BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S62–S70

Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners

G. Pollastri[1] and P. Baldi[2,*]

[1]Institute for Genomics and Bioinformatics, Department of Information and Computer Science, University of California, Irvine, Irvine, CA 92697-3425, USA and
[2]Department of Biological Chemistry, College of Medicine, University of California, Irvine, Irvine, CA 92697-3425, USA

Received on January 24, 2002; revised and accepted on March 31, 2002

ABSTRACT

Motivation: Accurate prediction of protein contact maps is an important step in computational structural proteomics. Because contact maps provide a translation and rotation invariant topological representation of a protein, they can be used as a fundamental intermediary step in protein structure prediction.

Results: We develop a new set of flexible machine learning architectures for the prediction of contact maps, as well as other information processing and pattern recognition tasks. The architectures can be viewed as recurrent neural network implemantations of a class of Bayesian networks we call generalized input-output HMMs (GIOHMMs). For the specific case of contact maps, contextual information is propagated laterally through four hidden planes, one for each cardinal corner. We show that these architectures can be trained from examples and yield contact map predictors that outperform previously reported methods. While several extensions and improvements are in progress, the current version can accurately predict 60.5% of contacts at a distance cutoff of 8 Å and 45% of distant contacts at 10 Å, for proteins of length up to 300.

Availability: The contact map predictor will be made available through http://promoter.ics.uci.edu/BRNN-PRED/ as part of an existing suite of proteomics predictors.

Contact: gpollast@ics.uci.edu; pfbaldi@ics.uci.edu

Keywords: protein structure prediction; protein contacts, contact map; graphical models; recurrent neural networks.

INTRODUCTION

The 3D structure of a complex polymer molecule, such as a protein, can be captured to a large extent by its distance map, and its contact map approximation. The distance map is a 2D symmetric matrix where the entry (i, j) represents the distance between elements i and j along the chain and the contact map is the binary clipped version of the distance map, where contact is defined according to some distance cutoff. For proteins, such maps can be defined at different scales of resolution from the level of single atoms, to the level of amino acids (for instance distances between C_α atoms), to the level of secondary structural elements, as in protein cartoons (Westhead *et al.*, 1998). While these maps do not contain all the information about a protein–for instance a protein and its mirror image have the same maps–it is clear that the coarse contact map provides a good representation of the overall topology of a protein and the detailed map at the amino acid level does indeed capture most of the relevant structural information. In fact, these contact maps can be used in a number of structural proteomics tasks, for instance as protein fingerprints for rapid comparison of two protein structures (Godzik *et al.*, 1992) and in protein structure and folding prediction (Selbig and Argos, 1998). Knowing the correct positions of residue contacts in proteins has proven to be extremely useful in determining the three-dimensional structure of a given protein, as demonstrated in the CASP3 and CASP4 experiments (see http://predictioncenter.llnl.gov/ and Ortiz *et al.* (1999); Lesk *et al.* (2001)). Likewise, the number of stabilizing contacts that residues make in the protein-folded globule (see Dill (1999) for a review) is a fundamental aspect of protein structure that is well worth predicting.

One of the fundamental problems in protein structure prediction is that protein structures are invariant under translations and rotations. These fundamental invariances must be addressed by the representations used in any algorithm that attempts to predict protein tertiary structures. A good representation that is translation and rotation invariant is particularly essential to machine learning approaches that attempt to learn how to predict protein structures from training examples extracted from the Protein Data Bank (PDB). It is precisely the rotation and

[*]To whom all correspondence should be addressed.

translation invariance of distance and contact maps that is leveraged in the protein-structure-prediction strategy we have outlined in Baldi and Pollastri (2002b).

The strategy decomposes the overall problem into three steps: (1) prediction of structural features, such as secondary structure or relative solvent accessibility from primary sequence; (2) prediction of distance and/or contact maps from primary sequence and structural features; and (3) prediction of 3D structure from 2D maps (Figure 1). Results obtained for the first step of this pipeline strategy have been described in Pollastri *et al.* (2001a,b). In addition, for the third step of recovering 3D coordinates from contact maps, several algorithms have been described in the NMR literature (Nilges *et al.*, 1988a,b) and elsewhere (Vendruscolo *et al.*, 1997), using distance geometry and stochastic optimization techniques, that seem to work reasonably well. Thus, the main focus of this paper is on the second and most difficult step, specifically the prediction of detailed contact maps at the amino acid level. Similar ideas, however, can be applied to distance maps, and to distance and contact maps at the level of secondary structure elements.

Beyond protein structure, accurate prediction of protein contact maps may be important for protein folding studies. Indeed, according to one of the dominant models folding occurs in the form of a hierarchical process where the ‘leaves’ correspond to the formation of local, possibly unstable and flickering, secondary structures that are progressively recruited in a hierarchical fashion during the folding process. Within this paradigm, we hypothesize that the hierarchical organization of the corresponding tree could be read from the contact map, the most distant contacts corresponding to the deepest levels in the hierarchy. Major contacts located on the same diagonal band of the map could be prioritized using energy, hydrophobicity, or coordination number considerations.

