Sequence analysis

Predicting disulfide bond connectivity in proteins by correlated mutations analysis

Rotem Rubinstein and Andras Fiser*

Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Received on August 14, 2007; revised on December 24, 2007; accepted on December 25, 2007

Advance Access publication January 18, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Prediction of disulfide bond connectivity facilitates structural and functional annotation of proteins. Previous studies suggest that cysteines of a disulfide bond mutate in a correlated manner.

Results: We developed a method that analyzes correlated mutation patterns in multiple sequence alignments in order to predict disulfide bond connectivity. Proteins with known experimental structures and varying numbers of disulfide bonds, and that spanned various evolutionary distances, were aligned. We observed frequent variation of disulfide bond connectivity within members of the same protein families, and it was also observed that in 99% of the cases, cysteine pairs forming non-conserved disulfide bonds mutated in concert. Our data support the notion that substitution of a cysteine in a disulfide bond prompts the substitution of its cysteine partner and that oxidized cysteines appear in pairs. The method we developed predicts disulfide bond connectivity patterns with accuracies of 73, 69 and 61% for proteins with two, three and four disulfide bonds, respectively.

Contact: rmubinst@aecom.yu.edu, andras@fiserlab.org

1 INTRODUCTION

The disulfide bond is the most frequent naturally occurring covalent cross-link in proteins. It is derived from the oxidation of the thiol groups of two cysteine residues. Proteins with disulfide bonds are usually secreted and rarely found in the cytoplasm, which has a reducing environment and lacks enzymes that promote disulfide bond formation (Kadokura et al., 2003). However certain archaea are rich in cytoplasmic proteins with disulfide bonds (Mallick et al., 2002). A number of studies have linked disulfide bonds to protein stability and to folding rate (Wedemeyer et al., 2000). It has been suggested that disulfide bonds stabilize the protein’s folded state by restricting the protein’s conformation, thereby reducing the entropy of the unfolded state (Harrison and Sternberg, 1994; Poland and Scheraga, 1965). Meanwhile, disulfide bonds increase the enthalpy of the folded state by stabilizing local interactions (Wedemeyer et al., 2000). Furthermore, disulfide bonds increase the protein’s half-life by enhancing protein protection against proteases by maintaining the integrity of protein structure against local unfolding events. Disulfide bonds have also been observed to contribute to protein function regulation (Hogg, 2003).

Disulfide bonds constrain the conformation of the protein structure and thus, knowledge of their location can facilitate protein structure prediction. In addition, disulfide bond connectivity patterns can be used to discriminate between protein folds and to accurately superimpose protein structures (Chuang et al., 2003; Gupta et al., 2004; Mas et al., 1998). The underlying assumption in these methods is that similar disulfide bond connectivity patterns place similar spatial constraints on proteins, resulting in similar protein structures. Finally, variation in disulfide bridge patterns may be used to infer variation of protein function (Cao et al., 2007).

There are two distinct steps in the process of predicting disulfide bond connectivity patterns. The first is the classification of bound (oxidized) and unbound (reduced) cysteines. The second is the correct pairing of all bound cysteines. Muskal and colleagues (1990) published the first method to identify bound and free cysteines by utilizing a neural network and reported a prediction accuracy of 82%. Fiser and colleagues (1992) observed that the sequence environments of bound and free cysteines have different compositions and they subsequently introduced a method to calculate disulfide-bond forming potential that is based on the amino acid composition of the sequential environment of cysteines. Later, Fiser and Simon analyzed the apparent difference between the conservation level of oxidized and reduced cysteines and developed a method to predict the oxidation states of cysteines from the conservation analysis of multiple sequence alignments. This simple approach reached a prediction accuracy of 82% (Fiser and Simon, 2000). In the same study it was noted that it is rare for a protein to have cysteines with mixed oxidation states. Mucchielli-Giorgi et al. (2002) developed a cysteine oxidation state predictor based on Fiser and Simon’s finding (Fiser and Simon, 2000) and on the global amino acid composition of proteins and attained an 84% prediction accuracy. Chen et al. (2004) trained a support vector machine on the local environment of cysteines as well as on global information of the protein and reported a 90% prediction accuracy. Given these high prediction accuracies for cysteines oxidation state prediction our current study focuses on the second step in prediction of disulfide bond connectivity patterns.
connectivity; the challenging problem of identifying the correct pairing of bound cysteines.

