensionality of the reduced feature subspace needs to be prescribed in advance of the feature selection process, which is often difficult or even impossible in practice [62].

To alleviate the above-mentioned two problems, Yan and Yang [62] recently proposed a generalized Joint Laplacian Feature Weights Learning algorithm, denoted as JLFWL. JLFWL can automatically determine the optimal dimensionality of the feature subspace and select the best feature components from the original feature space by iteratively learning the feature weights of components jointly and simultaneously. For the details of JLFWL, refer to [62]. Considering its advantages, we performed a JLFWL [62] feature selection on the original 601-D feature space, and the optimal dimensionalities of the selected feature subspace on the SP39 and PDBCYS-R are 526 and 462, respectively.

Table 3 summarizes the performance comparisons between without- and with- feature selection (JLFWL) on SP39 and PDBCYS-R datasets using SVR and RF classifiers over cross-validation. By observing Table 3, it can be seen that the prediction performance is improved after performing

TABLE 3
Performance Comparisons between without- and with- Feature Selection (JLFWL) on SP39 and PDBCYS-R Datasets Using SVR and RF Classifiers over Cross-Validation

Dataset	Method	Feature Selection	B = 2		B = 3		B = 4		B = 5		Overall	
			Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
SP39[a]	SVR	No	90.4	90.4	71.9	79.5	82.8	88.4	57.8	72.9	79.4	83.4
		Yes	92.9	92.9	72.6	79.7	83.8	88.9	60.0	73.8	**80.9**	**84.4**
	RF	No	92.3	92.3	78.1	83.6	82.8	87.9	62.2	76.9	**82.5**	**85.7**
		Yes	91.7	91.7	74.7	80.8	80.8	85.6	57.8	73.8	80.3	83.5
PDBCYS-R[b]	SVR	No	84.0	84.0	60.0	70.6	36.6	59.1	29.7	55.1	61.2	68.0
		Yes	84.0	84.0	62.4	72.5	43.9	64.0	24.3	55.7	**62.4**	**69.8**
	RF	No	83.0	83.0	76.4	82.4	53.7	65.9	21.6	52.4	**67.7**	**72.3**
		Yes	82.0	82.0	71.8	78.0	51.2	65.9	27.0	54.6	66.2	71.1

[a] *Results were obtained by performing a four-fold cross-validation.*
[b] *Results were obtained by performing a 20-fold cross-validation.*

TABLE 4
Performance Comparisons between TargetDisulfide and Other Predictors on the SP39 Dataset over Four-Fold Cross-Validation

B	DISULFIND [32]		Pair-wise SVM [33]		GASVM [21]		SS_SVR [22]		FS_SVR [23]		TargetDisulfide	
	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
2	75.0	75.0	79	79	85.7	85.7	86.5	86.5	85.3	85.3	92.3	92.3
3	46.6	55.7	53	62	74.6	79.7	67.1	72.6	69.9	75.6	78.1	83.6
4	50.5	63.4	55	70	63.2	77.1	78.8	84.8	79.7	86.8	82.8	87.9
5	17.8	42.7	58	71	47.6	71.4	46.8	64.0	55.9	71.0	62.2	76.9
Overall	54.5	60.2	63	70	73.9	79.2	74.4	77.9	76.0	80.3	**82.5**	**85.7**

feature selection using SVR as a classifier on both SP39 and PDBCYS-R datasets, which is consistent with the observation that we made previously [23]. However, the performance deteriorated after performing feature selection on both datasets when RF is used as the base classifier. By analyzing the RF algorithm, we found that the feature selection for each split in the classification tree is conducted from a small random subset of predictor variables (features) [37], [63]. In other words, the RF algorithm itself performs the feature selection procedure when constructing random forests. We speculate that this may be the underlying reason that accounts for the deterioration of performance of additional feature selection prior to RF.

The experimental results listed in Table 3 suggest that whether the feature selection will help to improve prediction performance depends not only on the redundancy of data itself but also on the underlying classifier used.

