A review of algorithmic techniques for disulfide-bond determination

Rahul Singh

Advance Access publication date 28 March 2008

Abstract
Disulfide bonds play an important role in understanding protein folding, evolution, and in studies related to determining structural and functional properties of specific proteins. At the state-of-the-art, a large number of computational techniques have been proposed for determining disulfide bonds. Operating across the gamut of input data, from pure sequence-based information to spectra from mass spectrometry, these techniques provide researchers with a variety of methodological choices and trade-offs. Techniques for disulfide-bond determination are also underpinned by a rich variety of algorithmic formulations. Analysis of these algorithms can provide valuable cues towards choosing a particular technique and understanding its results. Further, their study is critical in developing the next generation of techniques. This paper discusses the importance and applicability of disulfide-bond determination in understanding protein structure and function and provides a detailed review of computational approaches to this problem.

Keywords: disulfide bonds; proteomics; mass spectrometry; machine learning; review

INTRODUCTION
Disulfide bonds (also called disulfide bridges or SS-bond) are one of the main cross-linkages present in proteins and are formed by the oxidation of cysteine residues. Disulfide bonds have been broadly characterized to be either structural or catalytic. Recent evidence [1, 2] also suggests a third type of disulfide bonds, termed allosteric disulfides, which regulates protein function in non-enzymatic ways by triggering a conformational change when it breaks and/or forms.

Structural disulfide bonds play an important role in the folding and stabilization of proteins as indicated by theoretical [3–5] and experimental studies [6, 7]. These bonds are also involved in the formation of certain structural motifs such as the cysteine knot [8] and in redox-active proteins, the CxxC motif [9]. On the other hand, catalytic bonds mediate thiol-disulfide interchange reactions in substrate proteins and are important for regulation of enzymatic activity [6, 10]. Examples of such bonds include those at the active sites of thiol-disulfide oxidoreductases [11]. Allosteric disulfide bonds, in contrast to catalytic disulfides, control the functioning of proteins by triggering changes in the intramolecular or inter-molecular protein structure. Essentially, these bonds act as switches for protein function [1] and have been implicated, among others, in the activation of the tissue factor (TF) glycoprotein [12].

An important property of the disulfide bonds is that they constrain the possible distance and angle values between $C^\beta$ and $S^\beta$ atoms of the cysteine residues participating in the bond. This reduces the entropy of the unfolded molecule [5, 13, 14] and influences the thermodynamics of protein folding. Disulfide bonds also stabilize the folded state of a molecule enthalpically by favourable local interactions [5]. The stabilization of the protein structure helps in reducing protein damage in presence of oxidants and proteolytic enzymes [1]. It has also been suggested that these bonds may sometimes decrease the stability of folded proteins by restricting energetically favourable conformations [1].
The relationship between SCOP superfamilies and disulfide-bond connectivity patterns has been studied in [15] and the connectivity patterns used to detect structural similarity [16]. Insights into several structural properties of proteins can also be obtained using disulfide-bond reactions [5]: first, they can be used to monitor the accessibility of reactive groups and the stability of the tertiary structure. This is because of the fact that disulfide-bond reactions are inhibited if the thiol groups or protein and mixed disulfide bonds are buried in the tertiary structure. Second, the proximity of specific sulfur atoms, as indicated by their participation in disulfide bonds, can be used to characterize protein structures in conjunction with other distance measuring methods. Finally, disulfide-bond reactions are sensitive to the electrostatic environment of the reactive groups, including the pH of the solvent and the pKₐ values of the thiol groups involved.

### BIOCHEMISTRY AND NOMENCLATURE

Disulfide bonds are usually formed when the sulphydryl (thiol) groups of cysteine residues become oxidized, forming a covalent bond derived from the coupling of thiol groups.

\[
R_1SH + R_2SH \rightarrow R_1SSR_2 + 2H^+ + 2e^- \quad (1)
\]

In proteins, the main biochemical reaction involved in disulfide bond formation is the thiol-disulfide exchange in which the thiolate anion \( R_1S^- \) displaces a sulphur atom of the disulfide bond \( R_2SSR_3 \) [5, 17, 18]. The displaced sulphur atom is released as a new thiolate and the new disulfide bond formed between the attacking thiolate and the sulphur atom.

\[
R_1S^- + R_2SSR_3 \rightarrow R_2S^- + R_1SSR_3 \quad (2)
\]

Two such thiol-disulfide exchange reactions with a redox reagent are involved in the formation and reduction of disulfide bonds with the first leading to the formation of a mixed disulfide bond between the protein and the redox reagent. It may be noted that a protein thiolate group may attack a disulfide bond in the same protein. Such intra-molecular thiol-disulfide exchange reactions lead to rearrangement or reshuffling of disulfide bonds within a protein. The reshuffling is characterized by a change in the location of the bonds. However, the number of such bonds remains unchanged.

Consider an arbitrary protein having \( n \) disulfide bonds. The corresponding disulfide bonded pairs are represented as \((x_1-x_{i+1}, x_2-x_{i+2}, \ldots, x_n-x_{2n})\), where \( x_i-x_{i+1} \) denote the sequence numbers of the cysteine residues constituting the \( i \)th disulfide bond. Similarly, the disulfide connectivity of the protein is given by \((N_1-N_{i+1}, N_2-N_{i+2}, \ldots, N_n-N_{2n})\), where \( N_i-N_{i+1} \) denote the relative order of the cysteine pair forming the \( i \)th disulfide bond. For a disulfide connectivity pattern \((N_1-N_2, N_2-N_3, \ldots)\), a related concept is that of its cysteine spacing pattern \((S_{ij}, S_{kl}, \ldots)\), where \( S_{ij} \) is defined to be the number of residues occurring between \( N_i \) and \( N_j \).

The pancreatic trypsin inhibitor protein (PDB ID: 1G6X) can be used to provide an illustration as shown in Figure 1. The ‘A’ chain of the protein is of length 58 and involves three disulfide bonds occurring between cysteine residues at positions 5 and 55, 14 and 38, and 30 and 51. This yields the set of disulfide bonded pairs: \( (5-55, 14-38, 30-51) \). The disulfide connectivity of this molecule is given by \((1-6, 2-4, 3-5) \) and its cysteine spacing pattern is \((49, 23, 20) \).