Given 2n cysteines that form n disulfide bonds, the number of possible connectivity patterns of all 2n cysteines is (2n − 1)!!.

The number of possible disulfide bond connectivity patterns (cysteine pairings) increases rapidly with the number of bound cysteines (e.g., for proteins with four, six and eight disulfide bonds there are 105, 10,395 and ~2 × 10^6 possible connectivity patterns, respectively). An exhaustive search for optimal pairing of cysteines is possible only when the number of bound cysteines is small. To overcome the combinatorial explosion problem, the problem of pairing bound cysteines was translated into the problem of finding the perfect match in a complete weighted and undirected graph (Fariselli and Casadio, 2001), which can be solved in polynomial time using the Edmund–Gabow algorithm (Gabow, 1976). In their approach, graph vertices, edges and the weights of edges represent bound cysteines, potential connectivity between two cysteines and confidence scores for the pairing of two cysteines, respectively (Fariselli and Casadio, 2001).

Most current methods that predict disulfide bond connectivity use graph representation. These methods typically differ in the way in which the weights of the edges are calculated. Fariselli and Casadio (2001) assigned contact potentials to edge weights based on the assumption that the nearest sequential neighbors of the paired cysteines were also in contact. Their calculation was limited to protein queries with up to five disulfide bridges as the process of calculating the contact potential employed time consuming Monte Carlo and simulated annealing procedures. In a more recent work, the authors increased the speed of the contact potential calculation by employing a neural network (Fariselli et al., 2002). Vullo and Frasconi (2004) were able to significantly increase the accuracy of prediction by incorporating evolutionary information. The authors utilized a recursive neural network to score disulfide connectivity patterns. Ferre and Clote (2005) utilized a neural network with a unique hidden layer intended to examine bi-residue information. They incorporated evolutionary information in the form of Position Specific Scoring Matrices (PSSM) and also added secondary structure information. Tsai et al. (2005) used a Support Vector Machine (SVM) with evolutionary information and protein sequence separation of cysteines pairs as inputs. Cheng et al. (2006) created a complete platform for disulfide bridge prediction by predicting both the bound state of the cysteines and the disulfide bond connectivity pattern. The authors utilized kernel methods to predict the bound state of cysteines and a recursive neural network to predict disulfide bond pairing. The input to the neural network included evolutionary information, sequence separation of cysteine pairs, and solvent accessibility. Zhao et al. (2005) approached the problem of identifying the correct pairing of bound cysteines from a global perspective. Rather than scoring each possible pair of cysteines, the authors compared the query cysteine sequence separation profile to a database of similar profiles of proteins with known disulfide bonds. The limitation of the approach is that a novel cysteine pattern cannot be found. Chen and Hwang (2005) incorporated both local evolutionary information, in the form of PSSM, and global information, in the form of cysteine separation profile, as inputs for a SVM.

Lu et al. (2007) used a genetic algorithm to improve the optimal selection of input sources for disulfide bond prediction. Chen et al. (2006) utilized global and local information as inputs for a two layer SVM. The first layer had a SVM that utilized local information as the input. The inputs for the second layer were scores from the first layer of SVM along with global information such as the protein length, the cysteine separation profile and the disulfide connectivity frequency.

Fiser and Simon (2000) showed that it is possible to discriminate between the different oxidation states of cysteines based solely on the conservation analysis of the cysteines. However, this analysis cannot predict the correct pairing of oxidized cysteines because all the oxidized cysteines are expected to have a similar level of conservation (Fiser and Simon, 2002). Two cysteine residues in a disulfide bond form a strong interaction, which, in many cases, maintains both the protein’s structure and function. Such a strong interaction is expected to lead to interdependency between the two positions, which could be traced through evolution. In addition, it is difficult to maintain different redox conditions for the same protein environment, i.e., if a bridge forming an oxidized cysteine is mutated, its reduced cysteine partner will be under pressure to mutate as well. Correlated mutation algorithms aim to identify residue–residue linkages through identifying patterns of concerted variations in different positions in a multiple sequence alignment. A variety of correlated mutation algorithms have already been utilized for predicting residue contacts in a protein 3D structure (Dekker et al., 2004; Gobel et al., 1994; Hamilton et al., 2004; Larson et al., 2000; Neher, 1994; Shindyakov et al., 1994), but there has not yet been an attempt to utilize correlated mutation algorithms to automatically identify disulfide bond connectivity.

