PROTEINS: Structure, Function, and Bioinformatics 64:246–252 (2006)

Disulfide Connectivity Prediction with 70% Accuracy Using Two-Level Models

Bo-Juen Chen,[1] Chi-Hung Tsai,[1] Chen-hsiung Chan,[1] and Cheng-Yan Kao[1,2*]
[1]*Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan*
[2]*Institute for Information Industry, Taipei, Taiwan*

ABSTRACT **Disulfide bridges stabilize protein structures covalently and play an important role in protein folding. Predicting disulfide connectivity precisely helps towards the solution of protein structure prediction. Previous methods for disulfide connectivity prediction either infer the bonding potential of cysteine pairs or rank alternative disulfide bonding patterns. As a result, these methods encode data according to cysteine pairs (pair-wise) or disulfide bonding patterns (pattern-wise). However, using either encoding scheme alone cannot fully utilize the local and global information of proteins, so the accuracies of previous methods are limited. In this work, we propose a novel two-level framework to predict disulfide connectivity. With this framework, both the pair-wise and pattern-wise encoding schemes are considered. Our models were validated on the datasets derived from SWISS-PROT 39 and 43, and the results demonstrate that our models can combine both local and global information. Compared to previous methods, significant improvements were obtained by our models. Our work may also provide insights to further improvements of disulfide connectivity prediction and increase its applicability in protein structure analysis and prediction. Proteins 2006;64:246–252.**

Key words: disulfide connectivity; support vector machine; disulfide bonding pattern; hierarchical model

INTRODUCTION

When a protein folds into its native structure, two cysteine residues within a reasonable distance can be oxidized to form a disulfide bond. This kind of covalent bonding is commonly found in extracellular proteins, and is able to stabilize their conformations as it contributes to the stability of the three dimensional structures from the thermodynamics aspect.[1] It has been further described as a reduction of the conformational entropy of an unfolded polypeptide chain, leading to the transition from the unfolded state to the native state.[2] Furthermore, disulfide bonds impose length and angle constraints on the backbone of a protein. Therefore, the information of disulfide connectivity can be employed to reduce the search in conformational space dramatically and raise the accuracy for the prediction of the structure of a folded protein greatly.[3]

Usually there are two steps in predicting the disulfide connectivity patterns. First, the bonding state of each cysteine residue in a protein is inferred. Various algorithms with statistical methods[4] or machine-learning techniques[5] have been developed for this purpose. In 1999, Fariselli and coworkers[6] used neural network (NN) to obtain 80% accuracy by introducing evolutionary information (i.e., profiles extracted from multiple sequence alignments). With the promising results from evolutionary information, more methods based on NN[7,8] were further employed to improve the prediction accuracy. In 2004, a support vector machine[9] (SVM) was used to achieve an extraordinary accuracy of 90% for bonding states prediction.[10]

Once the bonding states of cysteine residues are known, the second step is to predict the bonding patterns of disulfide bridges. This article focuses on this stage with the prior knowledge of the oxidization states of cysteine residues. Different methods[11–18] have been developed to solve this problem. These methods can be classified into two categories: pair-wise or pattern-wise. The major difference between them is whether the methodology is developed to deal with the relation between cysteine pairs or the disulfide connectivity patterns alternatively. Methods classified as "pair-wise" focus on the bonding potential between two cysteines, whereas the "pattern-wise" ones rank connectivity patterns.

Focusing on the local environment of cysteine residues, the pair-wise methods attempt to predict the bonding between two cysteines. Stochastic global optimization[11] and neural networks[12] were adopted to assign contact potentials between cysteine pairs. Recently, similar formulation was employed by Baldi et al.,[15] Ferrè et al.,[16] and Tsai et al.,[18] with the bonding potentials output by variant NNs or SVMs. However, as the pair-wise models mainly consider the local information, the global knowledge of a protein is not easy to be utilized. As a result, the information provided for these methods is usually limited to the scope of local environments of cysteines, lacking the over-

Grant sponsor: the National Science Council; Grant number: NSC 94-2320-B-002-124.