### DISULFIDE-BOND DETERMINATION: FORMULATIONS

Determining the disulfide bonding pattern for a given protein is a complex technical problem. Three classes of techniques can be broadly distinguished in this context: (i) Crystallographic techniques producing high-resolution three-dimensional structures of proteins, where the disulfide bonds can be observed directly. (ii) Algorithmic techniques that predict (or infer) the disulfide connectivity based on sequence data. (iii) Techniques combining experimental and algorithmic methods to detect covalently linked peptides by analysing a mixture of peptides obtained by targeted digestion of an intact protein using mass spectrometry.

Determining the three-dimensional structure of a native intact protein is a highly complicated challenge. For this, a subdomain of the protein that may form crystals and is sufficiently soluble needs to be expressed and studied. In this review we do not focus on this set of techniques and refer the interested reader to the large volume of material in texts and papers [19–21] (and references therein). Techniques for algorithmic prediction of disulfide connectivity, falling in the second of the aforementioned classes, are based on utilizing local and global
sequence characteristics in conjunction with a machine-learning strategy. Essentially, the relationship between these characteristics and the disulfide connectivity (or related properties) is inferred by a machine learning algorithm based on a ‘training set’ consisting of sequences for which the disulfide connectivity is known. The connectivity for a new sequence is predicted based on its local and global features and the relationship between the features and the connectivity patterns learnt by the algorithm from the training data. The third class of techniques has attracted significant recent attention due to the availability of modern mass spectrometers having high sensitivity and mass accuracy. However, manual derivation of putative disulfide-bonding patterns and their comparison with mass spectra becomes complex as the number of cysteine residues increase [22]. This underlines the necessity for algorithmic solutions. In this paper, we focus on the latter two classes of techniques.

With respect to the use of sequence-based information, two correlated problem formulations are conceivable in the context of disulfide bond determination. These include:

- **Residue classification**: Classifying cysteine residues into those that are bonded and those that are not bonded. This formulation may subsume the problem of ‘chain classification’, i.e. distinguishing protein chains that contain disulfide bridges from those that do not.

- **Connectivity prediction**: Determining all pairs of cysteine residues that are connected to each other by disulfide bonds. A component of this formulation is the problem of ‘bridge classification’, which requires determining if a specific pair of cysteine residues are joined by a disulfide link or not.

### Residue classification

Since cysteines can occur in either oxidized or thiol forms, techniques for residue classification seek to predict the disulfide-bonding state of each cysteine from sequence information. Research in this area has broadly occurred along two lines; the first involves decision making based on analysis of the statistical frequency of amino acid residues in neighbourhoods around the cysteines [23, 24]. The second class of techniques employs encoding of the sequence environment followed by machine learning in supervised settings. A summary of the methods, their availability, data sets, testing procedure and reported accuracy is presented in Table 1.

### Methods based on analysis of amino acid frequency

Of techniques in the first category, the statistical frequencies of oxidized and reduced cysteines on the surface and in the interior of the proteins were compared in [24] on a dataset of 81 proteins that contained 51% free cysteines, 27% half-cysteines, and 5% both free and disulfide-bonded cysteines. Some key observations from this study were: (i) the distribution of half cystines and cysteines between the surface and the interior of the molecule were almost identical. (ii) In proteins containing both cystine and cysteine residues, the cysteines were typically bonded. (iii) A significant difference existed between the relative conservation scores between the bonded and the free forms of cysteines. The conservation scores reflected specific properties (hydrophobic, positive, negative, polar, charged, small, tiny, aliphatic, aromatic and proline and their negation) shared at a position. (iv) Statistical correlations appeared more pronounced when cysteines were grouped as bonded versus free rather than as cysteines versus cystines. The strategy in [24] was based on the observation that
Table I: Summary of methods for predicting the bonding state of cysteines (Due to differences in data sets, the prediction accuracy values should not be directly compared.)

<table>
<thead>
<tr>
<th>Method</th>
<th>Availability</th>
<th>Data set and testing procedure</th>
<th>Prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output of logistic function with local and/or global descriptors [25]</td>
<td></td>
<td>559 proteins with each sequence containing at least two cysteines. From a total of 2882 cysteines, 1270 were involved in disulfide bonds and 1612 were not. Predictions tested by 5-fold cross validation.</td>
<td>~78% using global descriptors (two or more cysteines) in sequences ~83% for sequences with only one cysteine 67.3% (local descriptor with binary encoding) 70.1% (local descriptor with encoding of amino acid occurrence frequencies)</td>
</tr>
<tr>
<td>Feedforward Neural Networks [26]</td>
<td></td>
<td>689 Cysteine-containing sequences from 128 protein structures (379 in disulfide-bonded state and 310 in non-disulfide-bonded state). Of these, 659 sequences used in seven independent training-testing sets for evaluation.</td>
<td>~81.4% for disulfide bonding and 80.0% for non-disulfide-bonded cases</td>
</tr>
<tr>
<td>Feedforward Neural Networks with multiple sequence profiles [27]</td>
<td>Cyspred, <a href="http://gpcr.biocomp.unibo.it/cgi/predictors/cyspred/pred_cyspred.cgi">http://gpcr.biocomp.unibo.it/cgi/predictors/cyspred/pred_cyspred.cgi</a></td>
<td>2452 cysteine containing sequences from 641 protein structures (842 in disulfide-bonded state and 1610 in non-disulfide-bonded state). Predictions tested using 20-fold cross validation. Training-test sample identity was maintained at &lt;30%.</td>
<td>71.8% using single sequence input 78.2% with multiple sequences ~80% with multiple sequences along with conservation weight and relative entropy 81% with a jury of networks</td>
</tr>
<tr>
<td>Hybrid classifier: Probabilistic combination of SVMs [30]</td>
<td></td>
<td>716 proteins containing 4859 cysteines (1820 in disulfide-bonded state and 3039 in non-disulfide-bonded state). Predictions tested by 5-fold cross validation.</td>
<td>83.64% with a two-stage classifier ~82.2% with SVM using a combination of local and global features</td>
</tr>
<tr>
<td>Hybrid classifier: HMM and Neural Networks [31]</td>
<td></td>
<td>969 proteins with 4136 cysteine-containing segments: 1446 in disulfide-bonded state and 2690 in non-disulfide-bonded state. Predictions tested using 20-fold cross validation.</td>
<td>88.0% on entire data set 87.4% on subset of the data with chains containing two or more cysteines</td>
</tr>
<tr>
<td>Hybrid classifier: Penalized regression and SVM [35]</td>
<td></td>
<td>639 cysteine-containing protein chains. Predictions tested by leave-one-out cross validation.</td>
<td>84.1% overall accuracy</td>
</tr>
</tbody>
</table>
oxidized and reduced forms of cysteines rarely co-occurred, except in cases when the cysteine was covalently bonded to heteroatoms, prosthetic groups or other amino acids in active sites. Therefore, if a majority of the predicted cysteines belonged to one oxidation state and had high conservation score to the group of oxidized cysteines and lower conservation scores to the group of reduced cysteines, then the other cysteines in the molecules were assigned to the same oxidation state. In cases where the predicted number of reduced and oxidized cysteines was equal, the logarithm of the averages of the relative conservation scores for the predicted bonded cysteines and predicted free cysteines were compared. The cysteines were predicted to be oxidized or reduced if the corresponding relative conservation score was larger. Incorrect predictions occurred if both forms of cysteines occurred in the same protein or the normalized conservation of the cysteines was near the prediction threshold. The strategy could also fail if certain free cysteines were strongly conserved for functional reasons. The prediction accuracy of the method, using leave-one-out crossvalidation, was reported in the 82–84% range depending on parameter selection.