Various algorithms for the prediction of contacts (Shindyalov *et al.*, 1994; Olmea and Valencia, 1997; Fariselli and Casadio, 1999, 2000; Pollastri *et al.*, 2001a), distances (Aszodi *et al.*, 1995; Lund *et al.*, 1997; Gorodkin *et al.*, 1999), and contact maps (Fariselli *et al.*, 2001) have been developed, in particular using neural networks. The best contact map predictor in the literature and at the last CASP prediction experiment reports an average off-diagonal precision (true positives/total positives) of 21% (Fariselli *et al.*, 2001). While this result is encouraging and well above chance level by a factor greater than 6, it is still far from providing sufficient accuracy for reliable 3D structure prediction. A key issue in this area is the amount of noise that can be tolerated in a contact map prediction without compromising the 3D-reconstruction step. While to the best of our knowledge systematic tests in this area have not yet been published, preliminary results appear to indicate that recovery of about 50% of distant contacts around an 8 Å distance cutoff ought to suffice for proper reconstruction, at least for proteins up to 150 amino acid long (Rita Casadio and Piero Fariselli, private communication and oral presentation during CASP4).

In this paper we introduce a new general class of graphical model architectures together with their associated implementations in terms of recurrent neural network that significantly improves the state-of-the-art in contact map prediction.

METHODS

The new architectures we introduce can be viewed as 2D generalizations of the BIOHMMs (bi-directional input-output HMMs) and BRNNs (bi-directional recursive neural networks) developed for several sequence analysis problems, including the prediction of protein secondary structure, coordination numbers, and solvent accessibility (Baldi and Brunak, 2001; Pollastri *et al.*, 2001a,b). The description of the architecture involves two steps: (1) the construction of a statistical graphical model (Bayesian network) for 2D contact maps; and (2) the reparameterization of the graphical model using artificial recurrrent neural networks. Additional theoretical details and generalizations can be found in Baldi and Pollastri (2002).

GIOHMM Bayesian networks for contact maps

To predict contact maps the key question is how can bidirectional IOHMMs be generalized from 1D to 2D? It turns out that there is a ‘canonical’ 2D generalization described in Figures 2 and 3. In its basic version, the generalization consists of a Bayesian network organized into six horizontal layers or planes: one input plane, 4 hidden planes, and one output plane. Each plane contains N^2 nodes arranged on the vertices of a square lattice. Thus in each vertical column there is an input unit $I_{i,j}$, four hidden units $H^{NE}_{i,j}$, $H^{NW}_{i,j}$, $H^{SW}_{i,j}$, and $H^{SE}_{i,j}$ associated with the four cardinal corners, and an output unit $O_{i,j}$ with $i = 1, \dots, N$ and $j = 1, \dots, N$. In each hidden plane, the edges are oriented towards the corresponding cardinal corner. In the NE plane, for instance, all edges are oriented towards the North or the East. It is easy to check that this defines a DAG (directed acyclic graph) and therefore a proper support for a Bayesian network probabilistic model. The parameters of this Bayesian network are the local conditional probability distributions:

$$\begin{cases} P(O_i | I_{i,j}, H^{NE}_{i,j}, H^{NW}_{i,j}, H^{SW}_{i,j}, H^{SE}_{i,j}) \\ P(H^{NE}_{i,j} | I_{i,j}, H^{NE}_{i-1,j}, H^{NE}_{i,j-1}) \\ P(H^{NW}_{i,j} | I_{i,j}, H^{NW}_{i+1,j}, H^{NW}_{i,j-1}) \\ P(H^{SW}_{i,j} | I_{i,j}, H^{SW}_{i+1,j}, H^{SW}_{i,j+1}) \\ P(H^{SE}_{i,j} | I_{i,j}, H^{SE}_{i-1,j}, H^{SE}_{i,j+1}) \end{cases}$$

with the obvious adjustments at the boundaries.