*Correspondence to: Cheng-Yan Kao, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 106, Republic of China. E-mail: cykao@csie.ntu.edu.tw

Received 24 October 2005; Revised 30 December 2005; Accepted 5 February 2006

Published online 13 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20972

TABLE I. The Number of Sequences in the Datasets, Divided According to the Number of Disulfide Bridges (B)

Datasets	B = 2	B = 3	B = 4	B = 5	B = 2. . .5
SP39[a]	156	146	99	45	446
SP39-template[b]	243	198	97	45	583
SP43[b]	119	116	43	35	313

[a]Dataset from the work of Vullo and Frasconi.[13]
[b]Dataset from Zhao et al.[14]

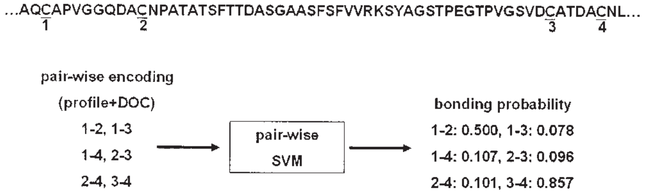


Fig. 1. Level-1: pair-wise model infers the bonding potential between cysteine residues.

view of the whole protein. Because a disulfide bridge is a long-range interaction between two sequentially distant cysteines, the performance of the pair-wise models built with only local information is restricted.

On the other hand, the pattern-wise approaches take a whole protein as a unit directly. The global information of a protein, such as the sequence length, amino acid contents, or the positions of all cysteines can be easily included into the data encoding. In 2004, recursive neural networks (RNN) were developed by Vullo and Frasconi[13] to score labeled undirected graphs that represent connectivity patterns. Other pattern-wise models, such as CSP proposed by Zhao et al.[14] and SVM by Chen and Hwang,[17] were also employed to predict disulfide connectivity. However, for these pattern-wise methods, the local information of cysteine residues is difficult to be fully examined. Moreover, the pattern-wise methods may suffer from the imbalance of data during the training phase (see the Material and Methods section). All these issues limited the accuracy of pattern-wise methods to around 50%.

In this article, a novel approach with two-level hierarchical framework was proposed to combine both the pair-wise and pattern-wise methods. In the first level, our models focus on the local relations of cysteine residues, whereas the second level incorporates the global information of proteins. Our models were validated with three datasets derived from SWISS-PROT[19] version nos. 39 and 43. For the dataset derived from SWISS-PROT 39, our method achieved a remarkable accuracy of 70%, which outperforms previous approaches. These results show our models take advantages of both pair-wise and pattern-wise methods and utilize the local and global information to achieve the optimal performance.

MATERIALS AND METHODS

Dataset

Three datasets, extracted from SWISS-PROT[19] releases 39 and 43, were used to evaluate the predicting power of our method. Because less than 20% of the filtered SWISS-PROT sequences in our datasets contain more than five disulfide bridges,[13] the experiments were conducted against the sequences with two to five disulfide bonds. The number of sequences in each dataset is summarized in Table I.

To compare our method to other approaches,[13,15] the same dataset extracted from SWISS-PROT database release no. 39 was employed (denoted as SP39). The same filtering procedure[11] was applied to ensure only high-quality and experimentally verified intrachain disulfide bridge annotations were included. Only the sequences containing information in the Protein Data Bank (PDB) were included in the filtered dataset. In addition, sequences with disulfide annotation described as “probable,” “potential,” or “by similarity” were excluded. For crossvalidation, this dataset was further divided into four subsets so that each two of them shared a sequence homology of $\leq 30\%$. Each subset contained an approximately equal number of sequences.

To further validate the predicting power of our method, the holdout-prediction assessment employed by Zhao et al.[14] was also adopted. Two datasets, SP43 and SP39-template, were built for this purpose. With the aforementioned filter, SP43 was extracted from SWISSPROT release no. 43, where sequences in release 39 were excluded. This dataset was further filtered so that the sequences share less than 25% identity with each other. SP39-template was built from SWISSPROT release 39, with the same filter except for the PDB match. Furthermore, SP39-template was filtered to contain only sequences that share similarity $<30\%$ with SP43. To assess the test accuracy, our model was trained with SP39-template to predict SP43, which contains the newly added proteins since release no. 39.