A study investigating the relative contribution of local features drawn from regions flanking the cysteine and global features (overall amino acid composition of the protein) to the prediction of bonding states of cysteines was conducted in [25]. A single logistic function equivalent to a neural network without a hidden layer was trained to perform the classification. The local features were captured using a window of size 2n + 1, centred on the cysteine of interest. Two alternate encodings of residues in the local neighbourhood were considered. In the first, each residue was encoded by a 20-dimensional binary vector, identical to the encoding scheme in [26], where the element corresponding to the specific residue present at a position was set to 1 and all other elements were set to 0. In the second encoding, each residue in the window was represented by a number reflecting its relative abundance in the dataset. This number was defined as \( \log \left( \frac{f_{l}}{f_{k}} \right) \), where for protein \( k \), \( f_{l} \) denoted the frequency of amino acid \( l \) at a given position in the window and \( f_{k} \) was the frequency of amino acid \( k \) in the entire dataset. Essentially, the second descriptor captured the over (under)-representation of an amino acid at a specific position in the neighbourhood of the cysteine. Additionally, the position of the cysteine within the sequence, in terms of its distance to the sequence extremities was also used. The global features were encoded by a 23-element vector. In it, the first 20 elements described the relative abundance of each residue. The 21st element corresponded to the normalized size of the protein. For a given protein, the last two elements contained the number and frequency of occurrences of cysteines in the protein. The descriptor vectors, individually or in combination, were passed to the logistic function. If the output exceeded a threshold, the cysteine was predicted to be oxidized otherwise it was predicted to be free. The threshold itself was defined by minimizing the prediction error rate. Prediction accuracies were tested using 5-fold cross-validation and varied across a broad range depending on descriptor choice and data complexity. For instance the accuracy was around 83% for sequences with one cysteine and 78% when global descriptors were used and two or more cysteines were present (Table 1).

**Methods based on sequence environment encoding and machine learning**

In one of the early works in this area Muskal, Holbrook and Kim [26] proposed the use of a neural network to learn the influence of the local sequence environment on whether a cysteine participated in a disulfide bond. Specifically, a feedforward network architecture was used with no hidden layers and the network weights were learned using back-propagation. A cysteine residue, flanked by symmetrical segments that constituted its environment, was classified as disulfide-bond forming or free, depending on the relative values of the two output neurons. The authors also proposed a novel encoding-representation scheme which has extensively been used in later works. In this scheme, each cysteine constituted the centre of a window. Each residue inside the window was encoded by a 21-node vector. In this vector, each of the 20 nodes corresponded to one of the 20 amino acids. The final node was used as an indicator in case the window overlapped a break in the chain or one of the termini. Within the window, a specific amino acid was encoded by assigning a value of 1.0 to its corresponding node and the value of 0.0 to all the other nodes (except the indicator node). A training cycle was preceded by random initialization of the network weights. Subsequently, each example in the training set being used was presented to the network and the network weights updated. Once the total
error over all samples in the training set converged to a minimum, the network weights were fixed and evaluations conducted on the remaining data. Different window sizes corresponding to different sizes of the flanking sequences were experimented with. The authors stopped after 14 flanking amino acid positions since longer flanking sequences lead to overfitting of the data. Prediction accuracy using independent training and testing sets was around 81% for disulfide-bonded cystines and 80% for non-disulfide-bonded cystines.

The learning framework proposed in [26] was extended in [27] by considering eight different encodings of the input sequences. Evaluations were based on a dataset containing 2452 cysteine-containing segments and disulfide-bond assignments were based on the DSSP program [28]. The basic encoding-representation scheme (employed in context of a single sequence input) was identical to that proposed in [26]. The other seven encodings involved different features captured from multiple sequence profiles. In the case of multiple sequences, the sequence profile of the cysteine-containing segments taken from the HSSP files of the proteins were used [29]. A 21-element vector was used to encode each residue as in [26]. However, the first 20 elements in the vector represented the frequency of occurrence of the 20 amino acid residues in the alignment. In computing the frequency, the central cysteine was also taken into account. In addition to the frequency of occurrence, other features that were considered included: charge, hydrophobicity, conservation weight and relative entropy, and their combinations. The model was evaluated using 20-fold cross-validation. During this process, identity between training and test samples was maintained to be less than 30%. The prediction quality was measured using a varied set of indices: (i) network accuracy (30%). The prediction quality was measured using validation. During this process, identity between the model was evaluated using 20-fold cross-validation and their combinations. The interpretation of $p(SS)$ and $o(SS)$ are analogous.