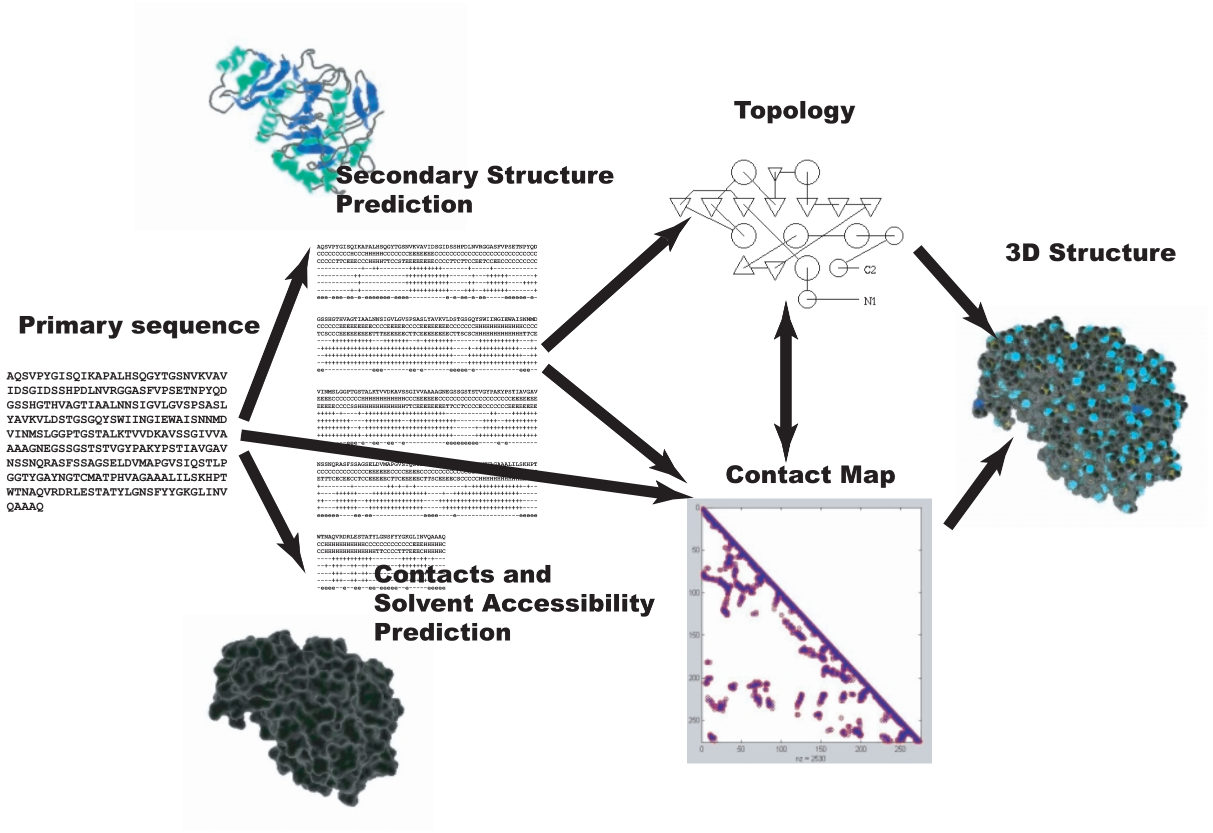
Topology
Secondary Structure Prediction
3D Structure
Primary sequence
AQSVPYGISQIKAPALHSQGYTGSNVKVAV
IDSGIDSSHPDLNVRGGASFVPSETNPYQD
GSSHGTHVAGTIAALNNSIGVLGVSPSASL
YAVKVLDSTGSGQYSWIINGIEWAISNNMD
VINMSLGGPTGSTALKTVVDKAVSSGIVVA
AAAGNEGSSGSTSTVGYPAKYPSTIAVGAV
NSSNQRASFSSAGSELDVMAPGVSIQSTLP
GGTYGAYNGTCMATPHVAGAAALILSKHPT
WTNAQVRDRLESTATYLGNSFYYGKGLINV
QAAAQ
C2
N1
Contact Map
Contacts and Solvent Accessibility Prediction
0
50
100
150
200
250
nz = 2530

Fig. 1. Overall pipeline strategy for machine learning protein structures. Example of 1SCJ (Subtilisin-Propeptide Complex) protein. The first stage predicts structural features including secondary structure, contacts, and relative solvent accessibility. The second stage predicts the topology of the protein, using the primary sequence and the structural features. The coarse topology is represented as a cartoon providing the relative proximity of secondary structure elements, such as alpha helices and beta-strands. The high-resolution topology is represented by the contact map between the residues of the protein. The final stage is the prediction of the actual 3D coordinates of all residues and atoms in the structure.

The architecture described can flexibly accommodate input sequences of any length and, in fact, is a special case of a much more general class of architectures we call GIOHMMs (generalized IOHMMs) introduced in Baldi and Pollastri (2002). The most general definition of a GIOHMM is a Bayesian network where the graph consists of a set of N input nodes, M output nodes, and a DAG hidden layer (possibly with multiple connected components), with additional connections running from the input to the output nodes, and from the input nodes to the hidden nodes. These architectures can process variable data structures such as sequences, trees, other DAGs, and images.

The connectivity of the 2D GIOHMM described above can easily be enriched. For instance, in the 2D hidden planes of Figure 2 one can add the diagonal connections associated with the triangular (or hexagonal) lattice. By doing so, the length of a diagonal path is cut in half in the 2D case (by D in the D-dimensional case) with only a very moderate increase in the number of model parameters. This is useful for contact map prediction where the main diagonal path is associated with distance 0, and the ability to align a sequence to itself easily (i.e. along a shorter path) may facilitate and accelerate learning.

2D recurrent neural network architectures

GIOHMMs are Bayesian networks and therefore the general propagation and learning algorithms for Bayesian networks can be applied to them (Pearl, 1988; Lauritzen, 1996; Heckerman, 1997, 1998). The 2D GIOHMMs we have described, however, contain many undirected cycles and therefore require propagation algorithms that are often computationally demanding, such as the junction-tree algorithm. In part to overcome this point, the architectures can be simplified by recasting them in terms of neural networks (or some other general set of functions) which