Method

In this work, a two-level framework is proposed to predict disulfide connectivity. Because the two aforementioned encoding schemes have different strength and weakness, the two-level framework we proposed attempts to integrate them to achieve better performance. The idea of the two-level framework is to extend the modeling from a local view (pair-wise) to a global perspective (pattern-wise). In the first level, our models focus on the relation between two cysteine residues. The resulting models infer the bonding probabilities for cysteine pairs (Fig. 1). On the other hand, the level-2 models take the pattern-wise encoding to solve the problem. In this level, the models take the results from the level-1 along with other global information of proteins to predict disulfide connectivity (Fig. 2). In addition, the framework proposed here is flexible; different machine-learning techniques, such as NN or SVM, can also be applied in both levels. In this article, we used SVM[9] and its probability output[20] in both levels.

Level-1: Pair-wise

In the first level, SVM models infer the bonding potential between two cysteines. Figure 1 illustrates the pair-wise modeling. Given a protein with oxidized cysteines,

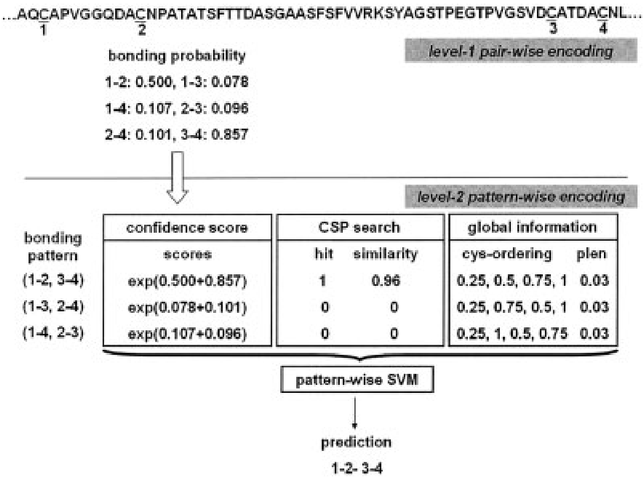


Fig. 2. Two-level framework: from pair-wise to pattern-wise encoding, models integrate various data to generate prediction. In CSP search, "hit" represents the returned bonding pattern; in global information, "cys-ordering" denotes cysteine-ordering, and "plen" stands for protein length (after scaling).

the data were first encoded with respect to each possible cysteine pair. For the protein with B disulfide bonds, there are B(2B-1) combinations of cysteines pairs to be encoded. (Note that although a cysteine pair can be symmetrically encoded as (i,j) or (j,i), where $i < j$, we simply used (i,j) for our models.) With these encoded data, SVM models were trained to infer the bonding probabilities for cysteine pairs.

To encode the data with respect to cysteine pairs, two descriptors were considered: (1) local sequence profiles (evolutionary information) around target cysteines from multiple sequence alignments, and (2) the sequential distance between oxidized cysteines (denoted as DOC).

Profiles

The sequence profiles were generated by performing multiple sequence alignments with PSI-BLAST.[21] For each cysteine pair *Cys* (i,j), profiles were extracted using a window centered at cysteines i and j. Each residue in the sequence window was encoded as a vector of 20 elements, which were extracted from the position specific scoring matrix generated by PSI-BLAST. The window size used in this work was set to 13. Therefore, for a cysteine pair, there are 520 features containing the evolutionary information.

DOC

Because the sequence separation between bonded cysteines correlates with specific connectivity patterns,[2] another feature used to encode data is the linear distance between two cysteines. In this article, DOC is defined as $\|i - j\|$, where i and j are the sequence indices of two cysteines. Moreover, as observed in our analysis, we found there are fewer instances of connectivity if DOC is larger than 100. For example, there are only 6.6% of bonded cysteine pairs with the sequential distance longer than 100 in SP39. Therefore, to encode the sequential distance into our data, the normalization min(DOC/100, 1) was applied.

Many previous works have employed pair-wise encoding to solve the disulfide connectivity.[13,14,17,18] With the bonding probabilities generated from the models, previous methods transform the disulfide connectivity problem to an undirected complete graph, where oxidized cysteines are considered as vertices and the probabilities of connectivity between cysteine pairs are assigned as the weights of the edges between corresponding vertices. The disulfide connectivity pattern can be solved by finding the maximum weight matching of this graph. However, this technique was not applied in this work, because our concern in level-1 is the bonding potential between cysteine residues. Instead of finding the disulfide bonding patterns in level-1, we generated the bonding probabilities between cysteines so that they can be further encoded in the second level of our framework.