Using a single sequence as an input, the highest accuracy (Q3 value of 71.8%) was obtained with window size of 13-residues. Marked improvement in performance was observed using information based on multiple sequence alignment. For a window size of 17 residues, the highest accuracy (Q3 value of 78.2%) was obtained. The clear advantage of using multiple sequences was also underlined by the fact that for a window size of 13-residues, the prediction accuracy was 77.5% and clearly exceeded the accuracy obtained for the same window size on single sequence input. The use of physico-chemical characteristics (charge, conservation weight, hydrophobicity and relative entropy) and their combinations were found, in certain cases, to improve prediction accuracy scores, correlation scores, and probability of correct bond predictions. Specifically, consideration of conservation weight and relative entropy, as well as their combinations with charge and hydrophobicity lead to prediction accuracies of ~80%. The authors also used a jury of networks to improve the prediction accuracy to 81%.

In [30], an approach was proposed that combined elements from techniques based on analysis of statistical frequency of residues as well as the use of a learning-based classification algorithm. A two stage classifier was proposed; in the first stage, global sequence descriptors were used to classify the sequence as having all, some, or none of its constituent cysteines participating in disulfide bonds. In the second stage, a classifier was trained to predict the bonding state of specific cysteines in sequences that were predicted to have a combination of free and disulfide-bonded residues. The classification framework was implemented as a probabilistic combination of support vector machines. The descriptors used were similar to those in [25] and consisted of a 24-element vector. The first 20 elements of the vector represented the relative abundance of specific amino acid residues in the sequence. The 21st element represented the ratio of

In Equations (3)–(6), $P$ denotes the number of correct predictions, $N$ the total number of predictions (correct and incorrect), $n$ is the number of correctly rejected assignments for a particular state and $p$ the number of correct predictions for that state, $u$ and $o$ are the number of under and over predictions for that state, and $p(SS)$ and $o(SS)$ are the number of correct predictions and over predictions for disulfide-bond forming cysteines. The interpretation of $p(SS)$ and $o(SS)$ are analogous.

$$Q^3 = \frac{P}{N}$$

$$C = \frac{(p \times n - u \times o)}{\sqrt{[(p + u)(p + o)(n + u)(n + o)]}}$$

$$Pt(SS) = \frac{p(SS)}{p(SS) + o(SS)}$$

$$Pt(SH) = \frac{p(SH)}{p(SH) + o(SH)}$$
the length of the sequence to the average length of sequences in the dataset. The ratio of the number of cysteines in the sequence to the maximum number of cysteines observed in the training set constituted the 22nd element. The 23rd element represented the ratio of the number of cysteines in the current sequence to the size of the sequence. The final element of the vector acted as a flag in case the sequence had an odd number of cysteines. A variety of parameters were evaluated in the experiments. Salient among these were the size of the local window and type of the classifier. Four different classifiers were evaluated: (i) A single SVM classifier using a polynomial kernel which took a local window of multiple alignment profiles as input. (ii) A single Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel using the aforementioned 24-element global descriptor as input. (iii) A polynomial kernel SVM using a combination of local and global descriptors as input and (iv) A two-stage classifier. Predictions were tested using 5-fold cross validation.

The highest accuracy of 83.64% was reported for the two-stage classifier with a window size of 9. Precision and recall values were also reported by using the DSSP program to identify disulfide bonds. For the parameters yielding the highest accuracy of prediction, the precision was 82.09% and recall 72.43%. The polynomial kernel classifier using combination of local and global descriptors as well as the two stage classifier consistently produced higher prediction accuracies than the first two classifiers. Interestingly, the first classifier which used a local window had an average accuracy of 79.3% (across window sizes). The reported results show it to outperform the second classifier (with the global descriptor), for which the average accuracy across all window sizes was 76.5%. Further, the prediction accuracy of this local descriptor-based SVM classifier was also better than the accuracy of 71.8% obtained by the neural–network based classifier in [27], which used single sequences as an input.

A hybrid classifier built using a hidden Markov model (HMM) and neural network was proposed in [31]. In the classifier architecture, the HMM probability parameters were estimated through outputs of state-specific neural networks [32]. The HMM in [31] was defined by four states and specific inter-state transitions to ensure that the number of cysteines predicted in the disulfide-bonding state is even. The neural network, in a setting similar to that in [27] was used to learn the local environment conductive to the bonding or non-bonding state of the cysteine residue on which the window is centred. The output of the neural network was then used as emission probabilities of the 4-state HMM. The key idea of this method was in recognizing that the prediction of the state of a specific cysteine in the sequence by a neural network, could not take into account the possible presence of other cysteines in the sequence and their respective states. Therefore, the HMM was used to enforce this global constraint. Experiments using the hybrid classifier reported prediction accuracies of 88% on the entire dataset and 87.4% on the dataset containing two or more cysteines based on 20-fold cross validation. These values were significantly higher than the corresponding prediction accuracies obtained just using the neural–network classifier (80.4% for the entire dataset and 80.1% for the dataset with two or more cysteines). The reader may note that using the neural network alone, was similar to the setting of [27] and yielded comparable performance. This method was used in [33] to study the disulfide content of *Escherichia coli, Aeropyrum pernix, Thermotoga maritima*, and *Homo sapiens*.

The idea of using both local and global descriptors was also used in [34], where an SVM was employed. In addition to local descriptors, global descriptors such as cysteine state sequences and amino acid composition were used. For *n* cysteines, there are $2^n - 1$ possible state sequences (when inter-chain bonds are excluded). Two approaches were used to combine the SVM with cysteine state sequence. In the first, a single SVM was trained with multiple feature vectors to produce the state probability for a given cysteine. In the second, multiple SVMs were each trained with a different feature vector to generate averaged state probability for the given cysteine. Subsequently, for both approaches a branch and bound method was used to find the transition sequence with the highest probability. Experimental results showed that when SVMs were coupled with cysteine state sequences, the overall prediction accuracy (based on 20-fold cross validation) ranged from 88 to 90%. Further, with this combination the specificity in predicting bonded cysteines significantly increased (88–91%), when compared with predictions using SVM and local sequence window (71%). Another example of the use of hybrid classifiers is provided in [35], where, using overall amino acid composition as the feature, penalized linear regression was used in the first stage to classify...
a protein into two classes, namely those containing reduced cysteines and those with oxidized cysteines. For the former class, all cysteines were predicted to be reduced cysteines. For the latter, a second SVM-based classifier with a polynomial kernel was used to predict the disulfide-bonding states of the cysteines. Experimental results were based on 639 cysteine-containing protein chains from PISCES Culled PDB [36], containing a representative dataset of accurately resolved non-homologous proteins. In a leave-one-out crossvalidation setting, the highest prediction accuracy obtained was 84.1%. For the corresponding parameter settings, the prediction accuracy for the class of oxidized cysteines was 87.8% and that for the class of reduced cysteines was 77.8%. The authors also presented a comparison of the protein structures in terms of the two classes of oxidized and reduced cysteines and their CATH classification. Of the molecules in the oxidized cysteine class, 53.0% were mainly \(\beta\), 29.3% \(\alpha\) and \(\beta\), and 16.8% mainly \(\alpha\). Of those in the reduced cysteine class, 48.9% were \(\alpha\) and \(\beta\), 30.9% mainly \(\alpha\), and 19.5% mainly \(\beta\). Since \(\beta\)-sheets are less stable, the observed frequencies underlined the role disulfide bonds play in maintaining the native structure of the protein. In [37] a linear discriminator based on dipeptide composition was used to classify the disulfide-bonding state of cysteines. Prediction accuracies of 89.1% were reported using leave-one-out crossvalidation on a dataset containing 8114 cysteine containing segments from 1856 non-homologous proteins with resolved structures.

