Enhancing Protein Disulfide Bonding Prediction Accuracy with Context-based Features

Ashraf Yaseen
Department of Computer Science
Old Dominion University
Norfolk, VA
ayaseen@cs.odu.edu

Yaohang Li
Department of Computer Science
Old Dominion University
Norfolk, VA
yaohang@cs.odu.edu

Abstract — **Accurately predicting protein disulfide bonds from sequences is important for modeling the structural and functional characteristics of many proteins. In this paper, we introduce a new approach to enhance disulfide bonding prediction accuracy. We firstly generate the first-order and second-order mean-force potentials according to the amino acid environment around cysteine residues from large number of cysteine samples. The mean-force potentials are integrated as context-based scores to estimate the favorability of a cysteine residue in disulfide bonding state as well as a cysteine pair in disulfide bond connectivity. These context-based scores are then incorporated as features together with other protein sequence and evolutionary information to train neural networks for disulfide bonding state prediction and connectivity prediction. Our computational results have shown that the context-based scores are effective features to enhance the prediction accuracies of both disulfide bonding state prediction and connectivity prediction. The 10-fold cross validated accuracy is 90.8% at residue-level and 85.6% at protein-level in classifying an individual cysteine residue as bonded or free, which is around 2% accuracy improvement. The average accuracy for disulfide bonding connectivity prediction is improved as well, which yields overall sensitivity of 73.42% and specificity of 91.61%.**

Index Terms — **Disulfide bonds, Context-based scores, Mean-force potentials, Neural Networks**

I. INTRODUCTION

Disulfide bonds are covalent bonds formed between two sulfur atoms in nonadjacent cysteine pairs of a protein structure. They play an important role in folding and enhancing thermodynamic and mechanical stability [1]. Furthermore, certain disulfide configurations provide mechanisms for sensing and responding to tensile forces, diversifying and functionalizing protein folds, minimizing aggregation, confining and coupling conformational changes, and controlling packaging and releasing for intercellular transport [2]. Therefore, correctly predicting the formation and connectivity of disulfide bonds can not only reduce the conformational space to aid modeling protein structures in three dimensions, but also help predict important protein functions.

Typically, most of the disulfide bonding prediction approaches include two stages. The first stage is the bonding state prediction, aimed at determining whether each cysteine residue in a protein sequence is involved in forming a disulfide bond or not. Afterward, the second stage carries out the connectivity prediction, where cysteine pairs likely to form disulfide bonds are identified.

The early methods of predicting the bonding states of cysteine residues used sequence information alone and small training sets. Muskal et al. [3] implemented a neural network to predict disulfide bonding states and achieved 81% accuracy. Fiser et al. [4] proposed a prediction method based on statistical analysis of residue frequencies near the cysteine residues and obtained 71% accuracy.

Substantial improvements were achieved by later disulfide bonding state prediction methods upon using evolutionary information contained in multiple sequence alignments (MSA). Fariselli et al. [5] designed a jury of neural networks trained by sequence profiles using MSA and resulted in 81% accuracy. Fiser and Simon [6] derived conservation scores from MSA to predict the oxidation state of cysteine residues and obtained an accuracy of 82%. More recent methods with enhanced strategies and additional features lead to continuing improvements of bonding state prediction accuracy. Mucchielli-Giorgi et al. [7] investigated the contribution of the overall amino acid composition of the protein and managed to increase the accuracy to 84%. Ceroni et al. [8] proposed a method using spectrum kernel in Support Vector Machines, which yielded 85% prediction accuracy. Martelli et al. [9] combined a hybrid hidden Markov model and a neural network in their prediction system and reached 84% and 88% accuracy measured on protein basis and cysteine basis, respectively. Song et al. [10] incorporated dipeptide composition as features in prediction and gained similar accuracy.

The pioneered method of connectivity prediction was proposed by Fariselli and Casadio [11] based on graph matching, where edges are weighted by residue contact potentials. The reported accuracy is 17 times higher than a random predictor, which is not comparable to the modern predictors with incorporation of evolutionary information in advanced machine learning technologies. Ceroni et al. [12] encoded MSA data into Recursive Neural Networks in their DISULFIND server with 54.5% pattern precision and 60.2% bonded pair accuracy. Ferre and Clote [13] took advantage of secondary structure encoding in their DiANNA server and reached 86% accuracy. Cheng et al. [14] performed large-scale prediction of disulfide connectivity using kernel methods, two-dimensional recursive neural networks, and weighted graph matching and obtained accuracy of 51% pattern precision. Vincent et al. [15] took advantage of decomposition kernels for classifying chains instead of individual residues and achieved prediction accuracy comparable to the other prediction methods.

