Soft Computing Methods for Disulfide Connectivity Prediction

Libertas Academica
FREEDOM TO RESEARCH

Alfonso E. Márquez-Chamorro and Jesús S. Aguilar-Ruiz
School of Engineering, Pablo de Olavide University, Seville, Spain.

ABSTRACT: The problem of protein structure prediction (PSP) is one of the main challenges in structural bioinformatics. To tackle this problem, PSP can be divided into several subproblems. One of these subproblems is the prediction of disulfide bonds. The disulfide connectivity prediction problem consists in identifying which nonadjacent cysteines would be cross-linked from all possible candidates. Determining the disulfide bond connectivity between the cysteines of a protein is desirable as a previous step of the 3D PSP, as the protein conformational search space is highly reduced. The most representative soft computing approaches for the disulfide bonds connectivity prediction problem of the last decade are summarized in this paper. Certain aspects, such as the different methodologies based on soft computing approaches (artificial neural network or support vector machine) or features of the algorithms, are used for the classification of these methods.

KEYWORDS: disulfide connectivity prediction, protein structure prediction, soft computing, support vector machines, neural networks

CITATION: Márquez-Chamorro and Aguilar-Ruiz. Soft Computing Methods for Disulfide Connectivity Prediction. *Evolutionary Bioinformatics* 2015:11 223–229 doi: 10.4137/EBO.S25349.

TYPE: Review

RECEIVED: March 02, 2015. **RESUBMITTED:** August 24, 2015. **ACCEPTED FOR PUBLICATION:** August 31, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,266 words, excluding any confidential comments to the academic editor.

FUNDING: This research was funded by the Project of Excellence P07-TIC-02611 "Intelligent Systems for discovering patterns of behavior. Application to biological database" and by the Spanish MEC under project TIN2011-28956-C02-01. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: amarcha@upo.es

Introduction: Background and Purpose

A disulfide bond, also known as disulfide bridge or SS-bond, plays an important role in the folding process, stability, and function of the protein. The oxidation of the thiol group (R–SH) of the cysteines is required for the formation of a covalent bond between cysteines, known as disulfide bond (S–S).[1] In this process, two atoms of hydrogen are released. These bonds are usually found in proteins that are secreted to the extracellular medium.

The prediction of disulfide bonds connectivity can help the protein structure prediction (PSP) problem that is an important challenge in structural bioinformatics. This issue can be considered as one of the subproblems to tackle the problem of PSP.[2] The tertiary structure of a protein is the result of the formation of disulfide bonds, hydrogen bonds, hydrophobic effect, and other interactions between the side chains of the amino acids. Disulfide bonds can be formed between cysteine residues in the same chain (intrabonded), separated by many amino acids or belong to different polypeptide chains of the protein (interbonding). Disulfide bonds stabilize protein native structures by lowering global-free energy and constraining the unfolded conformation.

Determining the disulfide bondings in an experimental way, such as X-ray crystallography, requires time-consuming procedures and expensive equipments. On the other hand, several computational approaches have been developed for the disulfide bond prediction problem, providing a fast and effective way to understand biological molecules. This problem can be divided into two different steps: disulfide bonding state prediction and disulfide connectivity prediction (DCP).[3] The aim of the methods of the first group is to classify cysteines according to their molecular state (bonded to another cysteine of the chain or to a free cysteine). Thus, we are addressing a binary classification problem, where the class labels are the states of the cysteines (reduced or oxidized). Predicting the disulfide state of each cysteine is a step toward the location of disulfide bridges in proteins. DCP tries to elucidate the different pairs of cysteines that are bonded in a protein sequence.[4] Currently, available predictors are mainly based on neural network (NN) approaches and support vector machines (SVMs) as well as other predictive methods.

This article presents relevant and ultimate DCP methods based on soft computing techniques. Section 2 introduces some basic concepts of the DCP problem. Section 3 shows the most relevant techniques and are briefly described. Finally, some conclusions are summarized.

Preliminary Concepts

In order to represent the protein disulfide bondings, we can use two different types of encoding: pairwise and pattern-wise models, depending on the extraction of local information or global information of the protein training data. Some of these information features used for the encoding are evolutionary

information, physicochemical properties of amino acids, prediction of secondary structures (SSs), cysteine separation distance, relative order of cysteines, protein length, and protein molecular weight.

