Prediction of disulfide connectivity in proteins

Piero Fariselli and Rita Casadio

CIRB Biocomputing Unit, Laboratory of Biophysics, Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy

Received on April 20, 2001; revised and accepted on July 12, 2001

ABSTRACT

Motivation: A major problem in protein structure prediction is the correct location of disulfide bridges in cysteine-rich proteins. In protein-folding prediction, the location of disulfide bridges can strongly reduce the search in the conformational space. Therefore the correct prediction of the disulfide connectivity starting from the protein residue sequence may also help in predicting its 3D structure.

Results: In this paper we equate the problem of predicting the disulfide connectivity in proteins to a problem of finding the graph matching with the maximum weight. The graph vertices are the residues of cysteine-forming disulfide bridges, and the weight edges are contact potentials. In order to solve this problem we develop and test different residue contact potentials. The best performing one, based on the Edmonds–Gabow algorithm and Monte-Carlo simulated annealing reaches an accuracy significantly higher than that obtained with a general mean force contact potential. Significantly, in the case of proteins with four disulfide bonds in the structure, the accuracy is 17 times higher than that of a random predictor. The method presented here can be used to locate putative disulfide bridges in protein-folding.

Availability: The program is available upon request from the authors.

Contact: Casadio@alma.unibo.it; Piero@biocomp.unibo.it

INTRODUCTION

Among the 20 amino acids, cysteine residues in proteins have a unique property since they may pair to form disulfide bridges that contribute to the thermodynamic stability of the 3D structure. This property was first highlighted by the pioneering work of Anfinsen on the pancreatic bovine ribonuclease. According to Anfinsen’s results protein unfolding in vitro occurs upon disulfide full reduction, whereas upon re-oxidation, correct disulfide connectivity (folding) and native structure is restored (Anfinsen, 1973). More recently kinetic characterization of the folding process indicated that correct disulfide connectivity parallels the correct formation of protein secondary structure (Creighton, 1993). This strengthens the view that disulfide bonds increase the conformational stability of the protein mainly by constraining the unfolded conformation, as many experimental and theoretical studies suggest (Harrison and Sternberg, 1994 and references therein; for review see Wedemeyer et al., 2000).

In protein-folding prediction, the location of disulfide bridges can strongly reduce the search in the conformational space (Skolnick et al., 1997; Huang et al., 1999). Therefore the correct prediction of the disulfide connectivity starting from the protein residue sequence may also help in predicting its 3D structure.

A necessary step to the prediction of disulfide connectivity is the prediction of the disulfide bonding state of cysteine in proteins. This has been tackled before (Muskal et al., 1990; Fiser et al., 1992) and recently refined (Fariselli et al., 1999; Fiser and Simon, 2000). The methods presently available discriminate between free and bonded state of cysteine with a high accuracy (about 80%) starting from the residue chain. This is possible, since the local sequence environments of the two type of cysteine (free and bonded) are quite different as highlighted by neural network-based predictors (Muskal et al., 1990; Fariselli et al., 1999). Analyzing the values of the neuron junctions of the predictor, it was possible to better understand which type of sequence environment was conducive to disulfide bond formation (Fariselli et al., 1999). More specifically when a segment centred in the cysteine to be predicted and including 11 residues is considered, the following features could be captured: (a) the presence of cysteine residues in the environment of the central cysteine strongly favours the disulfide bond formation, with the exception of position 3. This is in agreement with the fact that in proteins metal binding cysteines are typically found in position i and i + 3; (b) hydrophilic and/or charged residues in the environment are highly conducive towards disulfide bond formation as compared to hydrophobic residues which are poorly conducive.

The question then poses as to also whether cysteine pairing in the bridge is endowed with characteristic marks that can be captured and used to univocally establish the disulfide connectivity in proteins. In this work we will
complement the prediction of disulfide bonding state by focusing on the prediction of the disulfide connectivity.