It is well known that extracting and selecting "good" features are critical to the performance of the learning machines. In the literature, features influencing the formation of disulfide bonds, such as MSA, secondary structures, number of cysteine residues in a protein chain, etc., have been encoded in the machine learning algorithms and have achieved prediction accuracy improvement.

In this paper, we investigate the approaches of deriving context-based scores based on the mean-force potentials [21] derived from a large cysteine sample set. We consider the first-order and the second-order interactions. Afterward, context-based scores for cysteine residues considering nearby neighbors at different distances are generated. These context-based scores are then incorporated as features together with the MSA data to train neural networks for disulfide bonding state and connectivity predictions. 10-fold cross validations are performed. We also test our method on several popular protein benchmarks, including Manesh215 [24], Carugo338 [25], and CASP9 [26].

II. Material and Method

A. The Protein Data Sets

We use the protein chain dataset Cull16633 generated by the PISCES server [16] to collect cysteine samples to generate context-based statistics and for neural network training as well. Cull16633 contains 16633 chains with at most 50% sequence identity, 3.0A resolution cutoff, and 1.0 R-factor. Chains with less than 2 cysteine residues are eliminated. We also eliminate very short chains whose lengths are less than 40 residues since the PSI-BLAST program [22] is usually unable to generate profiles for very short sequences. The disulfide bond assignments are determined by the DSSP program [23]. Inter-chain disulfide bonded cysteines and cysteines with undetermined structures are excluded. After this elimination, the total number of protein chains remained in Cull16633 is 9781 and the total number of cysteine residues is 47655 where 21.27% of these cysteine residues are bonded. We refer to this protein set as Cull50.

We also use another dataset Cull7986 generated from PISCES server with maximum 25% sequence identity, 3.0A resolution and 1.0 R-factor. After filtering, the total number of protein chains is 4340 with a total of 20309 cysteine residues, where 21.28% of those are bonded. This protein chain set is referred to as Cull25. We compare the performance of our prediction methods when Cull50 and Cull25 are used as training sets.

The recent CASP9 targets [26] as well as the public protein data sets Manesh215 [24] and Carugo338 [25], which are popularly employed as benchmarks for secondary structure predictions, are used to benchmark our method. Therefore, any sequences with greater than 25% similarity with the test benchmarks sequences are excluded from the Cull50 and Cull25 when the neural networks are trained and also when the context-based scores are generated.

B. Context-based Statistics

It is well known that there exist general short range regularities in the primary structure of proteins [17]. Presumably, the neighboring residues have strong and probably deterministic influence to the chemical property of cysteine in forming disulfide bond [3, 28]. Actually, cysteine often forms particular motifs of biochemical functions with neighboring residues, such as Cys-X-X-Ser [18], Cys-X-X-Cys [19], Leu-X-Cys-X-Glu [20], Cys-X-X-Asp-X-X-Cys [27], etc.

In this paper, we generate the mean-force potentials [21] to estimate the favorability of a cysteine residue in a bonding state within its amino acid environment. The mean-force potential is based on the derived statistics of correlations between the cysteine residue and its nearby neighbors. In particular, the recent increasing number of experimentally determined protein structures in PDB has provided sufficient number of samples to enable derivation statistics for second-order mean-force potential. In our method, the first-order statistics estimate the correlations between a cysteine residue and one of its neighboring residues while the second-order statistics estimate the correlations between a cysteine residue and the coexistence of two neighboring residues. Both first-order and second-order statistics are extracted from protein chains in the Cull datasets. For a cysteine sample with window size of K, there are K - 1 position combinations for first-order statistics and $\binom{K-1}{2}$ position combinations for the second-order statistics of a cysteine residue in bonding state.

Similarly, the first-order and second-order statistics of a disulfide bonded cysteine pair and its neighboring residues are also extracted to estimate the probability of a cysteine pair in forming disulfide connectivity. Compared to the statistics in estimating a cysteine residue in a bonding state, the main difference lies in the different number of position combinations in second-order statistics since the two neighboring residues may belong to two different cysteine residues. As a result, considering a window size of K for both cysteine residues connected in a disulfide bond, there are totally $\binom{2K-2}{2}/2$ position combinations for the second-order statistics of a bonding cysteine pair.