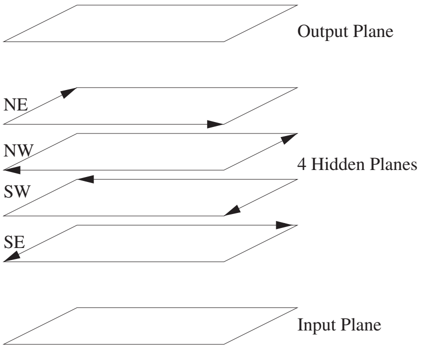

Fig. 2. General layout of Bayesian network for processing two-dimensional objects such as contact maps, with nodes regularly arranged in one input plane, one output plane, and four hidden planes. In each plane, nodes are arranged on a square lattice. The hidden planes contain directed edges associated with the square lattices. All the edges of the square lattice in each hidden plane are oriented in the direction of one of the four possible cardinal corners: NE, NW, SW, SE. Additional directed edges run vertically in column from the input plane to each hidden plane, and from each hidden plane to the output plane.

amounts to consider that the state of each node is a deterministic vector that depends on the value of the parent vectors through a function that is implemented by a neural network (Baldi and Chauvin, 1996; Frasconi *et al.*, 1998; Pollastri *et al.*, 2001a,b).

When the DAG(s) in the hidden layer of a GIOHMM have a regular structure (e.g. lattices, binary trees, etc) we can further introduce an assumption of spatial stationarity, or weight sharing, that considerably reduce the number of parameters. As a result, the neural networks used by different nodes are in fact the same. In 1D, this leads to recurrent neural network architectures that can intuitively be thought of as 'wheels' that are rolled over the hidden DAGs. It is not difficult to see that learning can then be achieved in these neural-network-parameterized GIOHMMs using gradient descent amounting to back-propagation through unfolded space, or structure.

The same ideas can be applied to the 2D GIOHMMs. Here the output and the hidden layer propagations are parameterized by 5 neural networks in the form

$$\begin{cases} O_{ij} = \mathcal{N}_O(I_{ij}, H^{NW}_{i,j}, H^{NE}_{i,j}, H^{SW}_{i,j}, H^{SE}_{i,j}) \\ H^{NE}_{i,j} = \mathcal{N}_{NE}(I_{i,j}, H^{NE}_{i-1,j}, H^{NE}_{i,j-1}) \\ H^{NW}_{i,j} = \mathcal{N}_{NW}(I_{i,j}, H^{NW}_{i+1,j}, H^{NW}_{i,j-1}) \\ H^{SW}_{i,j} = \mathcal{N}_{SW}(I_{i,j}, H^{SW}_{i+1,j}, H^{SW}_{i,j+1}) \\ H^{SE}_{i,j} = \mathcal{N}_{SE}(I_{i,j}, H^{SE}_{i-1,j}, H^{SE}_{i,j+1}) \end{cases} \qquad (1)$$

omitting possible diagonal connections if a triangular lattice is used instead of a square lattice.

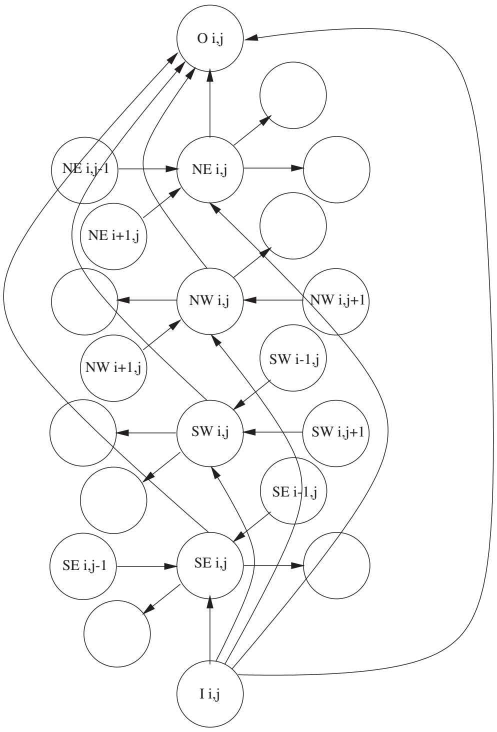

Fig. 3. Details of connections within one column of Figure 2. The input unit is connected to the four hidden units, one in each hidden plane. The input unit and the hidden units are connected to the output unit. $I_{i,j}$ is the vector of inputs at position (i, j). $O_{i,j}$ is the corresponding ouput. Connections of each hidden unit to its lattice neighbors within the same plane are also shown.

In the NE plane, for instance, the boundary conditions are set to $H^{NE}_{ij} = 0$ for $i = 0$ or $j = 0$. The activity vector associated with the hidden unit H^{NE}_{ij} depends on the local input I_{ij}, and the activity vectors of the units $H^{NE}_{i-1,j}$ and $H^{NE}_{i,j-1}$. Activity in NE plane can be propagated row by row, West to East, and from the first row to the last (from South to North), or column by column South to North, and from the first column to the last, or in zig-zag fashion running up and down successive diagonal lines oriented SE to NW.