3.3 Comparisons with Existing Predictors

According to the experimental results listed in Tables 2 and 3, it can be seen that performing the disulfide connectivity prediction with PSSM, PSS, CM, DOC, and PDTCR features and RF regression model is a better choice, at least based on the two considered benchmark datasets. In view of this observation, unless otherwise stated, all of the subsequent results listed below are obtained by taking the combination of PSSM, PSS, CM, DOC, and PDTCR as input features and RF as the regression model, and the corresponding method will be termed TargetDisulfide in the subsequent descriptions.

3.3.1 Cross-Validation Comparisons

Tables 4 and 5 compare the performances between the proposed TargetDisulfide and several other popular predictors on SP39 and PDBCYS-R over four-fold and 20-fold cross-validations, respectively. The results on SP39 listed in Table 4 demonstrate that the proposed TargetDisulfide significantly outperforms the other five considered predictors, i.e., DISULFIND [32], Pair-wise SVM [33], GASVM [21], SS_SVR [22], and FS_SVR [23], and acts as the best performer. Improvements of 6.5 and 5.4 percent for Q_P and Q_C were observed if they were compared with the second-best performer, i.e., FS_SVR [23]. Similarly, remarkable improvements were also achieved by TargetDisulfide on PDBCYS-R. For example, TargetDisulfide achieved over a 12 percent improvement for both Q_P and Q_C if it was compared with DISLOCATE [25]. In addition, improvements of 8.4 and 6.1 percent for Q_P and Q_C were also achieved by TargetDisulfide if it was compared with the second-best performer, i.e., DMC [26], which is one of the most recently released predictors for protein disulfide connectivity prediction.

3.3.2 Independent Validation Test

Validating a predictor with fresh independent data has been considered to be a mandatory procedure to objectively evaluate its generalization capability [45], [64], [65]. In view of this, the independent validation dataset, IVD-54, that we constructed using stringent inclusion criteria was taken to compare the generalization performance between the proposed TargetDisulfide and several other popular predictors, including DiANNA [38], DISULFIND [32], and EDBCP [24]. Note that the results of DiANNA [38], DISULFIND [32], and EDBCP [24] were obtained by feeding the sequences in IVD-54 to their corresponding web servers. In addition, we obtained the results of TargetDisulfide with two different strategies: in the first strategy, we assumed that knowledge of the bonding states for the cysteines in a test sequence is not available, and the bonding states of cysteines predicted by DiANNA [38], EDBCP [24], and DISULFIND [32] were utilized by TargetDisulfide for further disulfide connectivity pattern predictions; whereas in the second strategy, TargetDisulfide performed the connectivity pattern prediction according to the observed bonding states of the cysteines in a test sequence. Table 6 summarizes the performance comparisons between TargetDisulfide and the other considered predictors on the independent validation dataset IVD-54.

The results of the comparisons between DiANNA and TargetDisulfide (column (A) in Table 6), EDBCP and TargetDisulfide (column (B) in Table 6), DISULFIND and TargetDisulfide (column (C) in Table 6) demonstrate the superiority of the proposed TargetDisulfide on the independent validation dataset IVD-54. For example, the proposed

TABLE 5
Performance Comparisons between TargetDisulfide and Other Predictors on the PDBCYS-R Dataset over 20-Fold Cross-Validation

B	DISLOCATE [25]		DMC [26]		TargetDisulfide	
	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
2	75	75	76.0	76.0	83.0	83.0
3	48	60	55.3	62.8	76.4	82.4
4	44	57	51.2	67.7	53.7	65.9
5	19	46	32.4	58.9	21.6	52.4
Overall	54	60	59.3	66.2	**67.7**	**72.3**

TABLE 6
Performance Comparisons between TargetDisulfide and Other Predictors on the Independent Validation Dataset IVD-54