Level-2: Pattern-wise

In the second level, we considered the pattern-wise encoding to tackle the problem from a global perspective. An illustration of our two-level framework is in Figure 2. For each protein, all possible disulfide bonding patterns were generated for encoding. Three descriptors were considered to encode disulfide bonding patterns: (1) the confidence scores from the level-1 SVM, (2) the results of CSP search,[14] and (3) the global information of the protein. These descriptors are elaborated in the following.

Confidence scores from pair-wise SVM (denoted as S)

In level-1, each cysteine pair was assigned with a probability indicating the potential of bonding. The value of this probability can further be interpreted as how confident the level-1 model is toward the prediction. Let T_{ptn} be the set of cysteine pairs that comprise the disulfide bonding pattern *ptn*. For *ptn*, we can calculate a confidence score (S_{ptn}) from the bonding probabilities of cysteine pairs by

$$S_{ptn} = \exp(\sum_{t \in T_{ptn}} p(t)), \tag{1}$$

where t is a cysteine pair in T_{ptn}; $p(.)$ is the bonding probability generated in the level-1 SVM and exp(.) is the exponential function. In level-2, we use the confidence scores as measures for the probabilities of possible bonding patterns. In fact, if the problem is transformed into a graph in level-1, the predicted connectivity will be the one with the largest S_{ptn}.

CSP search (denoted as C)

CSP search[14] is a simple method based on the assumption that two proteins with similar cysteine separation share the same disulfide connectivity.[17,22] Given a protein with the positions of oxidized cysteines, this method searches the database for the protein that has the most resembled cysteine separation and returns its disulfide connectivity. The similarity of cysteine separation is decided by the one-norm distance. With CSP search, a predicted bonding pattern and the similarity of cysteine separation against the returned pattern can be acquired for each protein. Thus, in this level, we further used this information to encode data. We added a binary feature to indicate whether the encoded pattern is the predicted one from CSP. Following the binary, another feature was used to encode the similarity by

$$\left(1 + \log_{10}\left(1 + \frac{d}{10}\right)\right)^{-1}, \tag{2}$$

where d is the similarity computed during CSP search. These two features contain nonzero values only when the encoded bonding pattern is the predicted one from CSP search. Otherwise, these two features are set to 0.

Global information (denoted as G)

For the global information of a protein, two simple descriptors were used:

1. Cysteine ordering: to distinguish each possible disulfide pattern of a protein, the order of cysteines were encoded to represent the bonding. For example, if a protein has oxidized cysteines in positions 2, 5, 10, and 20, and is assumed to have 2–10 and 5–20 as disulfide bonds, we used the order of cysteines (1, 3, 2, 4) to represent the bonding pattern. Moreover, the values were normalized by the number of oxidized cysteines. Therefore, the pattern above was encoded as (0.25, 0.75, 0.5, 1). For a protein of unknown disulfide connectivity, all possible orderings were generated in the data pool for further prediction.
2. Protein length: because the length of a protein may affect the conformation,[2] it was also considered for data encoding. For each disulfide bonding pattern, the protein length was added as a feature. In addition, this feature was normalized by the maximum protein length in the dataset.

Using these descriptors, possible bonding patterns were encoded for our level-2 SVM models. Because the level-2 models employ pattern-wise encoding, the number of features vary according to the numbers of bonds. Therefore, SVM models were separately trained for proteins with different numbers of bonds in our experiments.

Reduction for Imbalance

A serious issue for the pattern-wise encoding is the imbalance between the number of positive and negative data. The imbalance issue is especially severe when the number of bonds is large. Taking a protein with five disulfide bonds as an example, there are 945 possible disulfide bonding patterns, but only one of them is the answer. This results in a positive/negative ratio of 1:944. Such an imbalance among the training data can attribute greatly to the difficulty of prediction.

To handle this issue, we took advantage of our level-1 SVM. As we used the level-1 SVM to generate confidence scores (S_{ptn}) for all the possible patterns, we can also use the confidence scores to ease the imbalance for the level-2 models. That is, using the confidence scores, we selected the top k possible patterns as candidates for the level-2 models. In our experiments, k is set to 15. For proteins with two bonds, there are only three possible disulfide bonding patterns included. Thus, for the proteins with more than two bonds, the imbalance ratio (positive/negative) was reduced to 1:14. One thing has to be clarified is that reducing the imbalance by candidate selection can also pose a risk of sacrificing the real answers. Because we only selected k patterns for each protein, the actual bonding pattern might not be included in the candidate set. This will be discussed later in more details.