Learning

Learning in neural-network-parameterized GIOHMMs can proceed by gradient descent for recurrent networks (Baldi, 1995), on typical error functions such as mean squared or relative entropy, by unfolding the structures through time or space. In practice however, it is not always trivial to get gradient descent learning procedures to work well in recurrent networks due, for instance, to the fact that error gradients can vanish rapidly as a function of time, and that learning procedure can become stuck in poor local minima. In learning contact maps, we also typically find large plateaux in the error function when we start training. After some experimentation, we find useful to use a clipped learning rule, applied on-line, protein per protein, where the update Δw_{ij} for a weight w_{ij} is piecewise linear in three different ranges according to

$$\Delta w_{ij} = \begin{cases} \eta \times 0.1 & : \text{ if } \delta w_{ij} < 0.1 \\ \eta \times \delta w_{ij} & : \text{ if } 0.1 < \delta w_{ij} < 1.0 \\ \eta & : \text{ if } 1.0 < \delta w_{ij} \end{cases}$$

where η is the learning rate, and δw_{ij} is the backpropagated error. The learning rate η is equal to 0.1 divided by the number of proteins in the training set. Prior to learning, the weights of each unit in the various neural networks are randomly initialized. The standard deviations, however, must be controlled in a flexible way to avoid any bias and ensure that the expected total input into each unit is roughly in the same range.

Data

The training and testing data sets are extracted from the Protein Data Bank (PDB) of solved structures using the PDB_select list (Hobohm *et al.*, 1992) of February 2001, containing 1520 proteins. The list of structures and additional information can be obtained from the following ftp site: ftp://ftp.embl-heidelberg.de/pub/databases. To avoid biases, the set is redundancy-reduced, with an identity threshold based on the distance derived in (Abagyan and Batalov, 1997), which corresponds to a sequence identity of roughly 22% for long alignments, and higher for shorter ones. This set is further reduced by excluding those chains whose backbone is interrupted. To extract 3D coordinates, together with secondary structure, solvent accessibility and information on beta-sheet partners, we run the DSSP program (Kabsch and Sander, 1983) on all the PDB files in the PDB_select list, excluding those for which DSSP crashes due, for instance, to missing entries or format errors. The final set consists of 1484 proteins. To speed up training and because most comparable systems have been developed and tested on short proteins, we further extract the subsets of all proteins of length less than 300 containing 1269 proteins, and all proteins of length less than 100 containing 533 proteins. The subset of 533 short proteins already contains over 2.3 million pairs of amino acids. Because contact maps strongly depend on the selection of the distance cutoff, we use different thresholds of 6, 8, 10 and 12 Å, yielding four different classification tasks. The number of pairs of amino acid in each class and each contact cutoff is given in Table 1. Note that contact maps are biased towards non-contacts– typically for small cutoffs a contact map of size N^2 contains a number of contacts that is linear in N.

Table 1. Data set composition, with number of pairs of amino acids that are separated by less (close) or more (far) than the distance thresholds in angstroms)

	6 Å	8 Å	10 Å	12 Å
far	2125656	2010198	1825934	1602412
close	202427	317885	502149	725671
total	2328083	2328083	2328083	2328083

Inputs

In contact map prediction, one obvious input at each (i, j) location is the pair of corresponding amino acids yielding two sparse binary vectors of dimension 20 with orthogonal encoding.

A second type of input consideration is the use of profiles and correlated mutation (Gobel *et al.*, 1994; Pazos *et al.*, 1997; Olmea *et al.*, 1999; Fariselli *et al.*, 2001). Profiles, essentially in the form of alignment of homologous proteins, implicitly contain evolutionary and 3D structural information about related proteins. This information is relatively easy to collect using well-known alignment algorithms that ignore 3D structure and can be applied to very large data sets of proteins, including many proteins of unknown structure. The use of profiles improves the prediction of secondary structure by several percentage points, presumably because secondary structure is more conserved than primary amino acid sequence. As in the case of secondary structure prediction, the input can be enriched by taking the profile vectors at positions i and j, yielding two 20-dimensional probability vectors. We use profiles derived using the PSI-BLAST program as described in Pollastri *et al.* (2001b).

When a distant pair (i, j) of positions in a multiple alignment is considered, horizontal correlations in the sequences may exist that are completely lost when each profile is entered independently of the other. These correlations can result from important 3D structural constraints. Thus an expanded input, which retains this information, consists of a 20×20 matrix corresponding to the probability distribution over all pairs of amino acids observed in the two corresponding columns of the alignment. A typical alignment contains a few dozen sequences and therefore in general this matrix is sparse. The unobserved entries can be set to zero or regularized to small non-zero values using standard Dirichlet priors (Baldi and Brunak, 2001).

While this has not been attempted here, even larger inputs can be considered where correlations are extracted not only between positions i and j but also with respect to their immediate neighborhoods including, for instance $i - 1$, $i + 1$, $j - 1$, and $j + 1$. This could compensate for small alignment errors but would rapidly lead also to intractably large inputs of size 20^{2k}, where k is the size of the neighborhood considered. Compression techniques using weight sharing and/or higher-order neural networks would have to be used in conjunction with these very expanded inputs.

Finally, specific structural features can also be added to each input such as the secondary structure classification and the relative solvent accessibility. These features increase the number of inputs by 10 for each pair of amino acids, five inputs for each position (α, β, γ,buried,exposed). The value of these indicators is close to exact when derived from PDB structures, but would be noisier when estimated by a secondary structure and accessibility predictor.

Previous studies Fariselli *et al.* (2001) have used somewhat comparable inputs but have failed to assess the contribution of each feature to the overall performance. To specifically address this issue, we use inputs of size $|I| = 40$ (just the two amino acids or the two profiles), size $|I| = 400$ (correlated profiles), as well as size $|I| = 410$ (correlated profiles, plus the secondary structure and relative solvent accessibility of each amino acid in the pair).