No. of Disulfide Bonds	No. of Sequences	DiANNA [38]		EDBCP [24]		DISULFIND [32]		TargetDisulfide (A)		TargetDisulfide (B)		TargetDisulfide (C)		TargetDisulfide (D)	
		Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C	Q_P	Q_C
2	29	10.3	20.7	10.3	22.4	13.8	19.0	13.8	24.1	6.9	25.9	13.8	15.5	69.0	69.0
3	15	0.0	22.2	13.3	20.0	13.3	20.0	20.0	37.8	13.3	28.9	13.3	28.9	66.7	75.6
4	7	0.0	17.9	0.0	10.7	0.0	14.3	14.3	28.6	28.6	35.7	14.3	28.6	14.3	39.3
5	3	0.0	0.0	0.0	33.3	0.0	6.7	0.0	20.0	0.0	6.7	0.0	6.7	0.0	40.0
Overall	54	5.6	18.5	9.3	20.5	11.1	17.8	**14.8**	**28.8**	**11.1**	**26.7**	**13.0**	**21.2**	57.4	62.3

(A): Results obtained with the predicted bonding state from DiANNA;
(B): Results obtained with the predicted bonding state from EDBCP;
(C): Results obtained with the predicted bonding state from DISULFIND;
(D): Results obtained with the observed knowledge of bonding state.

TargetDisulfide significantly outperformed DiANNA [38], and the improvements over 9 and 10 percent for Q_P and Q_C were achieved, respectively; TargetDisulfide achieved slightly performance for Q_P to that of EDBCP (i.e., 11.1 percent versus 9.3 percent), whereas it demonstrated a much better performance for Q_C: the overall Q_C of TargetDisulfide is 26.7 percent, which is 6.2 percent higher that of EDBCP; TargetDisulfide also remarkably exceeds DISULFIND, with improvements of 1.9 and 3.4 percent for the overall Q_P and Q_C, respectively. All of these comparison results demonstrate the good generalization capability of the proposed TargetDisulfide.

Furthermore, we find that the prediction performances were significantly improved to 57.4 and 62.3 percent for the overall Q_P and Q_C by replacing the predicted bonding states with the observed knowledge of the bonding states of cysteines, which suggests that developing accurate bonding state prediction methods is a very important step and promising route for further improving the performance of protein disulfide connectivity predictions.

As described in Section 2.2, the PDTCR feature extracted from a modeled 3D structure with higher reliability will be more accurate. However, to objectively evaluate the effectiveness of the PDTCR feature, previous experimental results listed in Section 3.1, 3.2, and 3.3 are performed by restricting the homology identity of the templates to be 40 percent. For reference, we also evaluated the performance by using only the PDTCR feature. In addition, the performances of the proposed method under different thresholds, from 30 to 70 percent with a step size of 10 percent, of homology identity for locating templates are also provided. Results on all the datasets of various thresholds demonstrate the superiority of the proposed method. Refer to Table R1, R2, and R3 in Supplemental Material 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2359451, for more details.

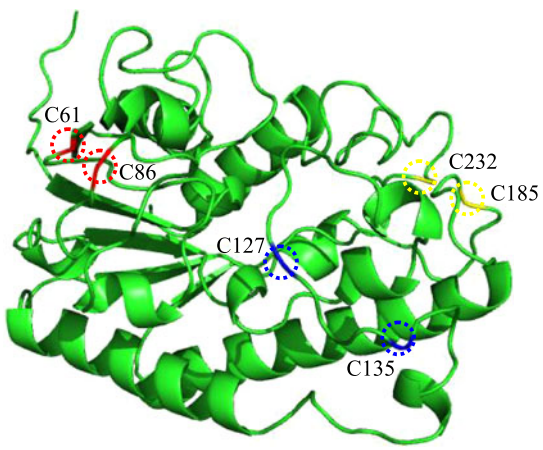


Fig. 2. Disulfide bonds of LIP_STRRM predicted by the proposed TargetDisulfide. Three bonds, i.e., C61-C86, C127-C135, and C185-C232, were correctly predicted and are highlighted in red, blue, and yellow, respectively, under using both the with- and without-PDTCR feature. The 3D structure of the LIP_STRRM was modelled by MODELLER software [6]. The pictures were made with PyMOL [66].