C. Context-based Potential

The context-based potential for cysteine bonding state is generated based on Sippl's mean-force potential method [21]. In this paper, we consider the first-order and the second-order mean-force potentials only. Currently, there is insufficient number of available protein structures in PDB to derive meaningful statistics for estimating higher order interactions.

According to the inverse-Boltzmann theorem, we introduce the first-order mean-force potential $U\left(R_{(k)}, Bonded\right)$ to treat the interaction between residue $R_{(k)}$ and cysteine in forming a disulfide bond,

$$U\left(R_{(k)}, Bonded\right) = -RTln\frac{P_{obs}\left(Bonded|R_{(k)}\right)}{P_{ref}\left(Bonded|R_{(k)}\right)}.$$

Here R is the gas constant, T is the temperature, $P_{obs}\left(Bonded|R_{(k)}\right)$ is the observed probability, and $P_{ref}\left(Bonded|R_{(k)}\right)$ is the reference probability.

Similarly, the second-order mean-force potential $U\left(R_{(k_1)}, R_{(k_2)}, Bonded\right)$ is calculated as

$$U\left(R_{(k_1)}, R_{(k_2)}, Bonded\right)$$
$$= -RTln\frac{P_{obs}\left(Bonded|R_{(k_1)}, R_{(k_2)}\right)P_{ref}\left(Bonded|R_{(k_1)}\right)P_{ref}\left(Bonded|R_{(k_1)}\right)}{P_{ref}\left(Bonded|R_{(k_1)}, R_{(k_2)}\right)P_{obs}\left(Bonded|R_{(k_1)}\right)P_{obs}\left(Bonded|R_{(k_1)}\right)}.$$

Influenced by all of its neighboring residues, the overall mean-force potential for the interactions of a cysteine residue in bonding state is the summation of all first-order and second-order potentials while the higher-order interactions are ignored

$$U = \sum_{k}^{k\neq0} U\left(R_{(k)}, Bonded\right) + \sum_{k_1}^{k_1\neq0}\sum_{k_2}^{k_2\neq0} U\left(R_{(k_1)}, R_{(k_2)}, Bonded\right).$$

The potential for a bonded cysteine pair can be obtained in a similar way. These potentials are used as context-based scores to be encoded in neural network training for bonding state and connectivity predictions.

D. Neural Network Model

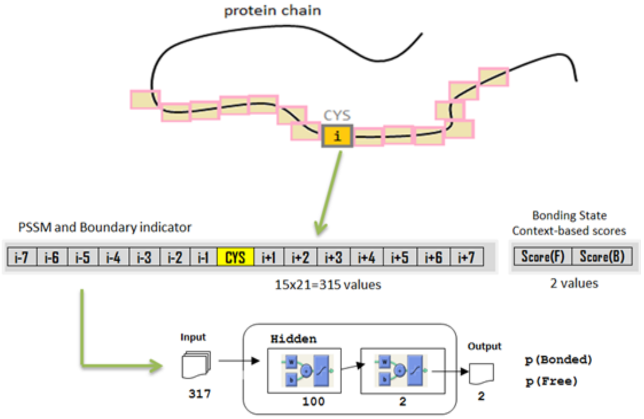

Fig. 1. Neural network architecture for disulfide bonding state prediction

We adopt the standard feed-forward back-propagation neural network architecture for both bonding state and connectivity prediction. The neural network (Fig. 1) for bonding state prediction uses a window size of 15 residues for input encodings. Each residue is represented with 20 values from the PSSM data and 1 extra input to indicate if the window overlaps C-terminal or N-terminal. When incorporating the context-based scores in training the neural network predictor, two more inputs specifying the scores of the cysteine residue being in free and bonding state are added. Hence, a total number of 317 values are used to describe each cysteine residue. 100 hidden nodes are used in the neural network for bonding state prediction.

The neural network for connectivity prediction incorporates two windows, each with size of 15 residues, for input encoding, each window for a cysteine residue in a cysteine pair. Each residue is encoded with 20 PSSM values and 1 boundary indicator. The predicted results (bonded or free) from the bonding state prediction for both cysteine residues and the context-based scores for connectivity are also encoded as input. As a result, there are totally 636 input values for each cysteine pair. 150 hidden nodes are used in the neural network for connectivity prediction.