RESULTS

In the simulations, we use a 2D GIOHMM approach with four hidden plane lattices with diagonal edges associated with four similar but independent neural networks. In each neural network we use a single hidden layer. Thus, a given architecture is described by three key parameters: (1) the number NHO of hidden units in the output neural network; (2) the number NHH of hidden units in each of the four hidden neural networks associated with lateral propagation in each of the four planes; and (3) the number NOH of units in the output layer of the four hidden neural networks, corresponding to the dimension of the vector encoding a hidden state in each of the four hidden planes. While we have experimented with several architectures, the results we report are for $NHH = NHO = NOH = 8$ corresponding to 17,114 parameters when an input of size 20×20 is used.

Results of contact map predictions at four distance cutoffs are provided in Table 2. In this experiment, the system is trained on half the set of proteins of length less than 100, and tested on the other half. These results are obtained with plain sequence inputs (amino acid pairs), i.e. without any information about profiles, correlated mutations, or structural features. For a cutoff of 8 Å, for instance, the system is capable of recovering 62.3% of contacts.

It is of course essential to be able to predict contact maps also for longer proteins. This can be attempted by training the recurrent neural network architectures on larger data sets, containing long proteins. While such experiments are in progress, it should be noted that because the systems we have developed can accommodate inputs of arbitrary lengths, we can still use a system trained on short proteins ($l \leq 100$) to produce predictions for longer proteins. In fact, because the overwhelming majority of contacts in proteins are found at linear distances shorter than 100 amino acids, it is reasonable to expect a decent performance from such a system. Indeed, this is what we observe in Table 3. At a cutoff of 8 Å, the percentage of correctly predicted contacts for all proteins of length up to 300 is still 54.5%.

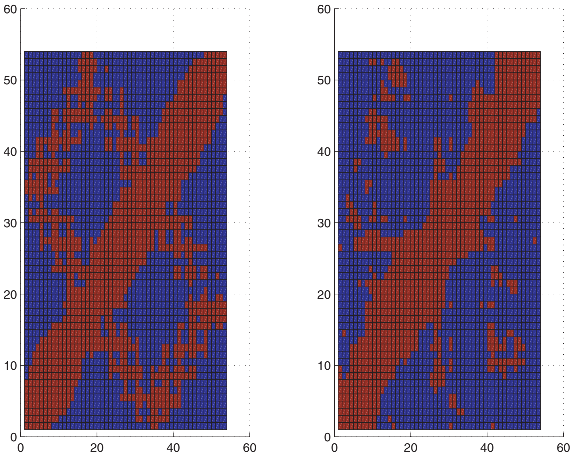

Fig. 4. Example of exact (left) and predicted contact map for protein 1DEEH, prior to symmetrization of the prediction. Color code: blue = 0 (non-contact), red = 1 (contact).

Table 2. Percentages of correct predictions for different contact cutoffs on the validation set. Model trained and tested on proteins of length less than 100. Inputs correspond to simple pairs of amino acids in the sequence (without profiles, correlated profiles, or structural features)

	6 Å	8 Å	10 Å	12 Å
Far	99.1%	98.9%	97.8%	96.0%
Close	66.5%	62.3%	54.2%	48.1%
All	96.2%	93.9%	88.6%	81.0%

A typical example of prediction is reported in Figure 4. In this example we display the raw output of the network which is *not* symmetric since symmetry constraints are not enforced during learning. A symmetric output is easy to derive from a non-symmetric output by averaging the output values at positions (i, j) and (j, i). Application of this averaging procedure yields a small improvement in the overall prediction performance, as seen in Table 4. A possible alternative is to enforce symmetry during the training phase.

The results of additional experiments conducted with larger inputs are displayed in Tables 5 and 6. When inputs

Table 3. Percentages of correct predictions for different contact cutoffs on the validation set. Model trained on proteins of length less than 100, but tested on all proteins with length up to 300. Inputs correspond to simple pairs of amino acids in the sequence

	6 Å	8 Å	10 Å	12 Å
Far	99.6%	99.6%	99.2%	97.8%
Close	64.5%	54.5%	45.7%	39.9%
All	98.3%	96.8%	93.7%	88.6%

Table 4. Same as Table 3 but with symmetric prediction constraints

	6 Å	8 Å	10 Å	12 Å
Far	99.1%	98.9%	97.8%	96.0%
Close	67.3 (+0.8)%	63.1 (+0.8)%	54.9 (+0.7)%	49.0 (+0.9)%
All	96.3 (+0.1)%	94.0 (+0.1)%	88.7 (+0.1)%	81.3 (+0.3)%

