Progress in Predicting Inter-Residue Contacts of Proteins With Neural Networks and Correlated Mutations

Piero Fariselli,1 Osvaldo Olmea,2 Alfonso Valencia,2 and Rita Casadio1*
1CIRB and Department of Biology, University of Bologna, Bologna, Italy
2Protein Design Group, CNB-CSIC Cantoblanco, Madrid, Spain

ABSTRACT This article presents recent progress in predicting inter-residue contacts of proteins with a neural network-based method. Improvement over the results obtained at the previous CASP3 competition is attained by using as input to the network a complex code, which includes evolutionary information, sequence conservation, correlated mutations, and predicted secondary structures. The predictor was trained and cross-validated on a data set comprising the contact maps of 173 non-homologous proteins as computed from their well-resolved three-dimensional structures. The method could assign protein contacts with an average accuracy of 0.21 and with an improvement over a random predictor of a factor greater than 6, which is higher than that previously obtained with methods only based either on neural networks or on correlated mutations. Although far from being ideal, these scores are the highest reported so far for predicting protein contact maps. On 29 targets automatically predicted by the server (CORNET) the average accuracy is 0.14. The predictor is poorly performing on all-\(\alpha\) proteins, not represented in the training set. On all-\(\beta\) and mixed proteins (22 targets) the average accuracy is 0.16. This set comprises proteins of different complexity and different chain length, suggesting that the predictor is capable of generalization over a broad number of features. Proteins 2001;Suppl 5:157–162. © 2002 Wiley-Liss, Inc.

Key words: protein structure predictions; contact maps; correlated mutations; neural networks; residue contacts

INTRODUCTION

A useful two-dimensional representation of a protein three-dimensional (3D) structure is its contact map.1 Secondary structures are easily detected from the contact map. \(\alpha\)-Helices appear as thick bands along the main diagonal involving contacts between residues in position \(i\) and \(i+4\), respectively. Offset parallel or perpendicular bands to the main diagonal are distinguished marks of parallel or antiparallel \(\beta\)-sheets. The remaining contacts in the representation are sparse and/or clustering in segregated areas, depending on the protein structural complexity.

In real proteins, the number of contacts linearly scales with the chains length.2–4 The slope of the linear dependence depends on the contact definition.3 Various ways have been used to define contacts. Routinely, a contact is said to exist between each pair of residues whenever the mutual distance is below a given arbitrary threshold. The distance involved in the different definitions of a contact can be that between the \(C_\alpha-C_\alpha\) atoms,\(^5\) between the \(\text{C}^\beta_\text{P}-\text{C}^\beta_\text{N}\),\(^2,5,6\) and the minimal distance between atoms belonging to the side chain or to the backbone of the two residues.\(^4\)

If the true physical contact map representation of a protein is known, it is possible to recover its 3D structure. The similarity to the native structure is still rather good [low root-mean-square deviation (RMSD) to the crystal] even when the number of true contacts is reduced by a factor of two.\(^3\)

A relevant issue is, therefore, whether it is possible to predict the contact map of a protein starting from the residue sequence and, most importantly, to which extent the prediction can be useful to reconstruct the protein structure. In this article we focus on the accuracy of the prediction of contact maps that can be obtained with our predictor (CORNET) and highlight some future perspectives for this ab initio procedure.

MATERIALS AND METHODS

We developed CORNET, a predictor that is essentially based on neural networks. The system was trained to learn the association rules between the covalent structure of each protein belonging to a selected database and its contact map. Complexity of the input coding, which is rather complex compared with others previously used for the same task, is new in the present version. CORNET was specifically designed to include evolutionary information in the form of sequence profile, sequence conservation, correlated mutations, and predicted secondary structures. We were prompted to modify the input coding by the results obtained at CASP3.\(^7\) A brief description of the method is outlined below.