Pairwise vs. pattern-wise model. Pairwise model uses the local information of the disulfide bond.[5] Generally, this encoding consists of two windows of residues centered around the two target cysteines. Local properties are based on the local environment of the cysteine residue, that is, composition of residues and physicochemical properties of the residues in the local environment of the target cysteine. On the other hand, pattern-wise model analyzes the global information of the whole protein for the encoding. Specifically, this encoding contains information such as the length of the protein, the position of the cysteines in the chain, the composition of amino acids, and the separation between cysteines.

Input data features. Input data are encoded according to several features based on the global and local properties of the cysteines. Evolutionary information, physicochemical properties, SS prediction, distance between cysteines, and protein length or protein molecular weight are some of the most used features in the literature.

Evolutionary information. Sequence alignment is a standard technique in bioinformatics for visualizing the relationships between residues in a collection of evolutionary or structurally related proteins. Existing DCP algorithms in the literature have used multiple sequence alignment, position-specific scoring matrices (PSSMs)[6,7] and correlated mutations[4] as input encoding.

The tendency of residue positions in proteins to mutate coordinately is called correlated mutation. For each cysteine residue, its frequency of being correlatively mutated with respect to all other cysteine residues present in the same chain is calculated. This is computed by counting the number of times the two cysteine residues are either present or absent together and dividing it by the total number of counts.

On the other hand, PSSMs are also obtained from sequence alignments. PSSMs determine the substitution scores between the amino acids according to their positions in the alignment. Each cell of the matrix is calculated as the $\log_2$ of the observed substitution frequency at a given position divided by the expected substitution frequency at that position. Thus, a positive score (ratio > 1) indicates that the observed frequency exceeds the expected frequency, suggesting that this substitution is surprisingly favored. A negative score (ratio < 1) indicates the opposite: the observed substitution frequency is lower than the expected frequency, suggesting that the substitution is not favored.

Physicochemical properties. The most direct information we can extract from the primary sequence of a protein are physicochemical characteristics of its residues. With this information, we can generate representations of, for example, how the hydrophobicity varies along the sequence of the protein and obtain information about hydrophobic areas, which may help the prediction of structural characteristics. Properties used in the literature are hydrophobicity, polarity, volume of residues, graph shape index, and isoelectric point, among others. Shilton et al.[8] and Song et al.[9] include amino acid properties as input data.

Secondary structures. SS prediction consists of predicting the location of α-helices and β-sheets and turns from a sequence of amino acids. The location of these motifs could be used by approximation algorithms to obtain the tertiary structure of the protein. SS is employed as input data by Lin and Tseng[7] and Song et al.[9] In particular, a relevant study presented by Song et al.[9] determines that the three most important features to enhance the DCP are SS, PSSMs, and normalized sequence distance between oxidized cysteines (DOC).

Cysteine separation distance. The separation distance between two cysteines is defined as $||i–j||$, where i and j are the sequence indices of two cysteines.[9] According to the sequential distance, we can estimate which pair of cysteines is bonded. The higher the distance (>100) between two cysteines, the lower the probability of being bonded. A second feature related with the positions of cysteines in the sequence describes the cysteine sequential ordering difference between each pair of cysteines.

Protein length and protein molecular weight. Protein length indicates the number of amino acids of each sequence. Molecular weight of a protein is the mass of this molecule. It can be calculated as the sum of the individual isotopic masses of all the atoms in the molecule. These features correspond to the representation of global information of a protein sequence and are used in several methods.[10]

Databases. The benchmark datasets used in the area of DCP are extracted from Swiss-Prot 39 (SP39).[11] SP39 includes 726 proteins of the Swiss-Prot database release no. 39, which include from two to five cysteine bonds. This dataset was experimentally verified and includes intrachain disulfide bridge annotations. The sequence homology between the proteins of this dataset is ≤30%. SP43 and SP56 are also employed in several proposals. SPX, an extended dataset of SP39 and SP41, is also used in the literature.

Other dataset used is called PDBCYS introduced by Savojardo et al.[10] This dataset was extracted from PDB (released May 2010) and contains 1797 Eukaryotic protein structures with resolution <2.5 with at least two cysteine residues and global pairwise sequence similarity $<25\%$. PDBCYS includes 7619 free and 3194 bonded cysteines. This dataset contains a high number of proteins, and its sequence similarity is very low. These characteristics make it a good candidate for the evaluation of a method to be used as training and test dataset.