SYSTEM AND METHOD
The protein data set
Protein sequences were extracted from the SWISS-PROT database (Bairoch and Apweiler, 2000). The extracted data set includes only the sequences containing information derived from the Protein Data Bank (PDB) and for which intra-chain disulfide bonds are annotated. In order to limit the noise due to wrong annotations, disulfide assignments that are described as ‘probable’, ‘potential’ or ‘by similarity’ are not retained. After this filtering procedure we ended up with 726 proteins out of the SWISS-PROT release no. 39, October 2000. A bar graph of the proteins grouped according to the number of disulfide bonds per protein, shows that most of the proteins (82%) are located in the 1–4 disulfide bond region (i.e. 2–8 cysteines in the protein forming disulfide bonds; Figure 1). Disulfide bonds in the selected database predominantly occur in coil structures (55%). Chains were grouped in four different sets with the constraint that sequence homology among the different sets was ≤30%. By this the sets included an approximate equal number of proteins. This grouping procedure allowed us to perform a 4-fold cross validation of the method.

Problem definition and the measure of accuracy
Given an even number of cysteines \(2B\), with \(B \in N\) believed to form disulfide bonds, the problem is to determine the correct connectivity pattern among all the possible alternatives. This problem becomes harder as the number of disulfide bridges (and the number of bonding forming cysteines) increases. For instance, the problem is trivial in the case of one bridge \((B = 1)\), while it may become unfeasible when the number of bridges increases. For this reason, only scoring the predictions as a function of the number of bridges \((B)\) can meaningfully assess the accuracy of the predictive method.

For the sake of clarification, the case of six cysteines in a protein chain forming three disulfide bridges \((B = 3)\) is depicted in Figure 2. On the top left side the correct connectivity pattern of the disulfide bridges is shown. On the right-hand side all the possible connectivity patterns are listed and the correct pairs are highlighted with shadowed boxes. An alternative representation is the symmetric connectivity matrix with the correct assignments shown with filled circles.

In order to evaluate the accuracy of the prediction two indexes can be used: \(Q_p\) and \(Q_c\). For a protein \(p\) \(Q_p\) is defined as:

\[
Q_p = \delta(\text{correct pattern, predicted pattern})
\]

where \(\delta(x, y)\) is 1 if and only if the predicted pattern coincides with the correct pattern.

Alternatively, \(Q_c\) is defined as:

\[
Q_c = \frac{\text{number of correctly predicted pairs}}{\text{number of possible pairs}}.
\]

The two indexes are equally suited and complimentary for measuring the accuracy of the prediction: \(Q_p\) is a measure of the predictive performance on each protein (either 1 or 0) and can be averaged over the number of predicted proteins to give a global measure of the accuracy of the method. \(Q_c\) quantifies the accuracy of the method based on the number of pairs correctly predicted with respect to the total number of possible pairs.

In order to score our method we can also compare its performance to that of a random predictor. The probability of a predictor randomly performing \((R_p)\) on the prediction of the connectivity patterns can be computed. In general, given \(2B\) cysteines, the number of possible connectivity patterns is:

\[
N_p = (2B - 1)!! = \Pi_{i \leq B}(2i - 1).
\]

The corresponding probability of \(R_p\) is:

\[
Q_p(R_p) = 1/N_p.
\]

For the random predictor \((R_p)\), \(Q_c\) is:

\[
Q_c(R_p) = 1/(2B - 1).
\]

In the connectivity matrix representation, \(Q_c(R_p)\) can be computed by the ratio of the number of correct pairs (filled circles in Figure 2) to the total number of cells minus the main diagonal.

It is evident that \(Q_p(R_p)\) (equation 5) and \(Q_c(R_p)\) (equation 4) depend both on the number of disulfide bridges, although in a different manner. By definition \(Q_p(R_p)\) is vanishing much faster than \(Q_c(R_p)\), which follows a hyperbolic pattern.

The prediction of disulfide connectivity as maximum-weight perfect matching
It is evident from Figure 2 that the problem of finding the correct connectivity pattern is equivalent to the problem of computing the maximum-weight perfect matching (for review see Papadimitrou and Steiglitz, 1982). If we attribute a weight greater than zero to each pair of cysteines (representing for instance an interaction potential between each pair) we obtain an undirected weighted graph \(G\). \(G\) consists of \(V = 2B\) vertices (the number of cysteines) and \(E = 2B(2B - 1)/2\) undirected edges that represent the strength of the interactions among the cysteines. Then a perfect matching in \(G\) is a subset of
Prediction of disulfide connectivity in proteins

**Fig. 1.** Distribution of the protein chains in the data set grouped according to the number of disulfide bonds in the 3D structure.