It should be noted that non-free cysteines that are not involved in disulfide bonds usually bind prosthetic groups (such as a metal ion). Techniques such as [38] have been developed to distinguish between disulfide or metal-bonding cysteines.

**CONNECTIVITY PREDICTION**

A disulfide-connectivity pattern can be represented by an undirected graph \(G(V, E)\), where the set of vertices \(V\) corresponds to the set of bonded cysteines and an edge \(e \in E\) corresponds to a disulfide bond between the cysteine residues (vertices) it joins. Since a cysteine can be bonded to only one other cysteine, the degree of any vertex, defined as the number of edges incident to it, is 1. Further, if \(B\) disulfide bonds are present, the number of cysteines is constrained to be \(2B\).

Given an even number of cysteines, putatively involved in disulfide bonds, the problem of connectivity prediction involves determining the correct connectivity pattern amongst all the possible alternatives [39]. In general, given \(2m\) cysteines, the number of possible connectivity patterns between them can be obtained as:

\[
\frac{(2m)}{2} \frac{(2m-2)}{2} \frac{(2m-4)}{2} \cdots \frac{2}{2} = (2m-1)! = \prod_{i=1}^{m} (2i-1)
\]

The combinatorial growth in the number of possible connectivity patterns, as the number of cysteines increases can be gauged from the fact that for \(m\) equaling 2, there are 2 possible patterns. However, for \(m\) equaling 5 and 6, the number of patterns increases to 945 and 10 395, respectively.

**Measures for performance assessment**

If the set of disulfide bonds are denoted by \(P\) and the set of cysteines not forming disulfide bonds by \(N\), then true positive (TP) predictions occur when disulfide bonds that exist are correctly predicted. False negative (FN) predictions occur when bonds that exist are not predicted as such. Similarly, a true negative (TN) prediction correctly identifies cysteine pairs that do not form a bond. Finally, a false negative (FP) prediction, incorrectly assigns a disulfide link to a pair of cysteines, which are not actually bonded. Based on these definitions, various measures can be used for evaluating the performance of techniques. These include:

Precision or Accuracy \((Q_p) = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}\) (8)

Recall or Sensitivity \((Q_r) = \frac{TP}{P}\) (9)

Specificity \((Q_s) = \frac{TN}{N}\) (10)

Fraction of correctly assigned proteins \((Q_p) = \frac{\text{Number of proteins with entire disulfide connectivity correctly predicted}}{\text{Total number of proteins}}\) (11)

\[\text{Matthew’s correlation coefficient } (c) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}\] (12)
Of these, $Q$, and $Q_p$, can be seen to have been broadly used in the literature. It should be noted that $Q$ is defined over all disulfide bonds while $Q_p$ measures the fraction of proteins for which the entire disulfide connectivity was correctly determined. Since, only the proteins for which the entire disulfide connectivity is correctly predicted are considered in computing $Q_p$, it is the more stringent performance measure.

In their seminal work, Fariselli and Casadio [39] used a graph-based approach to determine the disulfide connectivity pattern by solving the maximum weight matching problem. Their insight lay in constructing a completely connected graph $G$ on $m$ vertices, corresponding to the $m$ cysteines. Each of the $(m(m-1)/2)$ edges in this graph had non-zero weights corresponding to some measure of the propensity of bond formation between the corresponding cysteine residues. Subsequently, the disulfide connectivity was defined as the solution of the maximum weight perfect matching problem on $G$. The perfect matching of a graph is defined as a subset of its edges, such that each vertex in $G$ has only one edge from the subset incident to it. The maximum weight perfect matching, finds a perfect matching that has the largest total weight of the edges. This problem can be formulated as a linear programming problem and solved in $O(V^3)$ time by using the WMATCH implementation (WMATCH Solver: http://elib.zib.de/pub/Packages/mathprog/matching/weighted/) of the Edmonds–Gabow algorithm [40].

The data set used for evaluation consisted of non-redundant sequences derived from the SWISS-PROT database (release no. 39). To ensure quality it only included sequences containing information from PDB for which intra-chain disulfide bonds were annotated. Further, disulfide assignments described as ‘by similarity’, ‘probable’, or ‘potential’ were excluded. It may be noted that this dataset along with the rules for generating it, have been widely used by other authors and is considered a de facto benchmark. It is commonly abbreviated as SP39. Four different formulations of contact potential were used for weighting the edges in [39]: (i) The contact potential of Mirny and Shakhnovich [41], which is designed for protein folding and threading. (ii) A contact potential derived by constrained optimization that maximized the difference between scores for correct cysteine pairing and incorrect ones. (iii) An odds ratio contact potential and (iv) a contact potential obtained using Monte-Carlo simulated annealing. The last of the contact potentials, led to the best results, with $Q$ scores varying between 56% and 21% across four data sets having varying number of disulfide bonds. In subsequent research [42], a neural network was used for labelling edges of the graph $G$ by pair-wise interaction values, leading to improved prediction capability as well as reducing the training time.