of size 20 × 20 corresponding to correlated profiles are used, the performance increases marginally by roughly 1% for contacts (for instance, 1.4% at 6 Å and 1.3% at 12 Å) (Table 5). When both secondary structure and relative solvent accessibility (at 25% threshold) are added to the input, however, the performance shows a remarkable further improvement in the 3–7% range for contacts. For example at 6 Å contacts are predicted with 73.8% accuracy. The last row of Table 6 provides the standard deviations of the accuracy on a per protein basis. These standard deviations are reasonably small so that most proteins are predicted at levels close to the average. These results support the view that secondary structure and relative solvent accessibility are very important for the prediction of contact maps and more useful than profiles or correlated profiles. This is also confirmed by the results obtained by this model (trained on short proteins with correlated profile inputs augmented by structural features) when tested on proteins of length up to 300 (Table 7). At an 8 Å cutoff, the model still predicts over 60% of the contacts correctly, achieving state-of-the-art performance above any previously reported results. In terms of off-diagonal prediction, the sensitivity for amino acids satisfying $|i-j| \geq 7$ is 0.27 at 8 Å and 0.45 at 10 Å, to be contrasted with 0.21 at 8.5 Å reported in Fariselli *et al.* (2001). Finally, a further small improvement can be derived by combining the output of the four predictors using another similar architecture with $NHH = NOH = NHO = 5$ (Table 8) trained on the same data since no overfitting is detected. The global improvement is most visible at 12 Å with a 0.5% improvement over Table 6. But even at 6 Å, there is a non-trivial 0.4% improvement on the prediction of contacts.

Table 5. Percentages of correct predictions for different contact cutoffs on the validation set. Model trained and tested on proteins of length less than 100. Inputs of size 20 × 20 correspond to correlated profiles in the multiple alignments derived using the PSI-BLAST program

	6 Å	8 Å	10 Å	12 Å
Far	99.0%	98.9%	97.6%	96.1%
Near	67.9%	63.0%	55.3%	49.4%
All	96.3%	94.0%	88.5%	81.5%

Table 6. Percentages of correct predictions for different contact cutoffs on the validation set. Model trained and tested on proteins of length less than 100. Same as Table 5 but inputs include also secondary structure and relative solvent accessibility at a threshold of 25% derived from the DSSP program. Last row represents standard deviations on a per protein basis

	6 Å	8 Å	10 Å	12 Å
Far	99.6%	99.5%	98.5%	95.3%
Near	73.8%	67.9%	58.1%	55.5%
All	97.3%	95.2%	89.8%	82.9%
Std	2.3%	3.7%	5.9%	8.5%

CONCLUSION

We have introduced GIOHMMs, a broad class of adaptive graphical models for processing data structures, where outputs depends on inputs and hidden dynamic in an underlying hidden DAG. GIOHMMs can be reparameterized using neural networks. The topological structure of the hidden DAG is crucial and must be tuned to the problem at hand. When the hidden DAG has a regular structure, stationary assumptions lead to recurrent neural network architectures. In the cases of $d = 1$ and $d = 2$, we have shown that these architectures are efficient both in terms of training and addressing real world problems in biological sequence analysis. We believe these architectures are suitable for other applications for instance in image processing or the processing of chemical structures, but additional work is required in these directions. In particular, it is clear that GIOHMMs could be stacked hierarchically, for instance, to address scaling problems. If nothing else, GIOHMMs provide also a new paradigm for thinking about massive lateral information processing in neural architectures.

In the simulations we have reported for the problem of predicting contact maps, these architectures have led to encouraging results that appear to exceed the performance levels of other systems reported in the literature. The methods developed and the results immediately suggest several other experiments which are all in progress at the present time. These include:

- Training larger architectures and, for instance, training

Table 7. Percentages of correct predictions for different contact cutoffs on the validation set. Model trained on proteins of length less than 100, but tested on all proteins with length up to 300. Inputs include correlated profiles, secondary structure, and relative solvent accessibility (at 25%)

	6 Å	8 Å	10 Å	12 Å
Far	99.9%	99.9%	99.3%	95.1%
Near	70.7%	60.5%	51.3%	49.2%
All	98.8%	97.5%	94.4%	87.8%

Table 8. Percentages of correct predictions for different contact cutoffs on the validation set obtained by a network combining four predictors trained on each distance cutoff. Model trained on proteins of length less than 100 and tested on proteins with length up to 100. Inputs include correlated profiles, secondary structure, and relative solvent accessibility (at 25%)

	6 Å	8 Å	10 Å	12 Å
Far	99.6%	99.0%	97.4%	96.5%
Near	74.1%	70.8%	62.1%	54.8%
All	97.3%	95.2%	89.8%	83.4%

ensembles of architectures as is routinely done for other problems such as secondary structure prediction.

- Training architectures on larger data sets containing both short and long proteins.
- Training architectures to predict coarse contact maps, or distances rather than contacts.
- Using contact map prediction to recover and possibly refine predictions for other structural features such as secondary structure, pairing of beta strands, disulphide bridges or, more trivially, amino acid coordination number. The quality of these predictions will have to be compared to the predictions of the corresponding specialized predictors.

In the end, of course, the ultimate test is to combine the contact map predictor with a 3D reconstruction algorithm to produce a complete predictor of protein tertiary structures that is complementary to other approaches (Baker and Sali, 2001; Simons *et al.*, 2001), and can be used for large-scale structural proteomics projects. Indeed, most of the computational time in a machine learning approach is absorbed by the training phase. While training can take several weeks, once trained, a system can produce predictions almost faster than proteins can fold. A complete predictor is expected to be available on line soon and could be used to predict protein structures on a genomic scale. Not only the structure of all proteins in a given genome could be predicted, but also the structure of a large number of polymorphic variants. Furthermore, the same methods could be applied to synthetic proteins and to protein variants obtained by high-throughput techniques, such as phage display/shotgun scanning experiments (Weiss *et al.*, 2000).