3.5 Case Studies

In this section, two protein sequences (Entry names: LIP_STRRM and FUT7_HUMAN) in the independent validation dataset IVD-54, which have not been used for training our TargetDisulfide, were taken for case study.

In protein LIP_STRRM, there are three observed disulfide bonds, i.e., C61-C86, C127-C135, and C185-C232, as shown in Fig. 2. TargetDisulfide trained on the SP39 dataset can correctly predict all of the three disulfide bonds under both the with- and without-PDTCR feature, denoting that LIP_STRRM is an easy target for TargetDisulfide to perform a disulfide connectivity prediction.

However, among the three observed disulfide bonds in protein FUT7_HUMAN, only one (i.e., C318-C321, highlighted in blue) was correctly predicted, and the other two (i.e., C68-C76 and C211-C214) were mistakenly predicted as C68-C214 and C76-C211 (highlighted in red and blue, respectively) as shown in Fig. 3a) by the TargetDisulfide using the without-PDTCR feature. By carefully observing the mistakenly predicted cysteine pair C68-C214, we found that the spatial distance between the two residues in it (refer to Fig. 3) is far larger than that in any other potential cysteine pairs; thus, the likelihood of being an actual disulfide bond is very low for C68-C214. Hence, the prediction performance can potentially be improved by incorporating the spatial distance information encoded in the PDTCR feature. As we speculated, all three bonds of FUT7_HUMAN were correctly predicted by TargetDisulfide after incorporating the PDTCR feature, as shown in Fig. 3b.

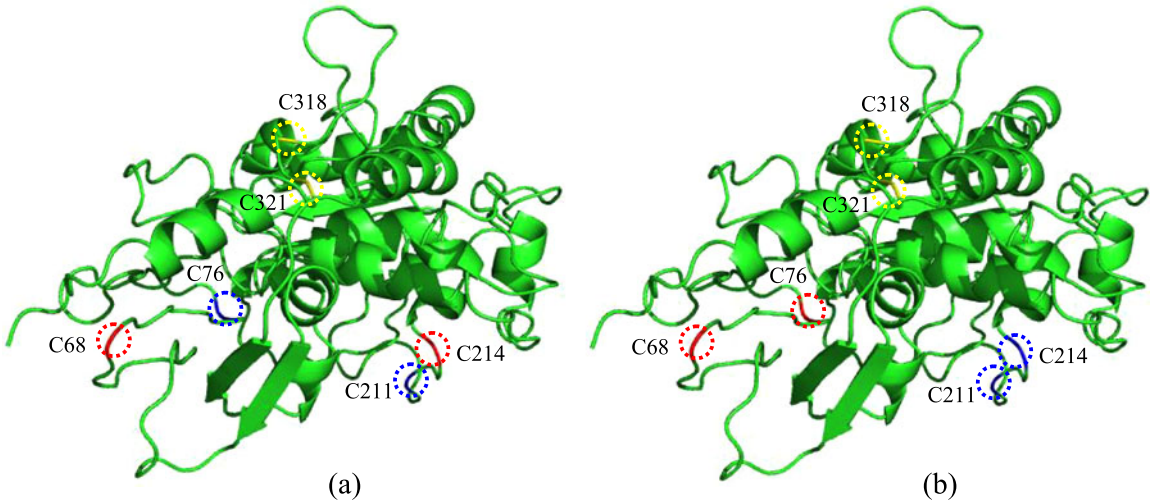


Fig. 3. Disulfide bonds of FUT7_HUMAN predicted by the proposed TargetDisulfide. (a) Two bonds, i.e., C68-C76 and C211-C214, were mistakenly predicted as C68-C214 and C76-C211 (highlighted in red and blue, respectively) using the without-PDTCR feature. (b) Three bonds, i.e., C68-C76, C211-C214, and C318-C321, were correctly predicted and are highlighted in red, blue, and yellow, respectively, under using the with-PDTCR feature. The 3D structure of the FUT7_HUMAN was modelled by MODELLER software [6]. The pictures were made with PyMOL [66].