Thornton (1981) analyzed 15 cases of non-conserved disulfide bonds and observed that when a disulfide bond is not conserved both cysteines are mutated in concert. Kreisberg et al. (1995) examined the trypsin-like serine proteases and phospholipase A2 protein families and demonstrated that cysteines forming disulfide bonds mutate in a correlated pattern. The authors noted that this correlated pattern could be used to predict disulfide bonds in proteins. These two studies observed trends in cysteine mutations with respect to conservation of disulfide bonds, although they utilized very small databases.

This study uses a large set of protein families to analyze non-conserved disulfide bonds. Subsequently, we introduce a novel method that predicts disulfide bond connectivity pattern using a correlated mutation algorithm.

2 METHODS

2.1 Conservation analysis of disulfide bonds

2.1.1 Data set In order to assess the conservation pattern of disulfide bonds, we examined protein domains from (SCOP) structural classification of proteins (Murzin et al., 1995) that had at least one disulfide bond. Our operational definition for a disulfide bond occurrence is when Sulfur gamma (SG) atoms of two cysteine residues fall within a 2.5 Å distance of each other. The expected SG–SG distance for disulfide bond is ~2 Å but this more generous definition accounts for inaccuracies in experimental data. We removed redundancy of
2.2 Disulfide bond prediction

2.2.1 Data set In order to benchmark the predictive power of our method we used the same version of annotated protein sequences of Swiss-Prot as other studies: release 39 (1999) (Boeckmann et al., 2003). Protein sequences were filtered by two requirements as described previously (Fariselli and Casadio, 2001). First, only proteins with known 3D structures were considered. Second, disulfide bonds annotation could not contain the words ‘by similarity’, ‘probable’ or ‘potential’. The test set had 435 proteins that were grouped by the number of disulfide bonds.

2.2.2 Generating multiple sequence alignment For each query protein, evolutionary related sequences were extracted from NR (Wheeler et al., 2007) by running five rounds of PSI-BLAST (Altschul et al., 1997). A representative multiple sequence alignment was generated by filtering the sequences from the PSI-BLAST output using BlastProfiler (Rai et al., 2006) with the following parameters: minimum e-value lower than 0.0001, hit-query alignment sequence identity of at least 15%, hit-query alignment coverage of at least 30% and 90% maximum sequence identity between any two hits.

2.2.3 Scoring scheme Our scoring scheme is based on the correlated substitution pattern observed for positions participating in disulfide bonds. In most cases of non-conserved disulfides both cysteines are substituted (see Section 3). Since it is unlikely that all disulfide bonds will always mutate simultaneously, we search for a simple correlation pattern of concerted appearing and disappearing of cysteines in order to predict disulfide bonds. Given a multiple sequence alignment, we examined only those sequence positions (columns) that correspond to disulfide forming cysteine positions in the query. For each sequence in the alignment, we divided the examined positions into two sets based on their amino acids composition; the first set is composed of positions with cysteine residues, while the second set is composed of positions with a gap or any residue other than a cysteine. The score for each possible pair in a sequence is a number between zero and one, and corresponds to our expectation that this pair of positions form a disulfide bond in the query protein based solely on the current sequence examined. If two positions are part of different sets (only one position is a cysteine) then the correlation score is zero, because our observation demonstrated that it is unlikely that only one of the positions that formed a disulfide bond in the query is substituted. If two positions are part of the same set (either a set of all cysteines or a set of anything but cysteines) then the score for a pair of such positions is 1/ (size of set −1), which is the probability of selecting the correct position pair randomly and with equal chance, assuming that pairing is possible between only those positions that are part of the same set. Those sequences were ignored in the alignment that contained either completely conserved or completely varied all cysteines as no correlated mutation information can be extracted from these. Also, sequences with an odd number of cysteines, at the sequence positions examined, were removed as they were assumed to be a product of a misalignment.