Performance metrics. The quality measures used to evaluate the accuracy of the connectivity patterns prediction methods are mainly two.[10]

R_b indicates the number of correctly predicted bonds (N_c) divided by the total number of disulfide bonds (N_b) in test proteins. This measure is also named P_b and Q_c in the literature.

$$R_b = \frac{N_c}{N_b} \quad (1)$$

Q_b is the number of proteins whose connectivity patterns are correctly predicted (N_{prot}) divided by the total number of proteins (N_t) in the test set. This measure is also named Q_p in the literature.

$$Q_b = \frac{N_{prot}}{N_t} \quad (2)$$

Methods

Support vector machines. SVMs are based on the transformation of the input space into a feature space of higher dimensionality. SVM techniques then build a hyperplane, or a set of hyperplanes, in this space trying to maximize the margin between each pair of classes. The function that performs the transformation of the space is called kernel function. SVMs are used as a machine learning tool to predict tertiary structure from the primary sequence. On the other hand, support vector regression (SVR) machine is a regression model based on SVM.

The four following methods employ information about the cysteine pairs (local information) and the whole sequence protein (global information) indistinctly. Savojardo et al.[4] incorporate evolutionary information derived from correlated mutations as feature encoding for a SVR machine. Correlated mutations are represented in the form of corrected mutual information (MIp) and inverse of covariance matrix (iCOV). SVR was trained using local and global information. As encoding features, they employ two PSSM-based windows centered on the pair of cysteines, the relative order of the cysteines in the sequences, the separation distance between each pair of cysteines, and the cited correlated mutation information. The predictions of the SVR constitute the weights of the edges of the graph formed by all possible cysteine pairs of the sequence. The Edmond–Gabow algorithm[12] is used to solve maximum weighted matching problem on this graph and obtain the most probable disulfide pattern. A 20-fold cross-validation is used to evaluate the SVR. Savojardo et al.[10] perform a two step-based algorithm, which includes bond state prediction and connectivity pattern prediction. They include a protein subcellular localization to improve the performance of the disulfide bond state predictor method. This model contains local and global information for the connectivity pattern prediction. Normalized protein length, protein molecular weight, and amino acid composition are the global features. Two PSSM windows of length 13, relative order cysteines, and the cysteine separation distances constitute the local features for the training of the SVR. The method described by Liu and Chen[13] combines global and local information. Cysteine separation profile (CSP) and evolutionary information profiles are encoded as input data for the SVM. CSP represents the distribution of the cysteines in the whole sequence (global information). SVM infers the potential of connectivity between each pair of cysteines with prior information of the bonding states. Later, Gabow's algorithm finds the disulfide connectivity pattern. Finally, Chen et al.[5] propose a two-level framework to predict the disulfide connectivity. This method combines two encoding schemes, pairwise and pattern-wise models. The bonding probabilities are the outputs of the first level, and this information is used as input data in the second level. As local information, the algorithm uses DOC and evolutionary profiles. On the other hand, as global information of the protein, the method employs the confidence scores of the pairwise probabilities, CSP, the cysteine ordering, and the protein length. This proposal used SVMs, but artificial neural networks (ANNs) can also be used.

Several methods have combined a SVM with a maximum weight perfect matching algorithm to predict the disulfide connectivity patterns. For instance, Lin and Tseng[14] introduce a method, called disulfide bonding connectivity pattern prediction web server (DBCP), based on SVM to predict the probabilities of the bonding pairs and the Edmond–Gabow algorithm to solve the maximum weight perfect matching problem. In this work, the atom coordinates of the C_α of cysteine amino acids are obtained by MODELLER (http://salilab.org/modeller/) to calculate the Euclidean distance of the cysteine pairs. These pair distances (PDs) are then used as input feature of the SVM. The method described by Tsai et al.[15] introduces a method based on SVM and DOC. They use three different normalized scaling schemes of DOC. After obtaining the potentials of connectivity between pairs of cysteines as outputs of the SVM, the Gabow's algorithm to solve the maximum weight matching problem is applied.