**Fig. 2.** Disulfide connectivity patterns for three disulfide bonds (six cysteines) in a protein chain. Top left side: the hypothetical correct connectivity pattern in the chain. Right-hand side: all the possible connectivity patterns are listed and the correct pairs are highlighted with boxes. Bottom left: the symmetric connectivity matrix (filled circles represents the correct cysteine assignments).

edges such that each node in $G$ is met by only one edge in the subset. Given a weight $w_e$ for each edge $e$ of $G$, the **maximum weight perfect matching** problem is to find a perfect matching $M$ of maximum weight ($\Sigma w_e : e \in M$).

Happily, in order to search for the best matching we do not have to evaluate all the possible connectivity patterns (equation 3). Indeed, the problem can be formulated as an instance of a Linear Programming (LP) (Edmonds, 1965) whose worst-case running time is $O(V^4)$, where $V$ is the number of vertices (here $V = 2B$ number of cysteines).

Gabow (1975) improved the solution and presented an algorithm that can solve the problem in a worst-case time bound of $O(V^3)$. Although there are other algorithms with different running time (Cook et al., 1997), we adopted the Gabow’s implementation (EG algorithm), since the graph $G$ at hand is complete and in this case, the Gabow’s implementation is still the best time performing.

The problem of selecting the correct connectivity pattern can be regarded as the problem of finding the optimal potential to generate the weight edges of our graph $G$.

**RESULTS AND DISCUSSION**

**Computing weight edges with residue contact potentials**

Given a protein sequence containing $B$ disulfide-bond candidates, we need a rule to assign the edge values. Starting from the protein covalent sequence, the information we search for can only be found in the local cysteine environment. For this reason we make a very basic assumption: we assume that for each cysteine all its sequence nearest neighbours make contacts. Actually, we cannot know in advance if only some of them could be involved, nor can we know whether there was a specific pairing. Following this assumption, the weight edge $(i, j)$
The first and straightforward approach is to compute equation (6) selecting for $U$ a classical contact pair potential. We chose the potential described by Mirny and Shakhnovich (1996), that is optimized for protein-folding and threading problems. We re-scaled it, since we need that $w_{ij} \geq 0$ (the larger the weight the stronger the interaction). The results are reported in Table 1. Predictions are scored grouping the chains according to the number of disulfide bonds (from 2 to 5). Only well-represented classes are shown (1 disulfide bond is trivial since it is always correct). Noticeably, the results are not distinguishable from that obtained using a random predictor. This indicates that a mean force potential computed by taking into consideration the whole protein structure (such as MS) is not suited to capture the features, if any, of the local environments of the bond-forming cysteines (Fariselli et al., 1999).

### Predicting the disulfide connectivity with the MS contact potential

The first and straightforward approach is to compute equation (6) selecting for $U$ a classical contact pair potential. We chose the potential described by Mirny and Shakhnovich (1996), that is optimized for protein-folding and threading problems. We re-scaled it, since we need that $w_{ij} \geq 0$ (the larger the weight the stronger the interaction). The results are reported in Table 1. Predictions are scored grouping the chains according to the number of disulfide bonds (from 2 to 5). Only well-represented classes are shown (1 disulfide bond is trivial since it is always correct). Noticeably, the results are not distinguishable from that obtained using a random predictor. This indicates that a mean force potential computed by taking into consideration the whole protein structure (such as MS) is not suited to capture the features, if any, of the local environments of the bond-forming cysteines (Fariselli et al., 1999).