4 Conclusions

In this study, a new predictor, TargetDisulfide, is proposed for protein disulfide connectivity predictions. It was found that the prediction performance can be significantly improved by integrating the newly developed PDTCR feature with the traditional features. Cross-validation and independent validation comparisons on different benchmark datasets demonstrate that the proposed TargetDisulfide can achieve good performance and outperforms many existing sequence-based predictors. To improve the applicability of the proposed method, we provide a web-server implementation that is freely available for academic use at http://csbio.njust.edu.cn/bioinf/TargetDisulfide. Our future work will focus on further improving the protein disulfide connectivity prediction performance by uncovering new effective features and by applying powerful classification algorithms.

Acknowledgments

The authors would like to thank the anonymous reviewers for suggestions and comments which helped improve the quality of this paper. This work was supported by the National Natural Science Foundation of China (No. 61373062, 61175024, 61222306, 61100116, and 61233011), The Natural Science Foundation of Jiangsu (No. BK20141403), Jiangsu Postdoctoral Science Foundation (No. 1201027C), China Postdoctoral Science Foundation (No. 2013M530260, 2014T70526), "The Six Top Talents" of Jiangsu Province (No. 2013-XXRJ-022), the Fundamental Research Funds for the Central Universities (No. 30920130111010). Dong-Jun Yu and Hong-Bin Shen are the corresponding authors for this paper.

References

[1] G. Casari, C. Sander, and A. Valencia, "A method to predict functional residues in proteins," *Nature Struct. Biol.*, vol. 2, no. 2, pp. 171–178, Feb. 1995.
[2] C. B. Anfinsen, "Studies on the principles that govern the folding of protein chains," *Nobel Lecture*, Dec. 1972.
[3] J. Garnier, "Protein structure prediction," *Biochimie*, vol. 72, no. 8, pp. 513–524, 1990.
[4] D. R. Westhead and J. M. Thornton, "Protein structure prediction," *Current Opinion Biotechnol.*, vol. 9, no. 4, pp. 383–389, 1998.
[5] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: A unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010.
[6] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali, "Comparative protein structure modeling using Modeller," *Current Protocols Bioinformatics*, pp. 15:5.6:5.6.1–5.6.30, 2006.
[7] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, no. 1, p. 40, 2008.
[8] M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani, and L. Kurgan, "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources," *Bioinformatics*, vol. 26, no. 18, pp. i489–i496, Sep. 15, 2010.
[9] S. J. Metallo, "Intrinsically disordered proteins are potential drug targets," *Current Opinion Chem. Biol.*, vol. 14, no. 4, pp. 481–488, 2010.
[10] M. E. Oates, P. Romero, T. Ishida, M. Ghalwash, M. J. Mizianty, B. Xue, Z. Dosztányi, V. N. Uversky, Z. Obradovic, and L. Kurgan, "D2P2: Database of disordered protein predictions," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D508–D516, 2013.
[11] J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky, and L. Kurgan, "RAPID: Fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale," *Biochimica et Biophys. Acta (BBA)-Proteins Proteomics*, vol. 1834, no. 8, pp. 1671–1680, 2013.
[12] T. Nugent and D. T. Jones, "Transmembrane protein topology prediction using support vector machines," *BMC Bioinformatics*, vol. 10, no. 1, p. 159, 2009.
[13] D. Yu, H. Shen, and J. Yang, "SOMRuler: A novel interpretable transmembrane helices predictor," *IEEE Trans. NanoBiosci.*, vol. 10, no. 2, pp. 121–129, 2011.
[14] Z. Aydin, Y. Altunbasak, and H. Erdogan, "Bayesian models and algorithms for protein β-sheet prediction," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 8, no. 2, pp. 395–409, Mar./Apr. 2011.
[15] F. E. Cohen, M. Sternberg, and W. R. Taylor, "Analysis and prediction of protein beta-sheet structures by a combinatorial approach," *Nature*, vol. 285, no. 5764, pp. 378–382, 1980.
[16] K. D. Kedarisetti, M. J. Mizianty, S. Dick, and L. Kurgan, "Improved sequence-based prediction of strand residues," *J. Bioinformatics Comput. Biol.*, vol. 9, no. 1, pp. 67–89, 2011.
[17] S. Wu, A. Szilagyi, and Y. Zhang, "Improving protein structure prediction using multiple sequence-based contact predictions," *Structure*, vol. 19, no. 8, pp. 1182–1191, 2011.
[18] M. Vassura, L. Margara, P. DiLena, F. Medri, P. Fariselli, and R. Casadio, "Reconstruction of 3D structures from protein contact maps," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 5, no. 3, pp. 357–367, Jul./Sep. 2008.