Physicochemical properties and prediction of protein SS are used as input features of the SVM approaches in the next three methods. The method proposed by Song et al.[9] adopts an SVR, method based on multiple sequence feature vectors and SS predictions as input features. Once the probabilities are obtained by the algorithm, a ranking of them is provided, determining the predicted disulfide bridges. The cited multiple sequence feature vector is composed of cysteine–cysteine coupling, amino acid compositions, cysteine separation distance, cysteine ordering, protein molecular weight, and protein sequence length. Finally, predicted secondary structure (PSS), is added to the encoding. Lin and Tseng[7] propose a method based on four features: PSSM, PSS, normalized bond lengths, and amino acid physicochemical properties indices. A SVM combined with a maximum weight perfect matching algorithm predict the disulfide connectivity patterns. To adjust the parameters of the SVM and the window sizes of the features, an evolutionary algorithm called

multiple trajectory search is employed during the SVM training phase. In a previous work,[19] the authors introduced a normalized PD vector as input feature information for the SVM. This vector includes the Euclidean distance between all the oxidized cysteines of the training proteins. Finally, Shilton et al.[8] elaborate an encoding scheme based on physicochemical properties and statistical features. These properties are hydrophobicity and polarity according to the scales described by Kyte and Doolitle[21] and Grantham,[22] respectively. As statistical feature, the probability of occurrence in SS based on Chou–Fassman scale is used. The algorithm uses a priori knowledge on their bonding states.

Lu et al.[17] develop a method that includes bonding state and connectivity pattern prediction using SVM. A genetic algorithm (GA) was implemented to optimize the feature selection (FS) and to adjust the parameters of the SVM. Each individual of the population of the GA is composed of three feature vectors, one to represent the different combination of features and the other two for the parameters of the SVM. A connectivity matrix, which includes the predicted cysteine states, is used as input encoding for the SVM to infer the disulfide connectivity patterns.

The proposal described by Chen and Hwang[16] implements an algorithm based on SVM. Local sequence environments with evolutionary information of cysteine pairs, cysteine sequences separation, and amino acid content constitute the biological features of the three input vectors of the SVM. This work determines the existence of a clear relationship between the disulfide patterns and cysteine sequence separations.

Zhu et al.[18] present a SVR method combined with FS to improve the performance and avoid the high-dimensional feature space. The following FS methods were employed: variance score, Laplacian score, and the Fisher score. They conclude that local features dominate the formation and the prediction of disulfide bridges.

The method proposed by Becker et al.[6] employs three different classification algorithms for the prediction of disulfide bonding probabilities: k-nearest neighbors, SVMs, and extremely randomized trees. Therefore, they propose a feature function selection, which determines a subset of feature functions and the best setting for associated window sizes. Finally, the best performance of the algorithm is obtained with the use of PSSM together with the CSP.

As limitations of SVM–SVR methods, we can argue that the kernel models overfit the model selection criterion, the selection of the optimal kernel function parameters is difficult, and for large-scale tasks, the algorithmic complexity and memory requirements are remarkable.

A summary of SVM-based methods for disulfide bond prediction is shown in Table 1. The first column refers to the name of the method in the literature. The second column shows the reference of the work. Third and fourth columns represent the accuracy values of R_b and Q_b (equations (1) and (2)). In case the value of accuracy is not provided by the authors, it is marked with a dash. The fifth column shows the data set used for the experimentation. The sixth column shows the main characteristics of the method. Finally, if the software is available, the URL is shown.

A real comparison of the presented methods is quite difficult. However, we have focused on those methods tested using the same recurrent data set (SP39). We can highlight the method presented by Lin et al.[19], which achieves a high level of accuracy (R_b 93.6 and Q_b 91.0). This method includes as input feature the distance between all the oxidized cysteines

Table 1. Summary of SVM-based methods for disulfide connectivity pattern prediction in chronological order.

METHOD	REF.	R_b(%)	Q_b(%)	DATASET	DESCRIPTION	SOFTWARE
	16	57.0	55.0	SP39	Local information	
	8	59.0	52.0	SPX	AA properties, PSS	
PreCys	15	70.0	63.0	SP39	DOC	http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/
	5	–	70.0	SP39, SP43	Probability outputs	
	13	71.0	65.0	SP39, SP43	CSP, evol. inf.	
	17	79.2	73.9	SP39	GA for FS	
	9	77.9	74.4	SP39, SP43	SVR	http://foo.maths.uq.edu.au/~huber/disulfide
DBCP	14	61.2	46.9	CHK25, SP56	MWPM	http://biomedical.ctust.edu.tw/edbcp/
	18	80.3	76.0	SP39	Feature selection	
DISLOCATE	10	60.0	54.0	PDBCYS	Local information	http://dislocate.biocomp.unibo.it/dislocate
	19	93.6	91.0	SP39	NPD	
	7	–	74.4	SP39	MTS, PSSM	
	6	–	58.3	PDBCYS, SPX	NN, ERT, PSSM, CSP	http://m24.giga.ulg.ac.be:81/x3CysBridges
	4	66.2	59.3	PDBCYS	Corr. mutations	

of the training proteins. According to Song et al.[9], this is one of the most relevant features for the DCP.