Grant sponsor: Ministero della Università e della Ricerca Scientifica e Tecnologica; Grant sponsor: Italian Centro Nazionale delle Ricerche (Target: Biotechnology).
O. Olmea’s present address is Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, NY.
*Correspondence to: Rita Casadio, Department of Biology, Via Irnerio 42, I-40126, Bologna, Italy. E-mail: casadio@alma.unibo.it
Received 27 March 2001; Accepted 2 July 2001
Database

For training the network, we use a large set of non-homologous proteins of known 3D structure. The list includes all proteins in the PDB-select list of non-sequence-redundant protein structures whose chain was not interrupted and for which alignments with >15 sequences were obtained: in total our set included 173 proteins. This constraint was introduced to compute correlated mutations. By this, the structural classes represented in the set are α + β proteins (27%) and α/β proteins (73%); apparently, all-α and all-β proteins are not retained with this stringent constraint. Distributed by length, 37% of the proteins have a length <100 residues, 33% comprise from 100 to 169 residues; 17% from 170 to 299 residues; 13% have a length >300 residues. The length distribution is described because it was shown before that accuracy is affected by the protein length (a rough estimate of the protein structural complexity).4,5

Contact Definition

The contact map of a protein is a binary representation of its tertiary structure.4 Two residues in a protein that come into contact with each other form a non-covalent (with the exception of the cysteines that form disulfides) interaction (hydrogen bonds, hydrophobic, or polar pairs, etc.). More formally, a contact map is a symmetric matrix C, whose values C(i,j) (i and j go from 1 to the protein length N), are set to 1 or 0, depending on the distance between the residue i and j d(i,j). We set C(i,j) to 1 when the Euclidean distance between the coordinates of the corresponding Cα atoms is lower than 8 Å (|r_i − r_j| < 8).5 We predict contacts only when the sequence separation between residues is more than six positions (|i − j| ≥7 residues).

Computing Sequence Conservation and Correlated Mutations

Sequence variability was taken from the HSSP database6 or evaluated accordingly. In the HSSP definition, variability is 0 when positions in the multiple sequence alignments are completely conserved, and it increases proportionally to the number of amino acid changes occurring at that position.

Correlated mutations are calculated as previously described.5 Briefly, a distance array is used to codify each position in the alignment. This position-specific array contains all the residue-residue distances between all the possible pairs of sequences at that position. The correlation value between each pair of positions in the alignment is computed as the correlation of the two arrays for each possible residue pair. Corresponding elements in the arrays contain the distance between the same two sequences in the two positions under comparison. The scoring matrix of McLachlan90 defines the distances between residues. Positions with a percentage of gaps >10% are set at a correlation value of −1 and completely conserved positions are set at a correlation value of 0. The similarity value of gaps is set to a dummy value of 0.

Neural Network Architecture and Input Codings

The search of the best topology of the neural network system for the problem at hand was discussed before.4 Here we briefly described CORNET, a predictor available at our web site and used to predict 29 of the CASP4 targets. The number is due to the fact that we entered late into the competition and some of the targets were already expired. A single output neuron codes for contact (output value close to 1) and non-contact (output value close to 0). The hidden layer consists of eight hidden neurons. A new type of input coding was previously introduced4 and it is also used here. Each residue pair in the protein sequence is coded as an input vector containing 210 elements (20 × (20 + 1)/2), representing all the possible ordered couples of residues (considering that each residue couple and its symmetric are coded in the same way). This was done to reduce the number of weight junctions. Evolutionary information is taken into account by changing the binary input used for single sequence coding in a frequency-based one. This means that for any two positions i and j the frequency of each type of ordered couple obtained from the same sequence in the alignment is computed.4 To take into account the sequence neighbors, we use a three-residue-long input window, considering both parallel and anti-parallel pairing of the two segments centered at positions i and j, respectively. This leads to the coding of the couples formed by the residues in positions [i − 1, j − 1], [i, j], [i + 1, j + 1] (parallel pairing) and [i − 1, j + 1], [i, j], [i + 1, j − 1] (antiparallel pairing) ending up with five possible combinations ([i − 1, j − 1], [i, j], [i + 1, j + 1], [i − 1, j + 1], [i + 1, j − 1]) of the ordered couples. This procedure requires 1,050 (210 × 5) input neurons. The neural network takes also into account for each position in the sequence alignment information derived from sequence conservation (1 input node) and correlated mutations (two input nodes for positions i and j, respectively) as previously computed.5 Moreover, the predicted secondary structures are also added as input to the network, with a three-residue-long window for each residue (18 input nodes). Therefore, a total of 1,071 input neurons are used.