In [43], a recursive neural network (RNN) was used for scoring undirected graphs that represent connectivity patterns by their similarity to the correct graph. Each vertex in the connectivity graph contains a description of the local environment of the bonded cysteine. Specifically, 20-element vectors corresponding to a bit-string environment was represented using a bit-string vector of amino acids. Evolutionary information was included by considering position specific amino
acid frequencies in profiles obtained using PSI-BLAST. Secondary structure information was encoded for helices, coils and sheets using a bit-string representation. Di-residue frequencies which reflected the occurrence of an amino acid at a position conditional to the occurrence of another amino acid at position in the sequence were also used. The values of (fraction of the protein for which the prediction was correct), was comparable or improved upon the results reported in [43]. Sensitivity or values for data sets containing 2, 3, 4 and 5 disulfide bonds were 68%, 60%, 73%, and 38%, respectively.

The idea of RNN pioneered in [43] formed the basis of the DISULFIND prediction server [46]. In it, the bonding state of cysteines was predicted using an SVM classifier based on local and global features. Subsequently in a refinement stage a bi-directional recurrent network was used to predict a globally correct sequence of bonding state assignments. Furthermore, the number of bonding states was enforced to be even and the most likely set of bonding states was computed using a Viterbi algorithm. Prediction was carried out by running a trained neural network on all possible connectivity patterns and by selecting the one having the highest score (in practice at most 5 disulfide bonds could be predicted). Using predictions in a leave-one-out setting, scores of and were reported on a dataset containing 2–5 bonds.

An integrated approach to disulfide-bond prediction was proposed in [47]. In it the chain classification problem was addressed using evolutionary information and kernel methods. Subsequently, the directed acyclic graph-RNN [48] was used to predict the bonding probability of each cysteine pair. During this stage, information about the secondary structure and solvent accessibility was also utilized. Finally, disulfide connectivity was inferred by graph matching: if the bonded states of the cysteines were known, the Edmond–Gabow algorithm was applied. The edge weights were the probabilities, computed by the predictor, of the corresponding bond. If the bonded state of the cysteines was not known, the number of bonds was predicted using simple regression. The Edmond–Gabow algorithm was subsequently run on the nodes and the final result pruned by removing a certain number of low-probability edges based on the estimated number of bonds. Experimental evaluations showed that the method could classify cysteines with equalling 89% and equalling 87%. The disulfide connectivity patterns were predicted with equalling 51%.

Since proteins with similar disulfide bonding patterns share similar folds, the separation between oxidized cysteine residues, called the cysteine separation profile (CSP), can be a useful predictor of disulfide connectivity. Approaches based on this idea encode the separation amongst cysteine residues as a vector. For a sequence with unknown disulfide connectivity, its CSP vector is compared with sequences where the connectivity is known. Similar separation profiles are assumed to have similar connectivity pattern. This methodology was used in the work of Zhao et al. [49]. Given a protein with cysteine residues , its CSP was defined as:

\[
\text{CSP}(P) = (C_2 - C_1, C_3 - C_2, \ldots, C_{2n} - C_{2n-1}) = (P_1, P_2, \ldots, P_{2n-1})
\]

The dissimilarity or divergence between the CSPs of two proteins, and , could then be found as:

\[
\text{CSP}(P) - \text{CSP}(Q) = \sum_i |P_i - Q_i|
\]

The disulfide connectivity of a protein that was being determined was predicted to be same as that of a database protein having the most similar CSP. In spite of its conceptual simplicity, results show the method to perform well. For instance, the average value for data sets containing 2–5 disulfide bonds, with redundant sequences removed and members having less than 30% sequence identity was 49%. CSPs were also used by Tsai et al. [50] to predict disulfide connectivity. In this work, SVM was employed to define the connectivity potential between cysteines. Two descriptors were considered: a local sequence profile and CSP. The disulfide connectivity was found by solving the maximum-weight matching problem. Various scaling of the distances between oxidized cysteines were empirically tested. The best results ( and equalling 70% on the dataset containing 2–5 bonds), were obtained when distances between cysteines were normalized using the logarithmic function.

In [51], the CSP was considered along with the amino acid content and the coupling between local sequence environments as three types of feature vectors. Each distinct disulfide pattern was considered as a class and multiple support vector machines were used to predict the disulfide patterns. A detailed experimental analysis of the efficacy of the features
Table 2: Summary of methods for connectivity prediction