Neural networks. An ANN is a computing system of interconnected elements, which process information by their dynamic state response to external inputs. The weights of the connections can be tuned based on the experience, making ANNs adaptive to inputs and capable of learning. ANN can be trained to recognize the disulfide connectivity patterns.

PSIPRED[20] and PSI-BLAST[23] are employed by Ferre and Clote[24] for the encoding scheme of a diresidue NN. The method consists of two phases, one for the bonding state predictor and the other for the connectivity predictor. Diresidue PSSMs are also used as evolutionary information. ANN provides a likelihood of forming a disulfide bond for each cysteine pair. Finally, a Rothberg's implementation of the Gabow's algorithm (http://elib.zib.de/ppub/Packages/mathprog/matching/weighted) is applied to determine the disulfide connectivity. Other methods are based on a two-dimensional recursive neural network (2D-RNN). In particular, the method proposed by Cheng et al.[3] presents an algorithm based on a 2D-RNN and the prediction of SS and solvent accessibility. The outputs of the RNN are the probabilities of existence of a cysteine bridge. This method can be applied when the information of the bonding states is known or unknown and is useful for chains with more than five bonds in the sequence. The majority of the algorithms only predict sequences with two to five cysteine bonds. The method showed by Vullo and Frasconi[25] uses evolutionary information in the form of multiple alignment profiles as input data of a 2D-RNN. The disulfide patterns are presented like graphs. The candidate graphs are compared to the correct graphs and are scored according to a similarity metric. This method, called DISULFIND, was implemented as a prediction server and described by Ceroni et al.[26] Finally, Yaseen and Li[27] perform an NN encoding based on PSSM and context-based statistics using two amino acid windows of 15 residues. They calculate the mean-force potentials as statistics to estimate the favorability of cysteine contacts. The cysteine bonding state is also predicted by this method, called Dinosolve.

Martelli et al.[28] present a hybrid system based on hidden neural networks that combine ANNs and hidden Markov models for the prediction of bonding states. A window of 27 residues centered on the cysteine residue is used as input feature.

NNs provide a high degree of flexibility. Besides encoded input vectors of pair of amino acids, we may include neurons with additional information, for example, sequence length or evolutionary information. On the other hand, NNs have certain limitations, for example, constraints on the encoding of input data, the use of appropriate parameters of the ANN, and overfitting. Comparing NNs and SVMs, we can state that ANNs follow a heuristic path, while SVMs are theoretically founded. ANNs can find multiple local minima solutions, while SVM classifiers converge in global and unique solutions. On the other hand, ANNs consume less storage and computational resources than SVMs.

A summary of NN methods for disulfide bond prediction is shown in Table 2. According to the results, we can assume that the different benchmarks make difficult a real comparison. However, Dinosolve clearly achieves the best results. This is due to the use of PSSM and statistics that calculates the probabilities of each disulfide bond connectivity as input features of the ANN to enhance the predictions. We can conclude that statistics and evolutionary information provide a differentiating factor for the DCPs.

Other predictive methods. In addition to the aforementioned approaches, there are other important approaches to tackle the disulfide connectivity problem, such as nearest neighbor and Monte Carlo simulated annealing (MCSA) approaches. In this section, we will cover some of these strategies.

Two proposals are based on nearest neighbor. The first method is proposed by Vincent et al.[29] and consists of two phases. In the first phase, a binary classifier determines the prediction of cysteine bonding states and whether or not the disulfide bridges correspond to intra- or interchain. The second phase is formed by a simple 1-nearest neighbor (1-NN) algorithm based on separation distances between cysteines and evolutionary profiles. The second proposal, described by Niu et al.[30], is based on a nearest neighbor algorithm using a FS method for the intra- and inter-disulfide bond prediction. They use an incremental FS to determine the optimal number of features. Sequence distance, PSSMs, residual disorder, and amino acid factor were used for the encoding.

Table 2. Summary of ANN-based methods for disulfide connectivity pattern prediction in chronological order.