### Predicting the disulfide connectivity with a contact potential derived by constrained optimization

In order to derive specific values, we need to generate a potential that maximizes the difference between the scores obtained for the correct cysteine pairs with respect to the wrong one. We define the score $S(U, C_p)$ of a connectivity pattern $C_p$ in a protein sequence $p$ as:

$$S(U, C_p) = \sum_{(i,j) \in C_p} w_{ij}$$  \hspace{1cm} (7)$$

where $w_{ij}$ is computed by equation (6), using a set of potential values $\{U(k, l)\}$. In the 20 amino acidic alphabet (AA), equation (7) becomes:

$$S(U, C_p) = \sum_{i \in AA} \sum_{j \geq i} U(i, j)n(C_p, i, j)$$  \hspace{1cm} (8)$$

where $n(C_p, i, j)$ is the number of times in which the residue pair $(i, j)$ is found in the connectivity pattern $C_p$. In order to maximize the native disulfide connectivity pattern ($C_p^*$) with respect to the average value of the alternatives, the difference $Z_p$ between the score of the correct connectivity patterns and the average of all the possible ones is evaluated:

$$Z_p = S(U, C_p^*) - \langle S(U, C_p) \rangle_{C_p^*}$$  \hspace{1cm} (9)$$

For a given set of $P$ proteins the total sum of $Z_p$:

$$Z = 1/P \sum_{p} Z_p$$  \hspace{1cm} (10)$$

is carried out. After some calculations and considering equation (7), we obtain:

$$Z = \sum_{i \in AA} \sum_{j \geq i} U(i, j)A(i, j)$$  \hspace{1cm} (11)$$

where $A(i, j)$ is:

$$A(i, j) = 1/P \sum_{p} \{n(C_p^*, i, j) - n(C_p^A, i, j)/(2B_p - 1)\}.$$  \hspace{1cm} (12)$$

$B_p$ is the number of disulfide bridges in the protein $p$, $n(C_p^*, i, j)$ and $n(C_p^A, i, j)$ are the number of residue couples of type $(i, j)$ in the native and in the alternative connectivity patterns, respectively. We also impose some constraints to the potential $U$:

$$\langle U \rangle = 0$$

$$\sigma^2(U) = \langle (U - \langle U \rangle)^2 \rangle = 1.$$  \hspace{1cm} (13)$$

The first constraint sets the average interaction between amino acids to zero; (i.e. eliminates non-specific interactions). The second one sets the dispersion of the interaction to one, in order to limit the explosion of some potential values. Then using the Lagrange multipliers and equating the following derivative to 0:

$$\partial[Z - \alpha \langle U \rangle - \beta \sigma^2(U)]/\partial U(i, j) = 0$$  \hspace{1cm} (14)$$

### Table 1. Prediction accuracy of the disulfide connectivity using the MS potential

<table>
<thead>
<tr>
<th>No. of chains</th>
<th>$B$</th>
<th>$Q_p(\text{MS})$</th>
<th>$Q_p(\text{MS})/Q_p(R_p)$</th>
<th>$Q_c(\text{MS})$</th>
<th>$Q_c(\text{MS})/Q_c(R_p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>2</td>
<td>0.34</td>
<td>1.0</td>
<td>0.35</td>
<td>1.0</td>
</tr>
<tr>
<td>153</td>
<td>3</td>
<td>0.05</td>
<td>0.9</td>
<td>0.19</td>
<td>1.0</td>
</tr>
<tr>
<td>103</td>
<td>4</td>
<td>0.009</td>
<td>1.1</td>
<td>0.14</td>
<td>1.0</td>
</tr>
<tr>
<td>44</td>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.12</td>
<td>1.0</td>
</tr>
</tbody>
</table>

$B =$ number of bridges ($2B =$ number of cysteines), $Q_p(\text{MS}) =$ number of correctly assigned connectivity patterns divided by number of proteins in the set. $Q_c(\text{MS}) =$ number of correctly assigned couples divided by the total number of couples in the set. $Q_p(R_p)$ and $Q_c(R_p)$ are as in equations (2) and (3), respectively.
### Prediction of disulfide connectivity in proteins

#### Table 2. Prediction accuracy of the disulfide connectivity using the constrained optimization potential (CO)

<table>
<thead>
<tr>
<th>No. of chains</th>
<th>B</th>
<th>$Q_p(CO)$</th>
<th>$Q_p(CO)/Q_p(R_P)$</th>
<th>$Q_c(CO)$</th>
<th>$Q_c(CO)/Q_c(R_P)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>2</td>
<td>0.42</td>
<td>1.2</td>
<td>0.42</td>
<td>1.2</td>
</tr>
<tr>
<td>153</td>
<td>3</td>
<td>0.09</td>
<td>1.3</td>
<td>0.21</td>
<td>1.0</td>
</tr>
<tr>
<td>103</td>
<td>4</td>
<td>0.05</td>
<td>5.0</td>
<td>0.21</td>
<td>1.5</td>
</tr>
<tr>
<td>44</td>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.14</td>
<td>1.3</td>
</tr>
</tbody>
</table>