<table>
<thead>
<tr>
<th>Method</th>
<th>Availability</th>
<th>Data set and testing procedure</th>
<th>Prediction accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximal-weight graph matching [39]</td>
<td>By request from authors</td>
<td>SP39: Sequences with 2–5 verified intra-chain bonds selected from SwissProt release 39. Total 726 proteins. Sequences were divided into 4 sets with &lt;30% homology. Four-fold cross validation used.</td>
<td>Qp between 56% (158 chains) and 2% (44 chains) and Qc between 56% (158 chains) and 21% (44 chains) using Monte-Carlo derived potentials</td>
</tr>
<tr>
<td>Recursive Neural Network [43]</td>
<td><a href="http://neural.dsi.unifi.it/cysteines">http://neural.dsi.unifi.it/cysteines</a></td>
<td>SP39. All chains with more than 10 oxidized cysteines were excluded. Four-fold cross validation.</td>
<td>Qp = 44% and Qc = 49% on dataset with 2–5 disulfide bonds</td>
</tr>
<tr>
<td>Secondary structure and di-residue frequency [44, 45]</td>
<td><a href="http://clavius.bc.edu/~clotelab/DIANNA">http://clavius.bc.edu/~clotelab/DIANNA</a></td>
<td>SP39 and other data sets.</td>
<td>Qp = 51% and Qc = 58% on dataset with 2–5 disulfide bonds</td>
</tr>
<tr>
<td>SVM and recursive neural network [46]</td>
<td><a href="http://disulfind.dsi.unifi.it">http://disulfind.dsi.unifi.it</a></td>
<td>SP39 with leave-one-out cross validation.</td>
<td>Qp = 54.5% and Qc = 60.2% on dataset with 2–5 disulfide bonds</td>
</tr>
<tr>
<td>Cysteine separation profile [49]</td>
<td>By request from authors</td>
<td>SP39 dataset, SP39-ID30 dataset (sequences having &gt;30% identity removed). SP43 dataset: SwissProt release 43 (overlap from SP39 excluded). Sequences have &lt;25% identity.</td>
<td>Qp = 49%, Qc = 52% on dataset with 2–5 disulfide bonds</td>
</tr>
<tr>
<td>Cysteine separation [50]</td>
<td><a href="http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/">http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/</a></td>
<td>SP39 with 4-fold cross validation.</td>
<td>Qp = 63%, Qc = 70% on dataset with 2–5 disulfide bonds with logarithmic scaling of the distance between oxidized cysteines</td>
</tr>
<tr>
<td>SVM with three types of features: amino acid content (A), cysteine–cysteine coupling (S) and cysteine separation profile (D) [51]</td>
<td></td>
<td>SP39 4-fold cross validation used.</td>
<td>On dataset with 2–5 disulfide bonds: Qp = 50%, Qc = 54% using D Qp = 45%, Qc = 48% using S Qp = 39%, Qc = 42% using A Qp = 55%, Qc = 57% using linear feature combination: D + A + 0.001S Overall (using hybrid method): Qp = 73.9%, Qc = 79.2%</td>
</tr>
<tr>
<td>Cysteine pair representation. SVM based prediction of bonding state with genetic algorithm-based feature selection [52]</td>
<td></td>
<td>SP39 with 4-fold cross validation.</td>
<td>Qp = 70% (with 2-level SVM)</td>
</tr>
<tr>
<td>Two-level framework with both pair-wise and pattern-wise encoding. SVM used in both levels [53]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison with an annotated database [54]</td>
<td><a href="http://research.i2r.a-star.edu.sg/CysView/">http://research.i2r.a-star.edu.sg/CysView/</a></td>
<td>Prediction performed by comparison of query sequence with 2350 SwissProt entries (release 43.1) with experimentally determined disulfide bonds.</td>
<td>Qp = 78.6%. Method only works when input data is similar to any of the 2350 SwissProt entries used as reference</td>
</tr>
</tbody>
</table>
was also conducted. Of the three features, the CSP provided the best overall results in terms of $Q_p$ and $Q_c$. For instance, $Q_p$ equaled 50% and $Q_c$ equaled 54% for the set containing 2–5 disulfide bonds. The values for these metrics using cysteine–cysteine coupling were 45 and 48%, respectively. In contrast, the amino acid composition yielded a $Q_p$ score of 39% and a $Q_c$ value of 42%. The authors also experimented with linear combinations of features. This led to the highest scores of $Q_p$ = 55% and $Q_c$ = 57% for the data set containing 2–5 disulfide bonds. Since this method dealt with the prediction problem by treating each disulfide pattern as a class, it was directly affected by the combinatorial increase in the number of disulfide patterns as the number of cysteines increased.

This issue was addressed in the work of Lu et al. [52]. The central idea of this research consisted in representing disulfide connectivity in terms of cysteine pairs rather than disulfide patterns. A single cysteine pair and two cysteine pairs were used to represent disulfide bonds. For example, for a sequence having three disulfide bonds denoted by {C1C2, C3C4, C5C6}, 15 states define the disulfide pattern using a single cysteine pair to define the disulfide pattern (in the following ‘$+$’ denotes a disulfide bond and ‘×’ denotes the lack of a disulfide bond between the corresponding cysteines): C1+ C2, C1× C3, C1× C4, C1× C5, C1× C6, C2× C3, C2× C4, C2× C5, C2× C6, C3+C4, C3× C5, C3× C6, C4× C5, C4× C6, C5+C6. The reader may note that this reduced the number of classification classes. For instance, for 4 disulfide bonds the number of states using a single cysteine pair (two cysteine pairs) was 28 (70). For 5 disulfide bonds, the corresponding number of states was 45 (210). The number of states was significantly less than the number of disulfide patterns for 4 and 5 bonds (105 and 945, respectively). Lu et al. used the same features as in [51] namely, CSP, amino acid content and the coupling between local sequence environments. A genetic algorithm was used for feature selection and the bonding state of cysteine pairs was predicted using support vector machines. Finally, the disulfide patterns were determined from the connectivity matrix constructed from the predicted bonding states of the cysteine pairs. On the dataset used in [39, 43, 51], the best overall performance ($Q_p = 73.9\%$ and $Q_c = 79.2\%$) was obtained. In [53] a two level framework is used where both pair-wise and pattern-wise encodings are considered. Reported results show $Q_p$ score of 0.7 on the SP39 dataset.

Other approaches for predicting connectivity patterns include comparisons with an annotated database, as done in the CysView server [54]. Finally, we point the reader to a conceptually different approach to this problem which arose as an offshoot of determining the topology of β-sheets in proteins in the work of Klepeis and Floudas [55]. In it, bonds between cysteine residues (as well as other residue-to-residue contacts) were predicted by maximizing hydrophobic contact energy terms. The maximization was done using integer linear programming, which allowed for global optimization. However, unlike the other methods reviewed by us, the experimental validation was done on two short sequences having two and three disulfide bonds.

### DISULFIDE-BOND DETERMINATION BY MASS SPECTROMETRY

Advances in instrumentation allowing analysis of minute amounts of peptides in complex mixtures along with development of genomic databases that can be searched using mass spectrometry data have established mass spectrometry as the primary method for protein identification at the state of the art. A variety of algorithmic problems have been investigated in this context, with many of the solutions publicly available [56, 57]. For the problem of disulfide connectivity determination, however, the field is relatively speaking, still in its early stages.

The basic strategy for determining disulfide bonds using mass spectrometry consists of the following steps [58, 59]: (i) The protein of interest is cleaved in its non-reduced state between as many of the half cysteine residues as possible using proteases like pepsin or trypsin. For example, trypsin can be used to cleave a protein after each lysine and arginine residues. (ii) The resultant disulfide-linked peptides are separated and analysed by electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI). These mass spectrometry techniques allow peptide and protein molecular ions to be put into the gas phase without fragmentation [58]. The analysis is a two-step process and involves measuring the mass-to-charge ($m/z$) ratio of the ionized peptide (also called the parent or precursor ion) along with measurement of the $m/z$ ratio of the product ions. Subsequently, the peptides ions are fragmented to...
determine the identity and sequence of the protein and its modifications.