TABLE II. Results in Terms of Q_p (%) of Cross-validation on the Data Extracted from SWISS-PROT no. 39

Methods	B = 2	B = 3	B = 4	B = 5	B = 2. . .5
MC graph-matching[a]	56	21	17	2	29
NN graph-matching[b]	68	22	20	2	34
BiRnn-2 profile[c]	73	41	24	13	44
2D-Rnn profile[d]	74	51	27	11	49
dNN2[e]	62	40	55	26	49
CSP	72	54	33	18	52
Pattern-wise SVM[f]	74	61	30	12	55
Pair-wise SVM[g]	79	53	55	58	63
2-level SVM	85	67	57	58	70

[a] Reported by Fariselli and Casadio.[11]
[b] Reported by Fariselli et al.[12]
[c] Reported by Vullo and Frasconi.[13]
[d] Reported by Baldi et al.[15]
[e] Reported by Ferrè and Clote.[16]
[f] Reported by Chen and Hwang.[17]
[g] Reported by Tsai et al.[18]

Evaluation

Evaluation of our models focuses on the metric Q_p computed by

$$Q_p = \frac{N_c}{N_t}, \tag{3}$$

where N_c is the number of proteins whose bonding patterns are correctly predicted, and N_t is the total number of proteins.

RESULTS AND DISCUSSION

In our experiments, LIBSVM[23] was used for SVM implementation. Different models were built for the datasets. Because parameters are important for SVM, their values were selected by cross-validation procedures. The results are summarized in Table II and Table III.

Validation with SP39 and SP43

The dataset SP39 was used for fourfold cross-validation. In Table II, the accuracies of our SVM models along with the results from previous works are listed. As observed in the table, our models outperformed other approaches. In previous methods, either pair-wise[11,12,15,16,18] or pattern-wise[13,14,17] models, the overall accuracy is limited. Our two-level SVM, combining both pair-wise and pattern-wise

TABLE III. Results in Terms of Q_p (%) for the Holdout Test on the Dataset SP43

Methods	B = 2	B = 3	B = 4	B = 5	B = 2. . .5
CSP	71	49	30	28	53
Pair-wise SVM[a]	77	52	53	29	59
2-level SVM	77	59	56	31	63

[a] Reported by Tsai et al.[18]

encoding, can reach an overall accuracy of 70%. Especially for B = 2, the two-level SVM can obtain an outstanding accuracy of 85%. Moreover, compared to previous methods, our SVM models also have remarkable improvement to predict the proteins with more than three disulfide bonds: the accuracy for B = 4 and 5 is above 57%.

Using the dataset SP39-template as the training dataset, we further tested our models with the prediction for SP43. The accuracies of our models and CSP search are listed in Table III. For this dataset, our SVM models show its predicting power for the newly added proteins with an accuracy of 63%. Compared to the simple CSP search, which can be categorized as a pattern-wise method, the accuracy was improved by 10% on average.

The result of SP43 is only for comparison with the CSP method. For fair comparison with other previous methods, the SP39 dataset should be considered. In the SP39-template, the sequences were not filtered by the PDB database, and therefore might contain incorrect disulfide annotation. Such sequences, in turn, contributed noise to the model and affected the testing accuracy. This kind of noisy information in the SP39-template might be the cause of the 7% difference in accuracy for SP39 and SP43.

Effects of Descriptors

To evaluate the influence of the three descriptors used, we further built different models for testing. The results are shown in Figure 3. In the figure, effects of three descriptors are shown, with "S" standing for the *confidence scores from pair-wise SVMs*, "C" for the *CSP search* and "G" for the *global information*. The combination "S + C + G" denotes the models with all three descriptors, whereas "S + C" denotes the models using the confidence scores and the CSP result. In addition, predictions made from pair-wise models and CSP search were also shown in the figure, denoted as "S" and "C."