$B =$ number of bridges ($2B =$ number of cysteines), $Q_p(MS) =$ number of correctly assigned connectivity patterns divided by number of proteins in the set. $Q_c(MS) =$ number of correctly assigned couples divided by the total number of couples in the set. $Q_p(R_P)$ and $Q_c(R_P)$ are as in equations (2) and (3), respectively.

#### Table 3. Prediction accuracy of the disulfide connectivity using the odd-ratio potential (OR)

<table>
<thead>
<tr>
<th>No. of chains</th>
<th>B</th>
<th>$Q_p(OR)$</th>
<th>$Q_p(OR)/Q_p(R_P)$</th>
<th>$Q_c(OR)$</th>
<th>$Q_c(OR)/Q_c(R_P)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>2</td>
<td>0.46</td>
<td>1.4</td>
<td>0.46</td>
<td>1.4</td>
</tr>
<tr>
<td>153</td>
<td>3</td>
<td>0.17</td>
<td>2.4</td>
<td>0.29</td>
<td>1.4</td>
</tr>
<tr>
<td>103</td>
<td>4</td>
<td>0.11</td>
<td>11.5</td>
<td>0.31</td>
<td>2.2</td>
</tr>
<tr>
<td>44</td>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.17</td>
<td>1.5</td>
</tr>
</tbody>
</table>

$B =$ number of bridges ($2B =$ number of cysteines), $Q_p(MS) =$ number of correctly assigned connectivity patterns divided by number of proteins in the set. $Q_c(MS) =$ number of correctly assigned couples divided by the total number of couples in the set. $Q_p(R_P)$ and $Q_c(R_P)$ are as in equations (2) and (3), respectively.

#### Predicting the disulfide connectivity with an odd-ratio contact potential

An alternative and straightforward potential derivation is the use of the odd-ratio computation (Durbin et al., 1998). In this case, for each ordered residue pair (210) we can compute the frequency in the ‘true’ connectivity patterns ($P(i, j|T)$) and that in the ‘non’-correct ones ($P(i, j|N)$). The odd-ratio contact potential (OR) is defined as

$$U(i, j) = P(i, j|T)/P(i, j|N).$$

(15)

The results (using a 4-fold cross-validation) are listed in Table 3. In spite of the simple procedure to derive the OR potential, the prediction accuracy is improving, becoming higher than that of a random predictor. Noticeably, in the case of the four disulfide bond containing chains, the accuracy of the most stringent index ($Q_p(OR)$), scores 11 times higher than that of the random predictor. We obtain the potential (CO):

$$U(i, j) = [A(i, j) - (A(i, j))]/\sigma^2(A(i, j)).$$

(16)

After computing the 210 (i.e. the 20(20 + 1)/2) elements for the CO potential as indicated by equation (16), we apply (as in the case of the MS potential) a further rescaling to set all the weights $\geq 0$.

Table 2 show the results in cross-validation (taking four different subsets), using the CO potential to evaluate the weight edges. In this case the performance is higher than that obtained with a random predictor. This is particularly true when we look at the $Q_c(CO)$ score, indicating that the CO potential has captured some information about the local cysteine environment.

#### Table 4. Prediction accuracy of the disulfide connectivity using EG algorithm and Monte-Carlo derived potential (EG)

<table>
<thead>
<tr>
<th>No. of chains</th>
<th>B</th>
<th>$Q_p(EG)$</th>
<th>$Q_p(EG)/Q_p(R_P)$</th>
<th>$Q_c(EG)$</th>
<th>$Q_c(EG)/Q_c(R_P)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>2</td>
<td>0.56</td>
<td>1.7</td>
<td>0.56</td>
<td>1.7</td>
</tr>
<tr>
<td>153</td>
<td>3</td>
<td>0.21</td>
<td>3.1</td>
<td>0.36</td>
<td>1.7</td>
</tr>
<tr>
<td>103</td>
<td>4</td>
<td>0.17</td>
<td>17.8</td>
<td>0.37</td>
<td>2.6</td>
</tr>
<tr>
<td>44</td>
<td>5</td>
<td>0.02</td>
<td>22.7</td>
<td>0.21</td>
<td>1.9</td>
</tr>
</tbody>
</table>