2.2.4 Retrospective prediction of disulfide connectivity for year 1999 Multiple sequence alignments were generated as described earlier. Then sequences were removed if, in the NCBI protein flat-file, the description of the year of creation was later than 1999.

3 RESULTS

3.1 Disulfide conservation analysis

3.1.1 Cysteines of disulfide bonds mutate in concert In order to analyze the conservation of disulfide bonds, we constructed multiple structural alignments of proteins of SCOP families that contained at least one disulfide bond. From 189 families 1363 such proteins were analyzed. We examined whether cysteines of non-conserved disulfides mutate in concert by comparing the number of times disulfides were substituted by two non-cysteines residues (or gaps) to the number of times only one of the cysteines was substituted. Disulfide bonds were observed to vary in 4288 cases and in 3463 (81%) of these cases both cysteines mutated in concert. However, after manual investigation it turned out that, in most of the cases, the non-concerted disulfide substitution was a consequence of either a misalignment or a protein structural divergence that the alignment program could not account for. Upon removing such disulfides from our conservation analysis 99% of the cases showed disulfides that mutated in concert.

3.1.2 Disulfide bond forming cysteines are not always conserved Proteins of the same SCOP family have obvious evolutionary relationships, usually sharing >35% sequence identity. Nevertheless, we observed that, even within the same SCOP family, disulfide bonds are not always conserved. Of the families, 60% had at least one non-conserved disulfide bond and, overall, we observed that 66% of all disulfide bonds are not conserved. These results are in agreement with recent findings, which show that variations in the number of disulfide bonds in proteins of the same structural family are not unusual (Cheek et al., 2006).
3.2 Prediction of disulfide bond connectivity pattern

3.2.1 Performance of the prediction method  Our method can predict disulfide bonds in proteins with any number of bonds. However, results are reported only for proteins with 2–4 disulfide bonds because the prediction of proteins with one bond is trivial and sequence databases lack a sufficient number of protein sequences with five or more disulfide bonds for a statistically significant analysis.

Tables 1 and 2 summarize our results for predicting disulfide bond connectivity patterns and disulfide bonds, respectively. When reporting on disulfide bonds prediction we measure our ability to correctly predict any disulfide bond in a protein. Tables 3 and 4 list results of predictions obtained in previous studies as well as results from two trivial predictors: a random predictor and a frequency predictor. A random predictor predicts disulfide connectivity by random, while a frequency predictor predicts bridges by relying on the most common connectivity pattern observed in the database. Comparing our results to the random and frequency predictors demonstrates that correlated mutations can capture the evolutionary signal generated by the disulfide bond interactions. The strength of using a correlated mutation analysis is most apparent when predicting connectivity patterns for proteins with four disulfide bonds (105 possible ways to combine four bonds). The method presented here is capable of predicting all four disulfide bonds with 61% accuracy and can predict a subset of bonds out of the four bridges with an accuracy of 64%. Because, we predicted a subset of the data set utilized by other studies any direct comparison is limited. Nevertheless, in order to obtain a general insight, we evaluated our results along with the results obtained from other methods. With the exception of one method (Lu et al., 2007), our approach produces predictions with the highest accuracies for proteins with three and four disulfide bonds both in predicting disulfide bond connectivity and in predicting disulfide bonds.

3.2.2 Predicting disulfide bonds of proteins with a mixed state of cysteines  We also analyzed our prediction method using protein sequences with disulfide bonds but with an odd number of cysteines. In addition to providing information on the predicted disulfide pattern, we also identified the unbound cysteine. Results are taken from previous works and are reported for proteins with two, three and four disulfide bonds.

References for methods in Tables 3 and 4: (a) (Vullo and Frasconi, 2004), (b) (Fariselli and Casadio, 2001), (c) (Cheng et al., 2006), (d) (Ferre and Clote, 2005), (e) (Zhao et al., 2005), (f) (Tsai et al., 2005), (g) (Chen and Hwang, 2005), (h) (Chen et al., 2006) and (i) (Lu et al., 2007).