ACKNOWLEDGEMENTS

The work of PB and GP is in part supported by a Laurel Wilkening Faculty Innovation award, and a Sun Microsystems award to PB at UCI. The 2D lattice architectures were first presented at a Bioinformatics School in San Miniato, Italy in September 2001.

REFERENCES

Abagyan,R.A. and Batalov,S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.*, **273**, 355–368.

Aszodi,A., Gradwell,M.J. and Taylor,W.R. (1995) Global fold determination from a small number of distance restraints. *J. Mol. Biol.*, **251**, 308–326.

Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.

Baldi,P. (1995) Gradient descent learning algorithms overview: A general dynamical systems perspective. *IEEE Trans. Neural Networks*, **6**, 182–195.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: the Machine Learning Approach*, Second edition, MIT Press, Cambridge, MA.

Baldi,P. and Chauvin,Y. (1996) Hybrid modeling, HMM/NN architectures, and protein applications. *Neural Comput.*, **8**, 1541–1565.

Baldi,P. and Pollastri,G. (2002a) Generalized IOHMMs and recurrent neural network architectures. Submitted.

Baldi,P. and Pollastri,G. (2002b) A machine learning strategy for protein analysis. *IEEE Intelligent Systems. Special Issue on Intelligent Systems in Biology*, **17**, 28–35.

Dill,K. (1999) Polymer principles and protein folding. *Protein Sci.*, **8**, 1166–1180.

Fariselli,P. and Casadio,R. (1999) Neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.

Fariselli,P. and Casadio,R. (2000) Prediction of the number of residue contacts in proteins. *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*. AAAI Press, Menlo Park, CA, pp. 146–151.

Fariselli,P., Olmea,O., Valencia,A. and Casadio,R. (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng.*, **14**, 835–843.

Frasconi,P., Gori,M. and Sperduti,A. (1998) A general framework for adaptive processing of data structures. *IEEE Trans. Neural Networks*, **9**, 768–786.

Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet.*, **18**, 309–317.

Godzik,A., Skolnick,J. and Kolinski,A. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, **227**, 227–238.

Gorodkin,J., Lund,O., Andersen,C.A. and Brunak,S. (1999) Using sequence motifs for enhanced neural network prediction of protein distance constraints. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*

(ISMB99), La Jolla, CA. AAAI Press, Menlo Park, CA, pp. 95–105.

Heckerman,D. (1997) Bayesian networks for data mining. *Data Mining and Knowl. Discov.*, **1**, 79–119.

Heckerman,D. (1998) A tutorial on learning with Bayesian networks. In Jordan,M. (ed.), *Learning in Graphical Models*. Kluwer, Dordrecht.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative data sets. *Protein Sci.*, **1**, 409–417.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Lauritzen,S.L. (1996) *Graphical Models*. Oxford University Press, Oxford.

Lesk,A.M., Conte,L.L. and Hubbard,T.J.P. (2001) Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins*, **45 S5**, 98–118.

Lund,O., Frimand,K., Gorodkin,J., Bohr,H., Bohr,J., Hansen,J. and Brunak,S. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.*, **10**, 1241–1248.

Nilges,M., Clore,G.M. and Gronenborn,A.M. (1988a) Determination of three-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms. *FEBS Lett.*, **239**, 129–136.

Nilges,M., Clore,G.M. and Gronenborn,A.M. (1988b) Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett.*, **229**, 317–324.

Olmea,O., Rost,B. and Valencia,A. (1999) Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.*, **295**, 1221–1239.

Olmea,O. and Valencia,A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold. Des.*, **2**, S25–S32.

Ortiz,A.R., Kolinski,A., Rotkiewicz,P., Ilkowski,B. and Skolnick,J. (1999) Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, **Suppl. 3**, 177–185.

Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interactions. *J. Mol. Biol.*, **271**, 511–523.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pollastri,G., Baldi,P., Fariselli,P. and Casadio,R. (2001a) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, In press.

Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2001b) Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, In press.

Selbig,J. and Argos,P. (1998) Relationships between protein sequence and structure patterns based on residue contacts. *Proteins: Struct. Funct. Genet.*, **31**, 172–185.

Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) Can three-dimensional contacts of proteins be predicted by analysis of correlated mutations? *Protein Eng.*, **7**, 349–358.

Simons,K.T., Strauss,C. and Baker,D. (2001) Prospects for ab initio protein structural genomics. *J. Mol. Biol.*, **306**, 1191–1199.

Vendruscolo,M., Kussell,E. and Domany,E. (1997) Recovery of protein structure from contact maps. *Fold. Des.*, **2**, 295–306.

Weiss,G.A., Watanabe,C.K., Zhong,A., Goddard,A. and Sidhu,S.S. (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl Acad. Sci. USA*, **97**, 8950–8954.

Westhead,D.R., Hatton,D.C. and Thornton,J.M. (1998) An atlas of protein topology cartoons available on the World-Wide Web. *TIBS*, **23**.