$B =$ number of bridges ($2B =$ number of cysteines), $Q_p(MS) =$ number of correctly assigned connectivity patterns divided by number of proteins in the set. $Q_c(MS) =$ number of correctly assigned couples divided by the total number of couples in the set. $Q_p(R_P)$ and $Q_c(R_P)$ are as in equations (2) and (3), respectively.
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.077</td>
<td>0.901</td>
<td>0.892</td>
<td>1.073</td>
<td>1.155</td>
<td>1.111</td>
<td>1.097</td>
<td>1.018</td>
<td>1.143</td>
<td>1.244</td>
<td>1.344</td>
<td>1.180</td>
<td>0.882</td>
<td>1.081</td>
<td>1.087</td>
<td>1.107</td>
<td>0.897</td>
<td>1.122</td>
<td>1.202</td>
<td>0.840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.901</td>
<td>0.566</td>
<td>0.765</td>
<td>0.619</td>
<td>0.798</td>
<td>0.949</td>
<td>0.757</td>
<td>0.971</td>
<td>0.811</td>
<td>0.570</td>
<td>0.914</td>
<td>0.905</td>
<td>0.780</td>
<td>0.758</td>
<td>0.759</td>
<td>0.926</td>
<td>0.845</td>
<td>0.907</td>
<td>0.641</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.952</td>
<td>1.526</td>
<td>0.876</td>
<td>0.828</td>
<td>0.889</td>
<td>0.716</td>
<td>0.990</td>
<td>0.858</td>
<td>1.054</td>
<td>1.205</td>
<td>0.936</td>
<td>0.865</td>
<td>1.145</td>
<td>0.761</td>
<td>1.318</td>
<td>0.998</td>
<td>1.019</td>
<td>1.129</td>
<td>1.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>1.089</td>
<td>1.238</td>
<td>1.028</td>
<td>0.998</td>
<td>1.034</td>
<td>0.944</td>
<td>0.539</td>
<td>0.945</td>
<td>1.007</td>
<td>1.171</td>
<td>1.193</td>
<td>0.941</td>
<td>0.737</td>
<td>1.123</td>
<td>1.106</td>
<td>0.956</td>
<td>0.950</td>
<td>1.024</td>
<td>1.130</td>
<td>0.751</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>1.153</td>
<td>0.625</td>
<td>0.833</td>
<td>1.232</td>
<td>1.195</td>
<td>0.783</td>
<td>0.962</td>
<td>0.787</td>
<td>0.959</td>
<td>1.308</td>
<td>1.359</td>
<td>0.868</td>
<td>0.960</td>
<td>1.082</td>
<td>1.287</td>
<td>1.031</td>
<td>0.888</td>
<td>0.969</td>
<td>1.545</td>
<td>0.879</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.909</td>
<td>1.103</td>
<td>0.877</td>
<td>1.046</td>
<td>0.893</td>
<td>1.304</td>
<td>1.013</td>
<td>0.856</td>
<td>0.822</td>
<td>0.937</td>
<td>0.948</td>
<td>0.936</td>
<td>1.097</td>
<td>1.052</td>
<td>1.035</td>
<td>1.043</td>
<td>0.910</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1.213</td>
<td>1.367</td>
<td>0.716</td>
<td>1.049</td>
<td>0.955</td>
<td>0.919</td>
<td>1.568</td>
<td>0.870</td>
<td>0.764</td>
<td>1.336</td>
<td>1.164</td>
<td>0.797</td>
<td>1.107</td>
<td>0.636</td>
<td>0.686</td>
<td>0.859</td>
<td>1.055</td>
<td>1.103</td>
<td>2.015</td>
<td>1.496</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1.011</td>
<td>1.166</td>
<td>0.982</td>
<td>0.750</td>
<td>1.090</td>
<td>0.845</td>
<td>1.252</td>
<td>0.992</td>
<td>0.799</td>
<td>1.132</td>
<td>1.277</td>
<td>1.117</td>
<td>0.982</td>
<td>1.016</td>
<td>0.828</td>
<td>0.863</td>
<td>1.025</td>
<td>0.838</td>
<td>1.188</td>
<td>0.856</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1.146</td>
<td>0.963</td>
<td>1.036</td>
<td>0.918</td>
<td>1.070</td>
<td>0.859</td>
<td>1.168</td>
<td>1.049</td>
<td>1.017</td>
<td>1.144</td>
<td>0.880</td>
<td>1.141</td>
<td>1.050</td>
<td>1.094</td>
<td>0.906</td>
<td>1.061</td>
<td>0.910</td>
<td>1.139</td>
<td>1.012</td>
<td>1.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>0.961</td>
<td>0.558</td>
<td>1.103</td>
<td>1.174</td>
<td>1.170</td>
<td>0.735</td>
<td>1.130</td>
<td>0.783</td>
<td>0.854</td>
<td>0.857</td>
<td>0.715</td>
<td>0.713</td>
<td>1.295</td>
<td>1.003</td>
<td>0.924</td>