The network is trained by using the back-propagation algorithm and a balancing procedure11 to avoid deterioration of the performance due to the large imbalance between contacts (less abundant and proportional to Lp) and non-contacts (more abundant and proportional to Lp). The weight junctions are randomly initialized in the range [−0.01, 0.01]; the learning rate and the momentum term are set to 0.1 and 0.9, respectively. A sorting procedure based on the network output values is adopted. Contacts are defined as the highest Lp prediction values for a protein of length equal to Lp. When CORNET was tested on the 173 proteins of the training set by adopting the cross-validation procedure, the output activation values were routinely >0.75–0.80.

For submitting inter-residue predictions, we chose to fix the number of contacts for each target equal to Lp/2, irrespective of the output value.
Sequence Alignments and Secondary Structure Prediction

Sequence profiles are computed directly from the outputs of PSI-BLAST\(^{12}\) run automatically over the non-redundant database without lower homology cutoffs to exclude doubtful alignments. Secondary structure prediction is performed with a neural network-based predictor implemented in house and available at our web site.\(^{13}\)

Evaluation of the Efficiency

The predictor accuracy is defined as

\[
A = \frac{N_{ep}^*}{N_{ep}} \quad (1)
\]

where \(N_{ep}^*\) and \(N_{ep}\) are the number of correctly assigned contacts and that of total predicted contacts, respectively.

A second index measures the difference in the distribution of the inter-residue distances in the 3D structure for predicted pairs compared with all pair distances in the structure.\(^{14}\) This index is defined as

\[
Xd = \sum_{i=1}^{n} \frac{(P_{ic} - P_{ia})}{d_i} \quad (2)
\]

where \(n\) is the number of bins of the distance distribution (15 equally distributed bins from 4 to 60 Å cluster all the possible distances of residue pairs observed in the protein structure); \(d_i\) is the upper limit (normalized to 60 Å) for each bin, for example, 8 Å for the 4–8 Å bin; \(P_{ic}\) and \(P_{ia}\) are the percentage of predicted contact pairs (with distance between \(d_i\) and \(d_{i-1}\)) and that of all possible pairs, respectively.

RESULTS AND DISCUSSION

A total of 29 contact maps were predicted for an equal number of targets. The results for all the 29 targets are shown in Table I. Targets are grouped according to their structural class, because it was possible to establish
during the evaluation procedure. The table lists the accuracy (A) and the $X_d$ value of the prediction of each target along with the classification by prediction difficulty (Class). Furthermore, for each target, the accuracy of the second-
ary structure prediction (Q3SS), the frequency of the predicted and observed secondary structure type (H or E), the protein length, and the number of chains in the alignment are also shown.

Grouping of the targets by protein structural classes best highlights values and pitfalls of the predictor. Mixed proteins (n/H11005 21) are predicted with an average accuracy equal to 0.16, which is comparable to that of the predictor in testing (Fig. 1). The average accuracy of the predictor on 173 proteins is 0.21, and a peak is detected at 0.15 (Table II). On the large testing set it can be shown that the inclusion of the secondary structure prediction as well as that of correlated mutations increases the predictor performance. However, it is difficult to assess a correlation between the accuracy of secondary structure prediction, the number of sequence in the alignments (if this number is below 15, correlated mutations cannot be effectively evaluated), and the accuracy of inter-residue contacts predictions. While most (n = 16) of the mixed proteins (n = 21) are satisfactorily predicted irrespective of the complexity, T0094 (FR/NF), T0101 (FR/H), T0105 (FR/NF), T0111 (CM), and T0116 (FR/H-A/NF) are poorly predicted. This is so even if the secondary structure is predicted with good accuracy (the secondary structure predictor predicts the 29 targets with (Q3) = 0.76, similar to that of the best performing methods at CASP4).

CORNET outperforms methods based only on correlated mutations particularly for large proteins (Table II). However, contacts are poorly predicted when the protein length largely exceeds 300 residues (T0101, T0111, and T0116). One should notice that even in these cases the network outperforms a random predictor (from 7- to 10-fold, depending on the protein length).