Table 1. Summary of the accuracy [(TP/(TP+FP)] and coverage (predicted queries/all queries) of disulfide connectivity predictions for proteins with two, three and four disulfides

<table>
<thead>
<tr>
<th>Number of disulfides (number of query proteins)</th>
<th>Current 2007 NR</th>
<th>1999 NR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Coverage</td>
</tr>
<tr>
<td>2 (148)</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>3 (146)</td>
<td>0.69</td>
<td>0.33</td>
</tr>
<tr>
<td>4 (98)</td>
<td>0.61</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 2. Accuracy and coverage of disulfide-bond predictions for proteins with two, three, four and five disulfide-bonds

<table>
<thead>
<tr>
<th>Number of disulfides (number of query bridges)</th>
<th>Current 2007 NR</th>
<th>1999 NR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Coverage</td>
</tr>
<tr>
<td>2 (296)</td>
<td>0.73</td>
<td>0.68</td>
</tr>
<tr>
<td>3 (438)</td>
<td>0.71</td>
<td>0.43</td>
</tr>
<tr>
<td>4 (392)</td>
<td>0.64</td>
<td>0.34</td>
</tr>
<tr>
<td>5 (215)</td>
<td>0.59</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 3. Accuracy of predicting disulfide bond connectivity by other methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of disulfide bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Random</td>
<td>0.33</td>
</tr>
<tr>
<td>Frequency^a</td>
<td>0.58</td>
</tr>
<tr>
<td>FC^b</td>
<td>0.56</td>
</tr>
<tr>
<td>VE^c</td>
<td>0.73</td>
</tr>
<tr>
<td>CSB^d</td>
<td>0.74</td>
</tr>
<tr>
<td>FCL^d</td>
<td>0.62</td>
</tr>
<tr>
<td>CSP^e</td>
<td>0.74</td>
</tr>
<tr>
<td>TCCLK^f</td>
<td>0.79</td>
</tr>
<tr>
<td>CH^g</td>
<td>0.74</td>
</tr>
<tr>
<td>CTCK^h</td>
<td>0.85</td>
</tr>
<tr>
<td>GASVM^i</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 4. Accuracy of disulfide bond predictions by other methods

<table>
<thead>
<tr>
<th>Methods</th>
<th>Number of disulfide bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Random</td>
<td>0.33</td>
</tr>
<tr>
<td>Frequency^a</td>
<td>0.58</td>
</tr>
<tr>
<td>FC^b</td>
<td>0.56</td>
</tr>
<tr>
<td>VE^c</td>
<td>0.73</td>
</tr>
<tr>
<td>CSB^d</td>
<td>0.74</td>
</tr>
<tr>
<td>CSP^e</td>
<td>0.74</td>
</tr>
<tr>
<td>TCCKH^f</td>
<td>0.79</td>
</tr>
<tr>
<td>CH^g</td>
<td>0.74</td>
</tr>
<tr>
<td>GASVM^i</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Results are cited from previous works and reported for proteins with two, three and five disulfide bonds.
The sequences we analyzed rarely had cysteines with mixed oxidation states in agreement with observations of earlier studies (Fiser and Simon, 2000). Out of the 137 cases, we studied only 12 could be confirmed to have both oxidized and reduced cysteine, for 38 it was not possible to identify the origin of the extra cysteine (not even after consulting the original literature) and 87 cases came from separating cysteines into intra and inter domain disulfide bonds. In these cases the cysteine of the interdomain disulfide bond shows as an unbound one because the crystal structure presents the monomeric state only. This latter set presents a more difficult task to the prediction algorithm, as the conservation levels of intra and inter disulfide bond forming cysteines are rather similar and these differ in their correlation pattern only. Our results for proteins with three, five and seven cysteines (i.e. proteins with one, two and three disulfide bonds and one free cysteine, respectively) demonstrate 91, 55 and 24% prediction accuracy, respectively. In terms of possible number of connectivity combinations, these numbers can be compared to the prediction accuracies of 73, 69 and 61% of the connectivity patterns for four, six and eight cysteines, respectively, when all cysteines are known to be in disulfide bonds. This suggests that with an increasing number of combinations the accuracy of prediction is getting worse, possibly due to an extra task of identifying free/interdomain cysteines. However, the accuracies remain significant and comparable to our earlier results.

