Machine Learning Structural and Functional Proteomics

Pierre Baldi* and **Gianluca Pollastri**
Department of Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine
Irvine, CA 92697-3425
(949) 824-5809
(949) 824-4056 (FAX)
{*gpollast,pfbaldi*}*@ics.uci.edu*

Abstract

While new high-throughput experimental techniques are being developed for proteomics applications (e.g. mass spectrometry, protein chips), it is clear that given the fundamental importance of proteins to biology, biotechnology, and medicine, computer methods that can rapidly sift through massive amounts of data and help determine the structure and function of a large number of proteins in a given genome remain important. We provide a brief overview of the application of machine learning methods to proteomic problems. In particular, we outline a novel strategy for the complete prediction of protein 3D coordinates. The strategy relies on three main successive stages: prediction of structural features, prediction of topology, and prediction of actual coordinates. We provide a progress report under this strategy and describe the corresponding suite of web servers available through http://promoter.ics.uci.edu/BRNN-PRED/. Applications to functional proteomics are briefly discussed.

Keywords: protein structure prediction, secondary structure, protein contacts, contact map, recurrent neural networks, solvent accessibility, evolutionary information.

Introduction: Proteins and Proteomics

Proteins are polymer chains made of 20 simpler building blocks, or amino acids, that function as the molecular machines of living systems. While proteins are first characterized by their primary sequences (i.e. the corresponding sequence of amino acids), they generally fold into complex three-dimensional structures that are essential for their function. Some proteins serve as structural building blocks for the cell, but the majority can be viewed as molecular "processors" that interact with each other (e.g signaling networks), with smaller molecules (e.g. metabolic networks), and with genetic information contained in DNA (e.g. regulatory networks), to form the complex circuitry of biochemical reactions associated with life. To a very first approximation, a gene codes for a protein and there are of the order of 40,000 genes in a typical mammalian cell. Each corresponding protein can exist in multiple copies, as well as different chemical variants (posttranslational modification) so that a typical mammalian cell contains about 1 billion protein molecules.

Genome and other sequencing projects are producing a data deluge of DNA and protein sequence data. In current data bases and sequencing projects, roughly 30% of proteins do not resemble any other known sequence and have no structure or functions assigned to them. Another 20% are homologous to a known sequence, for which the structure and/or functions remain also largely unknown. Genomics, the large-scale analysis of complete genomes, has its counterpart at the protein level in what has become known as proteomics (Kahn, 1995). Proteomes contain the total protein expression of a cell at a given time. Proteome analysis not only deals with determining the sequence and function of protein-encoding genes, but also is strongly concerned with the precise biochemical state of each protein in its posttranslational form.

and Department of Biological Chemistry, College of Medicine, University of California, Irvine. To whom all correspondence should be addressed.

Traditional experimental techniques for determining the structure and/or function of a protein, such as X-ray diffraction or NMR methods, remain slow and laborious and do not scale up to current sequencing speeds. Furthermore, determining the function of many proteins experimentally is a daunting task because of the complex interactions and specificity of the native environment in which a particular protein operates that may be difficult to replicate in the laboratory. While new high-throughput experimental techniques are being developed for proteomics applications (e.g. mass spectrometry, protein chips), it is clear that given the fundamental importance of proteins to biology, biotechnology, and medicine, computer methods that can rapidly sift through massive amounts of data and help determine the structure and function of say all the proteins in a given genome are sorely needed.

One of the most powerful approaches for addressing these problems is homology, i.e. to use dynamic programming alignment methods to look for evolutionarily related hence similar sequences in the data bases of known sequences. Strong sequence similarity implies similar structure and similar function. This approach works well when something is known about the structure and function of the homologue sequences, and when the degree of homology exceeds 25% identical residues. Thus alignments methods remain extremely valuable and the method of choice when they work. In about half the cases, however, they currently do not work and other methods are necessary to fill the remaining gap (Baker & Sali, 2001).