<td>1.150</td>
<td>1.149</td>
<td>1.004</td>
<td>1.280</td>
<td>1.334</td>
<td>1.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>0.985</td>
<td>1.191</td>
<td>0.956</td>
<td>1.045</td>
<td>1.080</td>
<td>0.811</td>
<td>1.295</td>
<td>0.939</td>
<td>0.934</td>
<td>0.948</td>
<td>0.701</td>
<td>0.713</td>
<td>1.008</td>
<td>0.898</td>
<td>0.852</td>
<td>0.906</td>
<td>1.176</td>
<td>1.013</td>
<td>1.290</td>
<td>0.706</td>
<td>1.054</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>0.902</td>
<td>0.907</td>
<td>0.870</td>
<td>1.048</td>
<td>1.273</td>
<td>1.047</td>
<td>0.710</td>
<td>0.980</td>
<td>1.028</td>
<td>0.942</td>
<td>0.793</td>
<td>0.732</td>
<td>1.037</td>
<td>0.940</td>
<td>1.140</td>
<td>0.933</td>
<td>0.943</td>
<td>1.069</td>
<td>1.061</td>
<td>1.133</td>
<td>1.005</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>0.988</td>
<td>1.202</td>
<td>1.266</td>
<td>1.125</td>
<td>1.297</td>
<td>1.047</td>
<td>1.235</td>
<td>0.746</td>
<td>1.292</td>
<td>1.240</td>
<td>0.917</td>
<td>0.838</td>
<td>1.248</td>
<td>1.149</td>
<td>1.166</td>
<td>1.003</td>
<td>1.052</td>
<td>1.296</td>
<td>1.327</td>
<td>1.070</td>
<td>1.115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>0.989</td>
<td>0.756</td>
<td>0.810</td>
<td>1.097</td>
<td>1.187</td>
<td>0.970</td>
<td>1.006</td>
<td>0.706</td>
<td>0.895</td>
<td>1.112</td>
<td>0.948</td>
<td>0.910</td>
<td>0.951</td>
<td>1.296</td>
<td>0.897</td>
<td>0.785</td>
<td>0.878</td>
<td>0.788</td>
<td>1.115</td>
<td>1.216</td>
<td>1.050</td>
<td>1.011</td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>0.901</td>
<td>1.180</td>
<td>1.104</td>
<td>1.055</td>
<td>0.930</td>
<td>1.087</td>
<td>1.053</td>
<td>1.068</td>
<td>1.048</td>
<td>1.038</td>
<td>1.160</td>
<td>0.971</td>
<td>0.941</td>
<td>0.967</td>
<td>0.900</td>
<td>1.111</td>
<td>1.111</td>
<td>1.080</td>
<td>1.723</td>
<td>1.032</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1.095</td>
<td>0.911</td>
<td>1.065</td>
<td>1.026</td>
<td>0.893</td>
<td>0.939</td>
<td>1.053</td>
<td>0.919</td>
<td>0.903</td>
<td>1.218</td>
<td>1.002</td>
<td>0.709</td>
<td>0.901</td>
<td>0.990</td>
<td>0.931</td>
<td>0.994</td>
<td>0.843</td>
<td>0.866</td>
<td>1.153</td>
<td>1.093</td>
<td>0.969</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>0.958</td>
<td>0.839</td>
<td>1.016</td>
<td>1.039</td>
<td>1.057</td>
<td>0.954</td>
<td>1.113</td>
<td>0.733</td>
<td>1.131</td>
<td>1.079</td>
<td>0.992</td>
<td>0.988</td>
<td>1.080</td>
<td>1.136</td>
<td>0.948</td>
<td>1.062</td>
<td>1.151</td>
<td>1.302</td>
<td>0.924</td>
<td>1.097</td>
<td>1.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0.980</td>
<td>0.899</td>
<td>0.932</td>
<td>0.921</td>
<td>1.123</td>
<td>0.937</td>
<td>1.988</td>
<td>0.780</td>
<td>0.708</td>
<td>0.725</td>
<td>0.932</td>
<td>1.026</td>
<td>0.725</td>
<td>0.974</td>
<td>0.751</td>
<td>0.943</td>
<td>0.988</td>
<td>0.983</td>
<td>0.857</td>
<td>0.858</td>
<td>1.115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1.157</td>
<td>0.607</td>
<td>0.997</td>
<td>1.248</td>
<td>1.080</td>
<td>0.938</td>
<td>1.512</td>
<td>0.849</td>
<td>0.952</td>
<td>0.972</td>
<td>0.989</td>
<td>0.749</td>
<td>0.911</td>
<td>0.912</td>
<td>1.106</td>
<td>0.945</td>
<td>0.953</td>
<td>1.019</td>
<td>0.784</td>
<td>0.870</td>
<td>0.781</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*EG* potential values are shown in the upper part.
*Odd-ratio potential values are shown in the lower part.*
Predicting the disulfide connectivity with a contact potential optimized with the EG algorithm and a Monte-Carlo simulated annealing