Pair-wise relation from Level-1 (S)

The descriptor *S* provides much information for the two-level SVM models. As shown in the figure, the pair-wise models can achieve a fair result. When combined with *C* and *G*, the prediction accuracies were further improved. Additionally, we tried to exclude *S* from the data encoding (using only *C* and *G*) and found the accuracy was seriously affected. The reason is that when *S* was not used, our SVM became simply pattern-wise models. Without the confidence scores representing the local relations between cysteine residues, the models can only rely on the CSP search and the global information. As a result, the accuracy is merely better than that of the CSP search. The overall accuracy is 55 and 54% for SP39 and SP43, similar to the result of previous pattern-wise models.[17] From the results, we observed that *S* is crucial to our two-level SVM, because it can provide the pair-wise information from a local aspect, to compensate the global view of the pattern-wise encoding.

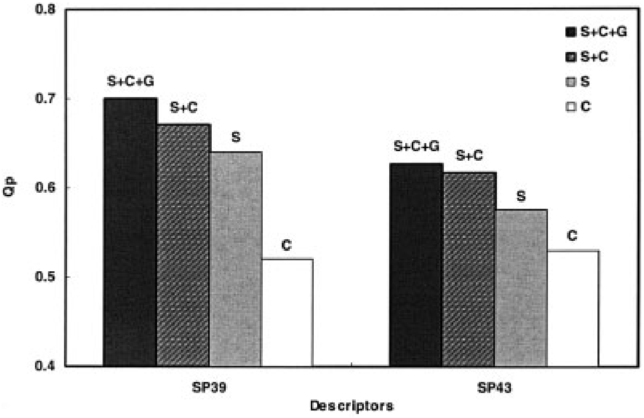


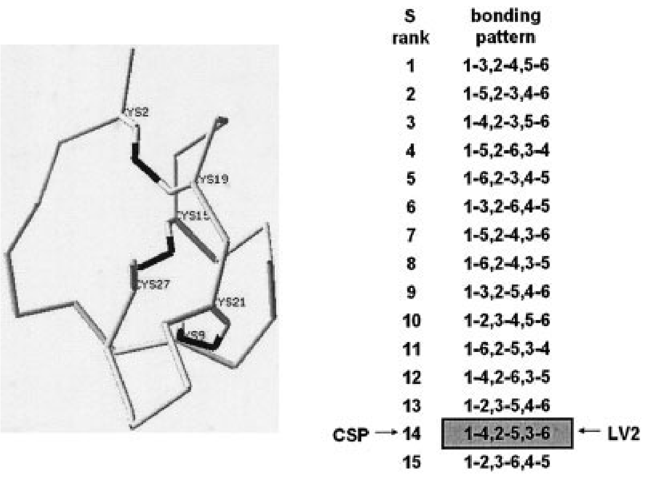
Fig. 3. Accuracies of models built with different descriptors (S: confidence scores from pair-wise SVM, C: CSP search, and G: global information): "S + C + G" represents models built with three descriptors; "S + C" is for models built with *S* and *C* information; "S" is for the predictions from the level-1 models, using maximum confidence scores to find bonding patterns; and "C" is for the results of CSP search.

CSP implication (C)

Figure 3 shows that inclusion of *C* enhanced the accuracies for both datasets. This indicates the information provided by *C* also has a major positive effect on the prediction. Moreover, *C* can provide complementary information to the confidence scores. For some sequences, their actual disulfide bonding patterns may not have high confidence scores from the pair-wise models. The models built with *S* and *G* may fail to predict the disulfide connectivity for these sequences. However, with the information provided by *C*, these bonding patterns may still be correctly predicted. One example is ITR2_ECBEL (PDB code: 2LET) in SP39. As shown in Figure 4, the actual disulfide connectivity is ranked as 14th among all the possible patterns according to the confidence scores from the level-1. If only *S* and *G* are used, the disulfide connectivity of this sequence cannot be correctly predicted. Nevertheless, because of the correct inference ([1–4, 2–5, 3–6]) by CSP search, our two-level SVM can predict the bonding pattern correctly.