In spite of the seeming simplicity of this process, the determination of disulfide bonds by matching putative models to mass spectra is complex in all cases other than the most trivial ones, where only a few disulfide bonds need to be considered. This is because the number of possible disulfide bonding models grows rapidly with the number of cysteines and the number of expected disulfide bonds. Specifically, given N cysteines and b disulfide bonds, the number of disulfide bonding models m is given by the recursion [22]:

$$m(b, N + 1) = m(b, N) + N \times m(b - 1, N - 1)$$  \hspace{1cm} (15)

Where $$m(0, N) = 1$$ for $$N \geq 0$$ and $$m(b, 0) = 0$$ for $$b > 0$$

Consequently, the total number of models for a given number of cysteine residues N, M(N) is:

$$M(N) = \sum_{b=0}^{N} m(b, N)$$  \hspace{1cm} (16)

In one of the early works on disulfide-bond analysis using mass spectrometry, Morris and Pucci [60] proposed a technique similar to paper-electrophoresis-based diagonal-peptide mapping with the exception that separation and analysis occurred simultaneously in the mass spectrometer. Disulfide-bonded peptides were identified by comparing spectra obtained with aliquots of the digests before and after reduction. Combinations of masses were computationally analysed and linkages were assigned based on them. These assignments were confirmed through Edman degradation. Intra-chain bonds were identified based on the mass increases associated with the reduction of the constituent half cysteines. Other early works include [61–64]. Further description of non-algorithmic techniques for disulfide-bond determination can be found in [59].

An early work in algorithmic determination of disulfide bonds using data from mass spectrometry is [65]. In [66], a method was proposed where all the linear peptides containing 1–3 cysteines were generated. Next, all combinations containing disulfide bonds between 2 and 6 cysteines were obtained and their theoretical mass determined. Mass spectrometric data were matched with the generated set of disulfide-linked peptides and the output screened using specificity of the proteolytic enzyme and/or further mass data from Edman degradation or carboxypeptidase treatment of the mixture. The approach in [66] was designed to work on a single manually entered sequence at a time. It was extended in a later work [67] to work with entire databases extracted from the PIR non-redundant reference protein database (ftp://ftp.pir.georgetown.edu/pir_databases/nref). The approach in [68], which is available as part of the PROWL software [69] was to list, based on the specificity of the enzyme used, all the possible linear reduced proteolytic digestion products containing cysteine residues. Subsequently, all possible disulfide linked peptides were identified and their mass was calculated. The calculated masses were compared with the experimentally derived masses to identify the sequences of the disulfide-linked peptides. Finally, the algorithm checked the disulfide-linked peptides to test if they contained component peptide chains having masses that corresponded to the measured masses of the reduced peptides. The implementation allowed for specifying the degree of completeness of digestion thus allowing modelling of scenarios with complete or partial digestion. An implementation and application of this algorithm was also used in [22]. Unlike the enumerative strategy of [67, 68], the computational complexity issues involved in generating and matching the set of all possible disulfide bonded peptide fragments with the mass spectrum was studied in [70, 71]. The authors exploited the structure of the problem and mass constraints from the spectrum to propose a hashing-based algorithm that reduces the complexity of making an identification from $$O(n^2)$$ to approximately $$O(n)$$, where n denotes the size of the mass space consisting of every disulfide-bonded fragment, given the peptides. In [72], this algorithm was used to derive a measure of how well the experimental data supported the hypothesis of a disulfide bond between two cysteines in the protein being analysed. This value was then used to weight an edge between the corresponding cysteines in a graph that modelled the overall disulfide connectivity. The maximum weight perfect matching problem was subsequently solved to obtain the final disulfide connectivity.

Other techniques that can also be used in this context include ProteinProspector (http://prospect- tor.ucsf.edu), X! Protein Disulphide Linkage Modeler (http://www.systemsbiology.ca/x-bang/ DisulphideModeler/disulphide.html), Peptidemap (http://prowl.rockefeller.edu/prowl/peptidemap.html), and MS2Assign. Of these only MS2Assign can be used to analyse disulfide linkages from MS/MS...
data. However, due to the fact that MS2Assign was primarily designed for analysis of results from cross-linking studies, it requires detailed information on the specific modifications expected. It is also not designed to be directly used in high-throughput settings.

Disulfide-bond determination can also be considered as a special case of determining cross linkages from mass spectra. In [73], a three step procedure is presented to algorithmically identify cross-linked peptides from tandem mass spectrometry. Given the mass of the parent cross-linked molecules obtained from mass spectrometry, the first step finds all peptide pairs whose mass (along with the linker mass) equals the parent mass. In the second step, given a pair of peptides, the cross-linked amino acid pair optimally correlated to the mass spectrum was found. Finally, the pair of peptides with the highest correlation score was reported. Given \( k \) protein sequences, each with \( n \) proteolytic amino acids that are targets for proteolytic digestion and a tandem mass spectrum having \( h \) peaks, the algorithm proposed in [73] solved the first step in \( O(kn^2 \log(kn)) \) time and the second step in \( O(m + h) \) time. Another approach for cross-linkage determination, used in MS2Assign (http://roswell.ca.sandia.gov/~mmyoung/ms2assign.html) is discussed in [74] which uses the ASAP (Automated Spectrum Assignment Program) [75] to suggest possible structures for cross-linked and non-cross-linked peptides resulting from the digestion of cross-linked proteins.

**CONCLUSIONS**

It is clear that the field of disulfide-bond determination using algorithmic approaches continues to be in a very dynamic state. Two classes of techniques can be seen to have evolved. Predictive approaches using machine-learning strategies and (typically) sequence-based information can be seen to have steadily improved in their accuracy. Yet, in using them, researchers should be aware of their dependency on the machine-learning algorithm, the features used, and the training data set. In contrast, techniques based on mass spectrometric analysis are direct. With increasing development of mass spectrometric technologies, these techniques can be expected to mature further and play an increasingly important role in solving the problem of determining disulfide connectivity.

---

**Key Points**

- A review of computational techniques for disulfide-bond determination is provided.
- Two major classes of techniques are noted to have evolved in this area: predictive techniques and computational methods based on analysis of mass spectrometric data.
- Published results indicate both classes of methods to be efficacious and accurate.
- Generalizability of the predictive methods depends on the features used in the learning formulation, the learning algorithm, and the training data. Mass spectrometry-based methods, in contrast, directly work with the data pertaining to the molecule being analyzed.

---

**Funding**

National Science Foundation IIS-0644418 (CAREER).

---

**References**