3.2.3 Applicability of the method  
Prediction of the disulfide bond’s connectivity pattern using the correlated mutation algorithm presented here requires that all but one disulfide bond is not fully conserved. Our method cannot predict disulfide connectivity patterns of proteins that do not follow this requirement. We evaluated the applicability of our algorithm by measuring the coverage (the number of predicted proteins divided by the number of proteins tested, or the number of disulfide bond predictions divided by the number of disulfide bonds tested) (Tables 1 and 2). The number of predicted proteins with 2–4 disulfide bonds was high enough for a statistically significance analysis. However, there were only nine predicted proteins with five disulfide bonds, which limited the reliability of statistical analysis, and therefore, we did not analyze the accuracy of disulfide bond connectivity predictions for proteins with five or more disulfide bonds. When we tested the performance of the current algorithm using a sequence database from 1999, we found a considerable decrease in coverage, sometimes by half (Tables 1 and 2). This implies that our algorithm applicability will further increase in the future, as sequence databases expand.

3.2.4 Illustration of the prediction method  
The first example suggests that, in order to accurately predict all disulfide bonds in a query protein, our algorithm may require only a very few evolutionarily related sequences to the query as long as they are sufficiently diverse in their disulfide bond patterns. Pepsin-A precursor (pepa_human) is a human protein that belongs to the peptidase A1 family. This protein has three disulfide bonds located at the query sequence positions of 107–112, 268–272 and 311–344. We automatically generated a multiple sequence alignment with 152 sequences but only two of the 152 sequences could be used for correlated mutation analysis because, in the rest of the cases, all three disulfide bonds were completely conserved. However, the alignment with the two remaining protein sequences (AAA23476, and XP_61523) was sufficient to decipher the connectivity of all three disulfide bonds in the query protein. Each protein sequence in the alignment had a different non-conserved disulfide bond, which is the minimum required information to predict disulfide patterns properly in a protein with three disulfide bonds (Fig. 1).

A second example highlights the correlated mutation pattern observed in disulfide bond positions. β-lactamase (hcpB) from Helicobacter pylori has four disulfide bonds formed between cysteines at sequence positions of 22–30, 52–60, 88–96 and 124–132. Figure 2 illustrates the correlated mutation pattern observed in a multiple sequence alignment of hcpB proteins. The corresponding correlated mutation matrix scores generated by our algorithm highlight the simultaneous mutations of both cysteines in non-conserved disulfide bonds.

A third example illustrates the fact that a correlated mutation signal can drastically reduce the problem of disulfide connectivity prediction even if the prediction is partially ambiguous. Proproteinase E precursor (Cac3_bovine) is a bovine protein with 10 cysteines involved in five disulfide bonds located at query sequence positions 41–57, 100–103, 140–206, 171–187 and 196–227. A multiple sequence alignment reveals that two of the five disulfide-bonds are completely conserved (171–187 and 196–227). However, three disulfide bonds can be accurately predicted based on their correlated pattern of conservation. Although our algorithm cannot fully predict all the disulfide bonds of Cac3_bovine, it produces valuable information as it reduces the complexity of prediction from 945 possible combinations of 10 cysteines to three possible combinations of four cysteines. The multiple sequence alignment of Cac3_bovine is composed of 171 sequences of which nine are completely conserved and 16 have an odd number of cysteines at the examined positions (these were ignored as assumed to be product of misalignment). Out of the remaining 146 sequences 131 had one unconserved disulfide bond, which was always aligned with the query disulfide bond at sequence positions 100–103. Fifteen sequences had two unconserved disulfide bonds, of which 13 had neither the disulfides corresponding to the query sequence positions 100–103, nor to the sequence positions 140–206. Two sequences did not have
4 DISCUSSION

In the current study, we developed a correlated mutation algorithm to identify disulfide bonds in proteins using sequence information alone. We assumed that correlated mutation analysis is a suitable technique to predict disulfide bonds, because disulfide bond is a well defined residue–residue interaction that plays an important role for the protein structure and function. When such a strong relationship exists between two sequence positions it is expected to result in the coevolution of these positions.