[19] J. Eickholt and J. Cheng, "Predicting protein residue-residue contacts using deep networks and boosting," *Bioinformatics*, vol. 28, no. 23, pp. 3066–3072, 2012.

[20] M. M. Gromiha, "Influence of long-range contacts and surrounding residues on the transition state structures of proteins," *Anal. Biochem.*, vol. 408, no. 1, pp. 32–36, 2011.

[21] C. H. Lu, Y. C. Chen, C. S. Yu, and J. K. Hwang, "Predicting disulfide connectivity patterns," *Proteins: Struct., Funct. Bioinformatics*, vol. 67, no. 2, pp. 262–270, 2007.

[22] J. Song, Z. Yuan, H. Tan, T. Huber, and K. Burrage, "Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure," *Bioinformatics*, vol. 23, no. 23, pp. 3147–3154, 2007.

[23] L. Zhu, J. Yang, J. N. Song, K. C. Chou, and H. B. Shen, "Improving the accuracy of predicting disulfide connectivity by feature selection," *J. Comput. Chem.*, vol. 31, no. 7, pp. 1478–1485, May 2010.

[24] H.-H. Lin and L.-Y. Tseng, "DBCP: A web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines," *Nucleic Acids Res.*, vol. 38, no. suppl. 2, pp. W503–W507, 2010.

[25] C. Savojardo, P. Fariselli, M. Alhamdoosh, P. L. Martelli, A. Pierleoni, and R. Casadio, "Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization," *Bioinformatics*, vol. 27, no. 16, pp. 2224–2230, Aug. 15, 2011.

[26] C. Savojardo, P. Fariselli, P. L. Martelli, and R. Casadio, "Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations," *BMC Bioinformatics*, vol. 14, no. suppl. 1, p. S10, 2013.

[27] C. Mirabello and G. Pollastri, "Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility," *Bioinformatics*, vol. 29, no. 16, pp. 2056–2058, 2013.

[28] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *J. Comput. Chem.*, vol. 33, no. 3, pp. 259–267, 2012.

[29] K. Inaba, S. Murakami, M. Suzuki, A. Nakagawa, E. Yamashita, K. Okada, and K. Ito, "Crystal structure of the DsbB-DsbA complex reveals a mechanism of disulfide bond generation," *Cell*, vol. 127, no. 4, pp. 789–801, 2006.

[30] H. Kadokura, H. Tian, T. Zander, J. C. Bardwell, and J. Beckwith, "Snapshots of DsbA in action: Detection of proteins in the process of oxidative folding," *Science*, vol. 303, no. 5657, pp. 534–537, 2004.

[31] K. Inaba, "Structural basis of protein disulfide bond generation in the cell," *Genes Cells*, vol. 15, no. 9, pp. 935–43, Sep. 1, 2010.