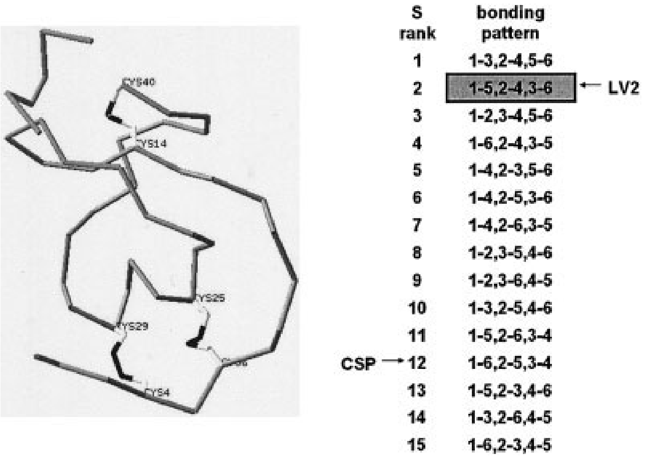
Global information (G)

As shown in Figure 3, adding the descriptor *G* also improved the prediction, compared to using only *S* and *C* ("S + C"). This suggests the descriptor *G* can supply further information to optimize our two-level SVM. In our experiments, we also found a few sequences whose predictions can be attributed to the global information. An example is shown in Figure 5 (LDTI_HIRME in SP39, PDB code: 1LDT). For this sequence, using the model built only with *S* and *C* cannot generate the correct disulfide



S rank	bonding pattern
1	1-3,2-4,5-6
2	1-5,2-3,4-6
3	1-4,2-3,5-6
4	1-5,2-6,3-4
5	1-6,2-3,4-5
6	1-3,2-6,4-5
7	1-5,2-4,3-6
8	1-6,2-4,3-5
9	1-3,2-5,4-6
10	1-2,3-4,5-6
11	1-6,2-5,3-4
12	1-4,2-6,3-5
13	1-2,3-5,4-6
CSP → 14	1-4,2-5,3-6 ← LV2
15	1-2,3-6,4-5

Fig. 4. ITR2_ECBEL: the structure is shown on the left; possible disulfide bonding patterns are on the right where "S rank" is the ranking of the confidence scores. "CSP" and "LV2" indicate the CSP and level-2 SVM prediction. The actual bonding pattern is boxed.



S rank	bonding pattern
1	1-3,2-4,5-6
2	1-5,2-4,3-6 ← LV2
3	1-2,3-4,5-6
4	1-6,2-4,3-5
5	1-4,2-3,5-6
6	1-4,2-5,3-6
7	1-4,2-6,3-5
8	1-2,3-5,4-6
9	1-2,3-6,4-5
10	1-3,2-5,4-6
11	1-5,2-6,3-4
CSP → 12	1-6,2-5,3-4
13	1-5,2-3,4-6
14	1-3,2-6,4-5
15	1-6,2-3,4-5

Fig. 5. LDTI_HIRME: the structure is shown on the left; possible disulfide bonding patterns are on the right where "S rank" is the ranking of the confidence scores. "CSP" and "LV2" indicate the CSP and level-2 SVM prediction. The actual bonding pattern is boxed.

connectivity. Note that the result from the CSP search is also incorrect ([1–6, 2–5, 3–4]) for this sequence. However, after including the global information, the two-level SVM finally predicts the connectivity ([1–5, 2–4, 3–6]) correctly.

Comparing these three descriptors used, we find that the most influential ones are S and C, because their information weight more for the prediction. As for G, although its effect is less obvious than S and C, it can provide auxiliary information to further optimize the models. Moreover, from the effects of these three descriptors, we show our two-level framework successfully combines both pair-wise and pattern-wise methods. The local relations (suggested by S) and the global information (provided by C and G) complement each other and produce the optimal results. Therefore, we combine these three kinds of information to build the two-level models for disulfide connectivity prediction.

Effect of Candidate Selection

Candidate selection also plays an important role in our experiments. In our experiments, we used the confidence score S to filter some data entries. This filtering process might exclude some actual disulfide bonding patterns. Taking SP39 for an example, there were 28 sequences with $B = 4$, whose actual bonding patterns were excluded from the candidate set when testing. However, such exclusion of the actual bonding patterns did not affect the prediction much. Because the confidence scores of these bonding patterns are too small, they are not likely to be predicted as the connectivity anyway. In our experiments, we also built models without candidate selection. The resulting accuracies for SP39 and SP43 are 69 and 62%, respectively. Compared to the models with candidate selection, the accuracy dropped slightly. This is because without candidate selection, the imbalance of data may affect model training. Also, when all possible patterns are considered, the models take much more time to train. As a result, the candidate selection not only eases the severe imbalance of data but also speeds up the model training.

CONCLUSION

Previous solutions to disulfide connectivity have been restricted to either pair-wise or pattern-wise methods. Due to the data encoding scheme, the information used for modeling is also confined to the local view or global aspect. In this article, we devised a novel two-level framework to combine both encoding schemes. Our results show that our final models have outstanding performance for disulfide connectivity prediction.