Machine Learning Structural Proteomics

Let us then focus on the structure problem, which is related but distinct from the study of the protein folding process occurring over time scales of a millisecond or less. From the outset, it is important to note that proteins can be partitioned into two classes: membrane proteins and globular proteins (Figure 1). Membrane proteins are embedded in cell membranes and therefore live in a lipid environment. In contrast, globular proteins live in aqueous environments since they are secreted from the cell, or segregated to non-membrane compartments such as the nucleus or the cytoplasm. Membrane proteins often act as receptors allowing the cell to gather information about its external environment. As such, they are often the targets of drug development efforts. In most known genomes, the typical proportion of membrane proteins is in the 20-30% range.

Because of this environmental difference, these two classes of proteins have different structural characteristics. Although membrane proteins may seem more constrained (for instance the secondary structure of known membrane domains consists of all alpha helices or, in a few cases such as porins, all beta strands) and hence simpler, they are far more difficult to crystallize.

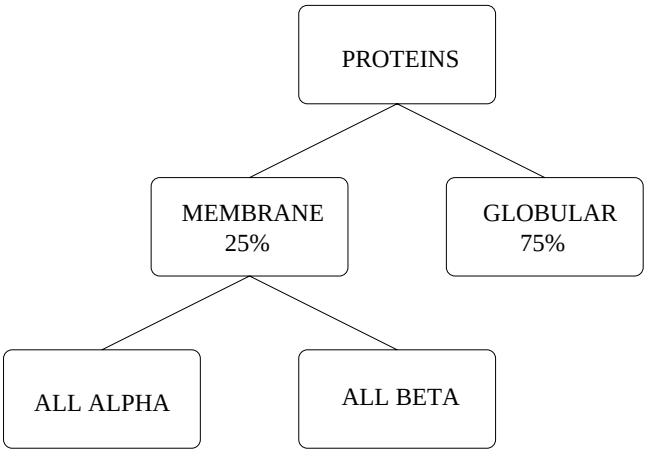

Figure 1: Proteins can be subdivided into two classes: membrane proteins and globular proteins. Membrane proteins are surrounded by membrane lipid bilayers and have peculiar structural properties. Roughly 25% of proteins in a typical genome are membrane proteins.

Hence very few membrane protein structures have been resolved and are available in the Protein Data Bank (PDB). Thus the prediction of the structure of membrane proteins per se is an important problem that remains largely unsolved. In the bioinformatics community at large, as well as in the rest of this paper, the main focus remains on the structure of globular proteins, which represent a larger fraction of all proteins and for which more data is available.

Several complementary computational approaches exist in order to predict structural features and three-dimensional structure of proteins (Sanchez & Sali, 1998; Jones, 2000) including: (1) ab initio; (2) homology modeling; (3) fold recognition; (4) "lego"; and (5) machine learning. Naturally these methods should not be viewed as exclusive of each other, but rather as complementary approaches that can be combined together in many ways. For example, the machine learning methods to be described heavily rely on the use of multiple alignments and homology.

Ab-initio approaches tackle the problem by the minimization of an energy potential derived from physicochemical, as well as statistical, considerations. The main obstacles in this approach are the derivation or approximation of the right potential and the speed up of the resulting formidable optimization problem that have prompted efforts such as IBM's BlueGene supercomputer and Stanford's protein folding@home distributed projects.

In homology modeling methods, a given protein is aligned to all its known homologues. If the 3D structure of one of the homologue sequences is known, then a structural model can be inferred for the given protein.

In fold recognition methods a similar approach is taken, but the new sequence is threaded through all the

existing folds in the protein structure data bases until an optimal match is found. Differences come not only from the alignment/threading phase but from the fact that occasionally homologous sequences have different structures, and non-homologous sequences have similar structure. [Likewise, at the functional level, occasionally proteins with the similar structure carry different functions and proteins with similar function have different structures].