[32] A. Ceroni, A. Passerini, A. Vullo, and P. Frasconi, "DISULFIND: A disulfide bonding state and cysteine connectivity prediction server," *Nucleic Acids Res.*, vol. 34, no. suppl. 2, pp. W177–W181, 2006.

[33] C. H. Tsai, B. J. Chen, C. H. Chan, H. L. Liu, and C. Y. Kao, "Improving disulfide connectivity prediction with sequential distance between oxidized cysteines," *Bioinformatics*, vol. 21, no. 24, pp. 4416–4419, 2005.

[34] J. Song, Z. Yuan, H. Tan, T. Huber, and K. Burrage, "Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure," *Bioinformatics* vol. 23, no. 23, pp. 3147–3154, 2007.

[35] R. Rubinstein and A. Fiser, "Predicting disulfide bond connectivity in proteins by correlated mutations analysis," *Bioinformatics*, vol. 24, no. 4, pp. 498–504, 2008.

[36] C. H. Tsai, C. H. Chan, B. J. Chen, C. Y. Kao, H. L. Liu, and J. P. Hsu, "Bioinformatics approaches for disulfide connectivity prediction," *Current Protein Peptide Sci.*, vol. 8, no. 3, pp. 243–260, Jun. 2007.

[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[38] F. Ferrè and P. Clote, "DiANNA 1.1: An extension of the DiANNA web server for ternary cysteine classification," *Nucleic Acids Res.*, vol. 34, no. suppl 2, pp. W182–W185, 2006.

[39] P. Fariselli and R. Casadio, "Prediction of disulfide connectivity in proteins," *Bioinformatics*, vol. 17, no. 10, pp. 957–964, Oct. 2001.

[40] C. Savojardo, P. Fariselli, M. Alhamdoosh, P. L. Martelli, A. Pierleoni, and R. Casadio, "Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization," *Bioinformatics*, vol. 27, no. 16, pp. 2224–2230, 2011.

[41] C. O'Donovan and R. Apweiler, "A guide to UniProt for protein scientists," *Methods Molecular Biol.*, vol. 694, pp. 25–35, 2011.

[42] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 1, 2006.

[43] J. Becker, F. Maes, and L. Wehenkel, "On the relevance of sophisticated structural annotations for disulfide connectivity pattern prediction," *PLoS One*, vol. 8, no. 2, p. e56621, 2013.

[44] A. A. Schaffer, "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Res.*, vol. 29, pp. 2994–3005, 2001.

[45] D. Yu, J. Hu, J. Yang, H. Shen, and J. Tang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 10, no. 4, pp. 994–1008, Jul./Aug. 2013.

[46] D. J. Yu, J. Hu, Z. M. Tang, H. B. Shen, J. Yang, and J. Y. Yang, "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing*, vol. 104, pp. 180–190, 2013.

[47] J. Song, K. Burrage, Z. Yuan, and T. Huber, "Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information," *BMC Bioinformatics*, vol. 7, no. 1, p. 124, 2006.

[48] J. Song and K. Burrage, "Predicting residue-wise contact orders in proteins by support vector regression," *BMC Bioinformatics*, vol. 7, no. 1, p. 425, 2006.

[49] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 17, 1999.

[50] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, vol. 6, Amsterdam, The Netherlands: North Holland, 1986 , pp. 233–243.

[51] S. Mukherjee and S. Mitra, "Hidden Markov models, grammars, and biology: A tutorial," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 491–526, Apr. 2005.

[52] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[53] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training SVM," *J. Mach. Learn. Res.* vol. 6, pp. 1889–1918, 2005.

[54] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[55] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 1, 2007.

[56] M. G. Tadesse, M. Vannucci, and P. Liò, "Identification of DNA regulatory motifs using Bayesian variable selection," *Bioinformatics*, vol. 20, no. 16, pp. 2553–2561, Nov. 1, 2004.