For the pair-wise methods,[11,12,15,16,18] the idea is to concentrate on the relation between two cysteine residues. Although the local information can be fully examined, the encoding is difficult to include the global information of a protein. On the other hand, the pattern-wise methods[13,14,17] have the global view of proteins. The global information, such as the positions of cysteines,[13] the protein length,[13] the composition of amino acids,[17] or the cysteine separation profiles[14,17] can be easily encoded for the models. Yet, the pattern-wise methods cannot fully explore the local environment of cysteine residues. In addition, the pattern-wise methods may suffer from the imbalance of training data. These conditions also restrict the performance of the pattern-wise models.

Our two-level framework takes the advantages of both encoding schemes and attempts to avoid the limitation of both methods. The local information of cysteine pairs is fully explored in the level-1 models, whereas the level-2 further incorporates the information from a global aspect. With the two-level framework, both the local and global information can be included to provide more comprehensive data to our models. As shown in our experiments, such combination indeed contributes to better performance than either pair-wise or pattern-wise models.

Furthermore, the proposed framework is flexible and extensible. The models employed in this study are SVMs, but other machine-learning algorithms, such as neural networks, can also be applied. The first level can be extended with multiple models to provide more accurate inference of the pair-wise relation between cysteines. In the level-2, more features can also be encoded to enrich the information provided to models. This encourages future research to extend the framework for disulfide connectivity prediction. The results from our method may also be useful for advanced studies in protein structure prediction, protein structure modeling, and protein engineering.

ACKNOWLEDGMENTS

We would like to thank Jianlin Cheng for generously sharing datasets.

REFERENCES

1. Wedemeyer WJ, Welker E, Narayan M, Scheraga HA. Disulfide bonds and protein folding. Biochemistry 2000;39:4207–4216.
2. Harrison PM, Sternberg MJE. Analysis and classification of disulphide connectivity in proteins. J Mol Biol 1994;244:448–463.
3. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
4. Fiser A, Cserzô M, Tüdôs E, Simon I. Different sequence environment of cysteines and half cysteines in proteins: application to predict disulfide forming residues. FEBS Lett 1992;302:117–120.
5. Muskal SM, Holbrook SR, Kim SH. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng 1990;3:667–672.
6. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins Struct Funct Genet 1999;36:340–346.
7. Fiser A, Simon I. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 2000;16:251–256.
8. Martelli PL, Fariselli P, Malaguti L, Casadio R. Prediction of the disulfide-binding state of cysteines in proteins at 88% accuracy. Protein Sci 2002;11:2735–2739.
9. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
10. Chen Y-C, Lin Y-S, Lin C-J, Hwang J-K. Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. Proteins 2004;55:1036–1042.
11. Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. Bioinformatics 2001;17:957–964.
12. Fariselli P, Riccobelli P, Casadio R. A neural network based method for predicting the disulfide connectivity in proteins. Amsterdam: IOS Press; 2002. p 464–468.
13. Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 2004;20:653–659.
14. Zhao E, Liu H-L, Tsai C-H, Tsai H-K, Chen C-H, Kao C-Y. Cysteine separations profiles on protein sequences infer disulfide connectivity. Bioinformatics 2005;21:1415–1420.
15. Baldi P, Cheng J, Vullo A. Large-scale prediction of disulphide bond connectivity. In: Saul LK, Weiss Y, Bottou L, editors. Advances in neural information processing systems 17. Cambridge, MA: MIT Press; 2005. p 97–104.
16. Ferrè F, Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. Bioinformatics 2005;21:2336–2346.
17. Chen Y-C, Hwang J-K. Prediction of disulfide connectivity from protein sequences. Proteins Struct Funct Bioinformatics 2005;61: 507–512.
18. Tsai C-H, Chen B-J, Chan C-h, Liu H-L, Kao C-Y. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. Bioinformatics 2005;21:4416–4419.
19. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
20. Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. Cambridge, MA: MIT Press. 2000.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
22. Vlijmen HWTv, Gupta A, Narasimhan LS, Singh J. A novel database of disulfide patterns and its application to the discovery of distantly related homologs. J Mol Biol 2004;335:1083–1092.
23. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. 2001. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/