III. Results & Discussion

A. Bonding State Prediction

Table I

Comparison of Prediction Performance of Bonding States Using PSSM Only and PSSM with Context-Based Scores on Cull25 and Cull50 Using 10-Fold Cross Validation

	Cull25		Cull50	
	PSSM Only	PSSM+Score	PSSM Only	PSSM+Score
Oc[1]	0.870	0.888	0.885	0.908
Op[2]	0.719	0.751	0.829	0.856
Sn[3]	0.554	0.616	0.655	0.720
Sp[4]	0.945	0.956	0.947	0.959
MCC[5]	0.574	0.646	0.734	0.801

1. Q_c (residue-level accuracy)=P_c/N_c, where P_c is the total number of correctly predicted cysteine residues and N_c is the total number of cysteine residues
2. Q_p (protein-level accuracy) =P_p/N_p , where P_p is the total number of proteins with all of its cysteine residues being correctly predicted and N_p is the total number of proteins in the data set
3. S_n(Sensitivity) =TN/(TP+FN)
4. S_p(specificity) =TN/(TN+FP)
5. MCC (Matthew's correlation coefficient)

$$= \frac{(\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}},$$

where TP, TN, FP, and FN are the number of true positives, the number of true negatives, the number of false positives, and the number of false negatives, respectively.

Table I compares the prediction qualities of bonding states with PSSM-only encoding and PSSM with context-based scores encoding after 10-fold cross validation. Compared to the one trained with PSSM data only, the neural network using context-based scores as additional features results in improvements in all performance indexes, including *Qc*, *Qp*, *Sn*, *Sp*, and *Mcc*. The residue-level prediction accuracy (0.908) and protein-level prediction accuracy (0.856) are higher than the reported accuracies in [3-15]. Table I also compares the prediction qualities when Cull25 and Cull50 are used as training sets. Cull50 has more than twice cysteine samples as Cull25, which leads to better prediction performance.

Table II

Comparison of Residue-Level Accuracies (Q_C) on Benchmarks of Manesh215, Carugo338, and CASP9 Using Cull25 and Cull50 as Training Sets

	Cull25		Cull50	
	PSSM	PSSM+Score	PSSM	PSSM+Score
Manesh215	0.830	0.848	0.879	0.900
Carugo338	0.808	0.821	0.872	0.884
CASP9	0.950	0.951	0.955	0.963

Table II shows the comparison of residue-level accuracies (Q_c) on the public benchmarks, including Manesh215, Carugo338, and CASP9. Cull50 training set yields better prediction performance than Cull25.

Moreover, the context-based scores are effective features for the training process. When context-based scores are incorporated, the prediction accuracies are improved on all three benchmarks compared to using PSSM data only.

B. Connectivity prediction

We compare the 10-fold cross validation for disulfide bond connectivity predictions on Cull50 using PSSM-only and PSSM with context-based scores for neural network encoding. Similar to bonding state prediction, incorporating the context-based scores as features in neural network training enhances the connectivity prediction accuracy, where sensitivity (*Sn*), specificity (*Sp*), and overall accuracy (Q_c) are improved from 73.07%, 91.03%, and 86.91% to 73.42%, 91.61%, and 87.34%, respectively, compared to PSSM only encoding. These prediction results are also higher than the reported disulfide connectivity accuracies in the popular disulfide bond prediction servers [11-15].

IV. Conclusions

An approach of deriving context-based scores based on the mean-field potentials for characterizing the favorability of cysteine residues in disulfide bond according to their amino acid environment is developed in this paper. Recently, the increasing number of experimentally determined protein structures in PDB has made sufficient number of cysteine samples available. This enables us to obtain reliable statistics for second-order mean-field potentials and thus leads to context-based scores with better accuracy. These context-based scores are selected as features together with other sequence and evolutionary information in neural network training for disulfide bonding state and connectivity predictions. The effectiveness of using context-based features has been demonstrated in our computational results in 10-fold cross validation as well as on benchmarks of Manesh215, Carugo338, and CASP9, where enhancements of prediction accuracies in both bonding state and connectivity predictions are observed.

A web server implementing our disulfide bonding prediction program is currently under development. Services of both bonding state and connectivity predictions will be provided.

Acknowledgements

This work is supported in part by NSF grant no. 1066471 and ODU SEECR grant.

References

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

[2] D. Fass, "Disulfide Bonding in Protein Biophysics," Annu. Rev. Biophys., 41:63–79, 2012.

[3] S. M. Muskal, S. R. Holbrook, S. H. Kim, "Prediction of the disulfide-bonding state of cysteine in proteins," Protein Engineering, 3(8): 667-672, 1990.