METHOD	REF.	R_b(%)	Q_b(%)	DATASET	DESCRIPTION	SOFTWARE
	28	–	88.0	4136, PDB	HNN, HMM	
	25	49.0	–	SP39	RNN, evolutionary information	
DiANNA	24	58.0	49.0	445	ANN, PSSM, PSS	http://clavius.bc.edu/~clotelab/DiANNA
DISULFIND	26	60.2	54.5	446	RNN	http://disulfind.dsi.unifi.it
	3	56.0	49.0	SP39, SP41, SPX	2D-RNN, PSS, SA	
Dinosolve	27	73.4	82.9	215, 338, CASP9	ANN, PSSM, statistics	http://hpcr.cs.odu.edu/dinosolve

Table 3. Summary of other methods for disulfide connectivity pattern prediction in chronological order.

METHOD	REF.	R_b(%)	Q_b(%)	DATASET	DESCRIPTION	SOFTWARE
	31	56.0	56.0	SP39	MCSA	http://gpcr.biocomp.unibo.it
	29	85.5	87.0	PDBSelect, SPX	1-NN	
	30	87.0	–	260 UniProt	nRMR, FS, k-NN, PSSM	
	32	–	–	PDBCYS, SP39	Random forest	http://csbio.njust.edu.cn/bioinf/TargetDisulfide

Fariselli and Casadio[31] present a method based on MCSA and the Gabow's algorithm to solve the problem of maximum weight matching problem. The contact potential between each pair of cysteines is calculated with the Edmond–Gabow's algorithm. Hydrophobic and charged amino acids are taken into account for the prediction.

A summary of these methods is shown in Table 3. The best results are obtained by the method described by Vincent et al.[29] This 1-NN approach also includes evolutionary profiles that enhance the prediction accuracy.

Conclusion

The DCP problem can be considered as a previous step for the PSP problem. Once the cysteine bridges are identified and established, the protein conformational search space is highly reduced. In this paper, we present a compilation of the DCP methods based on soft computing techniques. Soft computing methods have shown to be well suited for the treatment of massive amounts of biological data.[33] In this problem, those methods that use evolutionary information from sequence alignments obtain better results than others.

Comparing the performance of the different approaches, we cannot draw clear conclusions to determine which is the best methodology. It depends on the data set used, the input features of the machine learning algorithm, among other factors. However, according to their excellent results, we can highlight two previously described approaches: a SVM method presented by Lin et al.[19] and Dinosolve.[27] These methods include the use of PSSM and DOC, two important features in DCP.

Although the prediction accuracy is improved in the latest years, existing approaches fail to obtain accurate models in DCP. Several methods achieve accuracies of about 80%–90%; however, the size of the data sets used in the experimentations is quite small and a general model to predict any protein disulfide connectivity is not found yet. Nowadays, DCP is still considered an unresolved problem, in terms of nonspecific approaches. As future lines of work, it is becoming ever more evident the important role of evolutionary information as input feature for DCP algorithms. Latest methods combine cysteine co-evolutionary analysis as a feature to enhance the predictions.[34] High-quality alignments and phylogenetic trees are also recently used by Raimondi et al.[35]