[57] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouzé, and Y. Van de Peer, "Feature selection for splice site prediction: A new method using EDA-based feature ranking," *BMC Bioinformatics*, vol. 5, no. 1, p. 64, 2004.

[58] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, Mar. 1, 2005.

[59] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–e190, Jul. 15, 2006.

[60] O. D. Richard, E. H. Peter, and G. S. David, *Pattern Classification*, 2nd ed., New York, NY, USA: Wiley, 2001.

[61] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 507–514.

[62] H. Yan and J. Yang, "Joint Laplacian feature weights learning," *Pattern Recognit.*, vol. 47, no. 3, pp. 1425–1432, 2014.

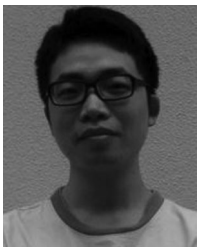
[63] M. Šikić, S. Tomić, and K. Vlahoviček, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLoS Comput. Biol.*, vol. 5, no. 1, p. e1000278, 2009.

[64] A. L. Boulesteix, "Over-optimism in bioinformatics research," *Bioinformatics*, vol. 26, no. 3, pp. 437–439, Feb. 1, 2010.

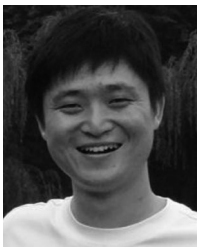
[65] T. D. Sterlinga, "Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa," *J. Amer. Statist. Assoc.*, vol. 54, no. 285, pp. 30–34, 1959.

[66] The PyMOL molecular graphics system [Online]. Available: http://www.pymol.org, 2001.

Dong-Jun Yu received the BS degree in computer science and the MS degree in artificial intelligence from the Jiangsu University of Science and Technology, in 1997 and 2000, respectively, and the PhD degree in pattern analysis and machine intelligence from the Nanjing University of Science and Technology in 2003. In 2008, he was an academic visitor at the University of York in United Kingdom. He is currently a professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include bioinformatics, pattern recognition, and data mining. He is a member of the IEEE and CCF.



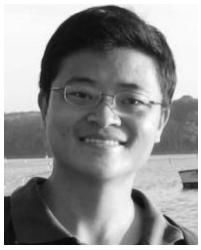
Yang Li received the BS degree in computer science from the Nanjing University of Science and Technology, China in 2014. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, data mining, and pattern recognition.



Jun Hu received the BS degree in computer science from Anhui Normal University, China in 2011. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include bioinformatics, data mining, and pattern recognition.

Xibei Yang received the BS degree from Xuzhou Normal University, China, in 2002, the MS degree from the Jiangsu University of Science and Technology, China, in 2006, and the PhD degree from the Nanjing University of Science and Technology, China, in 2010, both in computer applications. He is currently an associate professor at the Jiangsu University of Science and Technology, and also a postdoctoral researcher at the Nanjing University of Science and Technology. He has published more than 30 articles in international journals and international conferences. His research interests include pattern recognition and decision making.

Jing-Yu Yang received the BS degree in computer science from NUST, Nanjing, China. From 1982 to 1984 he was a visiting scientist at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. From 1993 to 1994 he was a visiting professor in the Department of Computer Science, Missouri University. And, in 1998, he was a visiting professor at Concordia University, Canada. He is currently a professor and chairman in the Department of Computer Science, NUST. He is the author of more than 150 scientific papers in computer vision, pattern recognition, and artificial intelligence. He has won more than 20 provincial awards and national awards. His current research interests are in the areas of pattern recognition, robot vision, image processing, data fusion, and artificial intelligence.



Hong-Bin Shen received the PhD degree from Shanghai Jiaotong University, China, in 2007. He was a postdoctoral research fellow of Harvard Medical School from 2007 to 2008. He is currently a professor at the Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University. His research interests include data mining, pattern recognition, and bioinformatics. He has published more than 60 papers and constructed 20 bioinformatics severs in these areas and he is the editorial members of several international journals.