[4] A. Fiser, M. Cserzo, E. Tudos, I. Simon, "Different sequence environments of cysteines and half cystines in proteins: Application to predict disulfide forming residues," FEBS Letters, 302: 117-120, 1992.

[5] P. Fariselli, P. Riccobelli, R. Casadio, "Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins," Proteins, 36: 340-346, 1999.

[6] A. Fiser, I. Simon, "Predicting the oxidation state of cysteines by multiple sequence alignment," Bioinformatics, 16(3): 251-256, 2000.

[7] M. H. Mucchielli-Giorgi, S. Hazout, P. Tuffery, "Predicting the disulfide bonding state of cysteines using protein descriptors," Proteins, 46(3): 243-249, 2002.

[8] A. Ceroni, P. Frasconi, A. Passerini, A. Vullo, "Predicting the disulfide bonding state of cysteines with combination of kernel machines," J. VLSI Signal Processing, 35: 287-295, 2003.

[9] P. L. Martelli, P. Fariselli, L. Malaguti, R. Casadio, "Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks," Protein Engineering, 15(12): 951-953, 2002.

[10] J. N. Song, M. L. Wang, W. J. Li, W. B. Xu, "Prediction of the disulfide-bonding state of cysteines in proteins based on dipeptide composition," Biochemical and Biophysical Research Communications, 318(1): 142-147, 2004.

[11] P. Fariselli, R. Casadio, "Prediction of disulfide connectivity in proteins," Bioinformatics, 17(10): 957-964, 2001.

[12] A. Ceroni, A. Passerini, A. Vullo, P. Frasconi, "DISULFIND: a disulfide bonding state and cysteine connectivity prediction server," Nucleic Acids Research, 34: W177-W181, 2006.

[13] F. Ferre, P. Clote, "DiANNA: a web server for disulfide connectivity prediction," Nucleic Acids Research, 33: W230-W232, 2005.

[14] J. Cheng, H. Saigo, P. Baldi, "Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching," Proteins, 62: 617-629, 2006.

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

[16] G. Wang, R. L. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19(12): 1589-1591, 2003.

[17] F. Vonderviszt, G. Y. Matrai, I. Simon, "Characteristic sequential residue environment of amino acids in proteins," Int. J. Peptide Protein Res., 27: 483-492, 1986.

[18] C. S. Sevier, C. A. Kaiser, "Formation and transfer of disulphide bonds in living cells," Nature Reviews Molecular Cell Biology, 3: 836-847, 2002.

[19] A. T. Washington, G. Singh, "Diametrically opposed effects of hypoxia and oxidative stress on two viral transactivators," Virology Journal, 7:93, 2010.

[20] Y. W. Kim, G. A. Otterson, R. A. Kratzke, A. B. Coxon, F. J. Kaye, "Differential specificity for binding of retinoblastoma binding protein 2 to RB, p107, and TATA-binding protein," Mol. Cell Biol., 14(11): 7256-7264, 1994.

[21] M. J. Sippl, "Calculation of conformational ensembles from potentials of mean force – an approach to the knowledge-based prediction of local structures in globular proteins." J. Mol. Biol., 213: 859-883, 1990.

[22] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, 25: 3389-3402, 1997.

[23] W. Kabsch, C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," Biopolymers, 22: 2577-2637, 1983.

[24] S. Ahmad, M. Gromiha, A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," Proteins, 50:629–635, 2003.

[25] O. Carugo, "Predicting residue solvent accessibility from protein sequence by considering the sequence environment," Protein Eng., 13:607-609, 2000.

[26] [26] L. N. Kinch, S. Shi, H. Cheng, Q. Cong, J. Pei, V. Mariani, T. Schwede, N. V. Grishin, "CASP9 target classification," Proteins, 79(Suppl 10):21-36, 2011.

[27] Y. S. Jung, C. A. Bonagura, G. J. Tilley, H. S. Gao-Sheridan, F. A> Armstrong, C. D. Stout, B. K. Burgess, "Structure of C42D Azotobacter vinelandii FdI. A Cys-X-X-Asp-X-X-Cys motif ligates an air-stable [4Fe-4S]2+/+ cluster," J. Biol. Chem., 275(47): 36974-36983, 2000.

[28] I. Rata, Y. Li, E. Jakobsson, "Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops," Journal of Physical Chemistry B, 114(5): 1859-1869, 2010.