In this section we introduce another type of potential that takes advantage of the fact that the EG algorithm is extremely fast. Fixing a potential $U$, the EG algorithm gives the connectivity pattern with the maximum weight ($C^0_p$) for the protein sequence $p$. Then we compute the function:

$$ F(U) = 1/P \sum_p \delta(C^*_p, C^0_p) $$

(17)

where $\delta(x, y)$ is 1 if and only if the connectivity pattern $x \equiv y$, and $P$ is the number of protein chains used. In this way we can score the performance of the potential $U$ for all proteins in the data set.

Starting with an initial potential $U$ value, a Monte-Carlo simulated annealing procedure evaluates a new potential $U'$. This is accepted or rejected, according to a probabilistic rule for each step:

$$ p = \min[1, \exp(F(U') - F(U))/T]. $$

(18)

In order to maximize $F(U)$, the OR potentials are used as a starting point. Then 50 new different potentials are generated by means of the same number of Monte-Carlo simulated annealing runs. The runs differ in that different starting and final pseudo-temperatures ($T$) and alternative cooling schemes are used. All runs gave a similar score. Table 4 shows the performance (in 4-fold cross validation) of one set of the best potential at the end of the annealing. Table 5 lists for each possible pair all the $F$ potential values, averaged on the four simulations and bioinformatics analysis. J. Mol. Biol., 300, 975–985.


Overall, the results indicate that with the $EG$ potential it is possible to predict with some extent of success patterns of disulfide connectivity in proteins. The predictor in this case is significantly over performing a random predictor, although still the scores are decreasing at increasing number of disulfide bridges. This finding agrees with the notion of the decrease in probability of correct predictions which can be evaluated directly from equations (4) and (5) for a random predictor as a function of the number of disulfide bridges in the protein chain. Our method is however largely outperforming a random predictor as shown in Table 4, and we propose it as a suited predictor to locate putative disulfide bridges in the protein-folding, particularly when the number of putative bridges is in the range of 2–3. We are confident that as soon as databases will contain a larger number of well annotated proteins with disulfide bridges it will be possible to apply different kind of potentials, taking into consideration also higher order residue interactions to improve the values of weight edges.