Two requirements have to be fulfilled in order to predict disulfide bonds with correlated mutation analysis. First, some disulfide bonds must be unconserved and second, cysteines of unconserved disulfide bonds have to substitute in a correlation manner. Our analysis of multiple structure alignments of proteins from the same SCOP family demonstrated that both conditions are met. In agreement with recent findings (Cheek et al., 2006), we observed that the number of disulfide bonds varied between evolutionary related proteins. We also demonstrated that multiple sequence alignment columns corresponding to the query disulfide bonds showed a correlated pattern of conservation, i.e. the simultaneous appearance and disappearance of cysteines.

In the current analysis we assumed that the cysteines participating in a disulfide bond are known and, therefore we focused on identifying the correct pairing of these residues. The reason behind this feasible assumption is twofold: (i) the bound state of cysteines can be predicted by several methods with a high accuracy of around 90% (see Section 1), and (ii) <5% of proteins contain cysteines with mixed oxidation states (Fiser and Simon, 2000). Furthermore, past findings have shown that unbound cysteines are significantly less conserved than bound one (Fiser and Simon, 2000), and consequently these will have little or no correlation with bound cysteine. Therefore, any error in predicting the bound state of cysteines should affect only a small fraction of proteins. When we tested our algorithm on a set of proteins with cysteines in mixed oxidation states the predictive power was sustained.

A limitation of our algorithm is that if more than one fully conserved disulfide bond exists, we cannot predict all disulfide bonds of a protein unambiguously. We demonstrated that the recent expansion of sequence databases made our algorithm applicable to more proteins by an average factor of 1.5 since 1999, which suggests an increasing and wide applicability of this approach in the future. We also examined whether our approach provides a unique aspect of disulfide connectivity prediction, in comparison to other methods. Therefore, we compared the overlap between true positive predictions on the same test set using both our approach and a method developed by Cheng and colleagues (2006), which is one of the best method that is publicly available. The protein test set was composed of proteins sequences from a recent release of Swiss-Prot SP51 (2007) (Boeckmann et al., 2003), filtered as described earlier. We retained only those proteins that shared <30% sequence identity with any other proteins from the training set that was used to train the neural network method developed by Cheng et al. (2006). The resulting test set was composed of 275 proteins with 2–4 disulfide bonds and was new to both methods. Our findings showed that our algorithm predicted 135 proteins, of which 83 were correct and 52 were incorrect predictions (61% accuracy). The method of Cheng et al.
correctly predicted 79 out of the 135 proteins (58% accuracy). When we compared the true positive predictions of both methods there were 57 overlapping cases. This indicates that an ideal combination of both methods could provide a maximum accuracy of 78%, suggesting a potential 17–20% increase over the current accuracies of these methods if used in combination.

Meanwhile it is very useful if one is able to accurately predict a subset of disulfide bonds that vary as this information can be used as a source input for meta predictors or as a complement for other indirect experimental studies that introduce cross-links. Disulfide bond prediction for proteins where only a subset of disulfides are unconserved has an important implication as it may suggest a structural/functional feature not shared by all members of a protein family. For instance, the T cell immunoglobulin mucin (TIM) protein family provides a recent and interesting example as these proteins were found to have two unique disulfide bonds on top of the canonical disulfide of the immunoglobulin domain. The two non-canonical disulfide bonds support the scaffold of a unique binding site in TIM proteins (Cao et al., 2007).

While past findings indicated that using multiple sequence alignment significantly increases the accuracies of disulfide bond prediction, it remained unclear as to how multiple sequence alignments support the prediction. In the current study we illustrated that part of the contribution of multiple sequence alignment is the identification of the correlated mutation patterns of the query-bounded cysteines. Future studies should evaluate the correlated mutation pattern of the sequence environment of bound cysteines to disulfide bond prediction.

ACKNOWLEDGEMENTS

We thank Joseph Dybas, Narcis Fernandez-Fuentes, Eduardo J. Fajardo, Dmitry Rykunov and Daniela Yaar for helpful discussions. Financials support was provided by NIH-NIAID HHSN266200400054C.

Conflict of Interest: none declared.

REFERENCES