The only example of all-β protein (T0114, FR) is fairly well predicted even if there is only one sequence in the alignment. One should consider that also for the all-β proteins there is no counterpart in the training set. Despite this, the networks during the training procedure on the mixed proteins easily capture features of inter-residue contacts within the β-strands, and this is apparently sufficient to generalize on the never seen before inter-residue contacts of the target. The target is of medium complexity (FR), and its length comprises 87 residues.

It is evident from the data (Table I) that a major limit of the predictor is that inter-residue contacts in all-α proteins are poorly predicted compared with the other types of structural classes (A = 0.07). This is possibly due to the

### Table II. Comparison of the Predictive Performance of CORNET With a Method Based on Correlated Mutations

<table>
<thead>
<tr>
<th>Method</th>
<th>Set</th>
<th>All proteins</th>
<th>L &lt; 100</th>
<th>100 ≤ L &lt; 170</th>
<th>170 ≤ L &lt; 300</th>
<th>L ≥ 300</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(A)</td>
<td>(X)</td>
<td>(A)</td>
<td>(X)</td>
<td>(A)</td>
</tr>
<tr>
<td>CORR</td>
<td>S173</td>
<td>0.09</td>
<td>4.3</td>
<td>0.11</td>
<td>3.69</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>CASP4(29)</td>
<td>0.06</td>
<td>1.9</td>
<td>0.07</td>
<td>-2.8</td>
<td>0.05</td>
</tr>
<tr>
<td>CORNET</td>
<td>S173</td>
<td>0.21</td>
<td>9.9</td>
<td>0.26</td>
<td>10.1</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>CASP4(29)</td>
<td>0.14</td>
<td>6.5</td>
<td>0.18</td>
<td>5.4</td>
<td>0.13</td>
</tr>
</tbody>
</table>

CORR, a method based on correlated mutations; CORNET, the method based on neural networks and including also correlated mutations; Set, the testing sets including the training set and the 29 targets. Performance is also presented by sorting the chains by length.

Fig. 1. Bar graph of the accuracy of the predictor over the proteins of the training set (173) adopting a cross-validation procedure.

![Accuracy](image.png)
lack of representation of all-\(\alpha\) proteins in the training set and can eventually be corrected for by using a more abundant training set. The only exception is target T0102, which being wrongly predicted is seen by CORNET as a mixed protein and predicted accordingly.

Two contact maps of NF are shown in Fig. 2. Predicted contacts were submitted irrespective of the network output values. Assessment indicates that in both test cases when a 0.8 threshold value is used to filter the network output, 19 and 11 contacts are predicted for T0087 and T0115, respectively (second half of the map). The selected contacts are within a 1.5-nm cutoff inter-residue distance. Among these, 6 are correct both for T0087 and T0115 (below 0.8-nm distance). Evidently, a filter would have increased the performance.

Did we do our best? It is evident from the evaluation of the data that several factors can be improved. First, we are confident that a larger training set possibly balanced with respect to the protein structural classes improves the prediction. All-\(\alpha\) proteins seem to have a rather peculiar pattern of inter-\(\alpha\)-helix contacts (large bands perpendicular to the main diagonal, similar to those obtained for membrane proteins; data not shown). This prediction can also be performed separately, by using a predictor especially trained on the structural class. Second, also the alignment procedure can be modified so that those proteins, which do not have a sufficient number of sequences in the alignment, are predicted with a second predictor not trained on correlated mutations. Possibly a jury of networks improves the predictions in all these cases.

Finally, we point out that our goal is an accuracy value equal to 0.5. In this case, we could reconstruct the folding with good accuracy (within 0.4-nm RMSD) for proteins whose chains are as long as 300 residues (our data not shown). For the time being, our best values (in the range of 0.2 accuracy), when used to reconstruct the folding, correspond to structures with a RMSD to the targets equal to approximately 1 nm. This is obviously not satisfactory for a folding method. However, the correct contacts predicted at this stage can be useful in fixing constraints in NMR experiments and/or threading procedures.

ACKNOWLEDGMENTS

Financial support for this work was provided to R.C. by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) delivered to the project “Structural, Functional and Appreciative Prospects of Proteins from Thermophiles” and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR). We thank the Italian Ministero della Università e della Ricerca Scientifica e Tecnologica and the Spanish Minister of the Research for supporting the joint collaboration between Italy and Spain.

REFERENCES