Author Contributions

Conceived and designed the experiments: AEMC. Analyzed the data: AEMC. Wrote the first draft of the manuscript: AEMC. Contributed to the writing of the manuscript: AEMC. Agree with manuscript results and conclusions: AEMC, JSAR. Jointly developed the structure and arguments for the paper: AEMC, JSAR. Made critical revisions and approved final version: JSAR. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. McMurry JE. *Organic Chemistry: With Biological Applications*. New Delhi: Brooks Cole; 2010.
2. Baldi P, Cheng J, Vullo A. Large-scale prediction of disulphide bond connectivity. *Adv Neural Inf Process Syst*. 2005;17:97–104.
3. Cheng J, Saigo H, Baldi P. Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*. 2006;62:617–29.
4. Savojardo C, Fariselli P, Martelli PL, Casadio R. Prediction of disulfide connectivity in proteins with machine-learning methods and correlated mutations. *BMC Bioinformatics*. 2013;14(1):10.
5. Chen BJ, Tsai CH, Chan CH, Kao CY. Disulfide connectivity prediction with 70% accuracy using two-level models. *Proteins*. 2006;64:246–52.
6. Becker J, Maes F, Wehenkel L. On the relevance of sophisticated structural annotations for disulfide connectivity pattern prediction. *PLoS One*. 2013;8(2):e56621.
7. Lin HH, Tseng LY. Prediction of disulfide bonding pattern based on a support vector machine and multiple trajectory search. *Inf Sci*. 2012;199:167–78.
8. Shilton AP, Parker MM, Palaniswami M. Prediction of cysteine connectivity using SVM. *Bioinformation*. 2005;1(2):69–74.
9. Song J, Yuan Z, Tan H, Huber T, Burrage K. Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics*. 2007;23:3147–54.
10. Savojardo C, Fariselli P, Alhamdoosh M, Martelli PL, Pierleoni A, Casadio R. Improving the prediction of disulfide bonds in Eu-karyotes with machine learning methods and protein subcellular localization. *Bioinformatics*. 2011;27(16):2224–30.
11. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 2000;28(1):45–8.
12. Gabow HN. An efficient implementation of Edmonds algorithm for maximum weight matching on graphs. In: Technical Report CU-CS-075-75. Department of Computer Science, Colorado University, Boulder, CO (USA); 1975.
13. Liu HL, Chen SC. Prediction of disulfide connectivity in proteins with support vector machine. *J Chin Inst Chem Eng*. 2007;38(1):63–70.
14. Lin HH, Tseng LY. DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding state of cysteines. *Nucleic Acids Res*. 2010;38:503–7.
15. Tsai CH, Chen BJ, Chan CH, Liu HL, Kao CY. Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. *Bioinformatics*. 2005;21(24):4416–9.
16. Chen YC, Hwang JK. Prediction of disulfide connectivity from protein sequences. *Proteins*. 2005;61:507–12.
17. Lu CH, Chen YC, Yu CS, Hwang JK. Predicting disulfide connectivity patterns. *Proteins*. 2007;67:262–70.
18. Zhu L, Yang J, Song JN, Chou KC. Improving the accuracy of predicting disulfide connectivity by feature selection. *J Comput Chem*. 2010;31(7):1478–85.

19. Lin HH, Hsu JC, Chen YF. Disulfide bonding pattern prediction server based on normalized pair distance by MODELLER. In: Proceedings on 2012 IEEE International Symposium on Computer, Consumer and Control (IS3C), Taichung (Taiwan);2012;581–84.
20. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292:195–202.
21. Kyte J, Doolittle R. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 1982;157:105–32.
22. Grantham R. Amino acid difference formula to help explain protein evolution. *J Mol Biol.* 1974;185:862–4.
23. Altschul SF, Madden TL, Schffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
24. Ferre F, Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics.* 2005;21(10):2336–46.
25. Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics.* 2004;5:653–9.
26. Ceroni A, Passerini A, Vullo A, Frasconi P. DISULFIND: a disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res.* 2006; 34:177–81.
27. Yaseen A, Li Y. Dinosolve: a protein disulfide bonding prediction server using context-based features to enhance prediction accuracy. *BMC Bioinformatics.* 2013;14(13):S9.
28. Martelli PL, Fariselli P, Malaguti L, Casadio R. Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.* 2002;11:2735–9.
29. Vincent M, Passerini A, Labbe M, Frasconi P. A simplified approach to disulfide connectivity prediction from protein sequences. *BMC Bioinformatics.* 2008;9(1):20.
30. Niu S, Huang T, Feng KY, He Z, Cui W. Inter-and intra-chain disulfide bond prediction based on optimal feature selection. Protein Pept Lett. 2013;20:324–35.
31. Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. *Bioinformatics.* 2001;17(10):957–64.
32. Yu DJ, Li Y, Hu J, Yang X, Yang JY, Shen HB. Disulfide connectivity prediction based on modelled protein 3D structural information and random forest regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2015;12(3):611–21.
33. Mitra S, Hayashi Y. Bioinformatics with soft computing. *IEEE Trans Syst Man Cybern C Appl Rev.* 2006;36(5):616–35.
34. Raimondi D, Orlando G, Vranken WF. Clustering-based model of cysteine co-evolution improves disulfide bond connectivity prediction and reduces homologous sequence requirements. *Bioinformatics.* 2015;31(8):1219–25.
35. Raimondi D, Orlando G, Vranken WF. An evolutionary view on disulfide bond connectivities prediction using phylogenetic trees and a simple cysteine mutation model. *Bioinformatics.* 2015;31(8):1219–25.