It is essential to notice that the universe of fold classes for natural proteins is believed to form a finite dictionary with only a few thousand words. The PDB is the main repository of protein structures, containing over 15,000 (redundant) structures and undergoing a phase of exponential growth, like most other biological databases. Nowadays, homology modeling and fold recognition approaches share the same weaknesses when a suitable target is not found in the PDB database. In time, however, as the dictionary of structures is completed (within a decade or so), these approaches will end up providing a consistent and effective solution to the structure prediction problem, albeit perhaps not as satisfactory for some as a purely ab inito approach.

Another approach to structure prediction is the "lego" approach used by David Baker (Simons *et al.*, 2001). In the lego approach, a structural dictionary is extracted from the PDB database for small fragments of proteins of length 9 or so. A new sequence is then broken into consecutive fragments, each snippet is aligned to the dictionary and a rough structure is derived and converted via some additional massaging into a final prediction. The field of protein structure prediction has its own Olympiad, which occurs every two years in Asilomar, CA. It is the CASP (Critical Assessment of Protein Structure) competition, a worldwide blind comparison of structure predictors. At the last CASP competition in December 2000, some of the best results in 3D prediction where obtained through the lego approach (Lesk *et al.*, 2001).

Finally, there are statistical or machine learning approaches. Machine learning approaches aim at extracting information from data, more or less automatically, via a process of training from examples, a modern version of statistical model fitting. They are ideally suited for domains characterized by an abundance of data and a lack of a clear theory, which is precisely the case in bioinformatics.

Machine Learning Secondary Structure

Observation of thousands of protein structures reveals the universal presence of three kind of structural motifs: (1) alpha-helices; (2) beta-sheets; and (3) coils, two of which (α and β) are characterized by periodic patterns of hydrogen bonding that can be detected from a PDB 3D file using a program such as DSSP (Kabsch & Sander, 1983). Machine learning approaches, and neural networks in particular, have been extensively used for over 15 years to predict protein secondary structure and have consistently led to the best secondary structure predictors.

Without going through an historical summary that can be found in (Baldi & Brunak, 2001), many successful secondary structure predictors have been built using feedforward neural networks (Rost & Sander, 1994; Jones, 1999), with local input windows of 9-15 amino acids. Over the years, performance has been steadily improving, by about one percentage point per year, thanks to the increase in available training data but also the use of a number of additional techniques including: (1) output filters to clean up predictions; (2) input or output profiles (associated with alignments of homologous sequences), especially at the input level; (3) ensemble of predictors. The main weakness of these approaches probably resides in the use of a local window that cannot capture long-ranged information such as those present in beta-sheets. This is in part corroborated by the fact that the beta class is always the one with the weakest performance results. Substantially increasing the size of the input window, however, does not seem to improve performance for reasons related to overfitting and the weak signal to noise ratio associated with long-ranged interactions that play an important role but are sparse, hence hard to detect.

The methods we use to try to overcome the limitations of simple feed-forward networks have been described in (Baldi *et al.*, 1999; Baldi *et al.*, 2000a) and (Pollastri *et al.*, 2001b) and consist of BRNNs (Bidirectional Recurrent Neural Networks) with the capability of capturing at least partial long-ranged information without overfitting. These architectures are based on the probabilistic graphical model depicted in Figure 2 where inputs are transformed into outputs using both forward and backward Markov chains of hidden states. This can be viewed as a generalization of hidden Markov models by addition of the input states and the backward chain. The backward chain is predicated on the fact that biological sequences are spatial objects rather than temporal sequences. Propagation of information and learning in these graphical models is somewhat slow due to the presence of numerous undirected loops in the graph. A faster architecture is obtained by reparameterizing the graphical model using neural networks that are stationary with respect to time, leading to BRNN architectures (Figure 3).

In these general architectures for sequence translation, translation or prediction at a given position depends on a combination of local information, provided by a standard feedforward neural network, and more distant context information. More precisely, letting t denote position within a protein sequence, the overall model outputs for each t a probability vector O_t representing the membership probability of the residue at position t in each one of the three classes. This output is implemented by three normalized exponential output units. The output prediction has the functional form:

$$O_t = \eta(F_t, B_t, I_t) \qquad (1)$$

and depends on the forward (upstream) context F_t, the

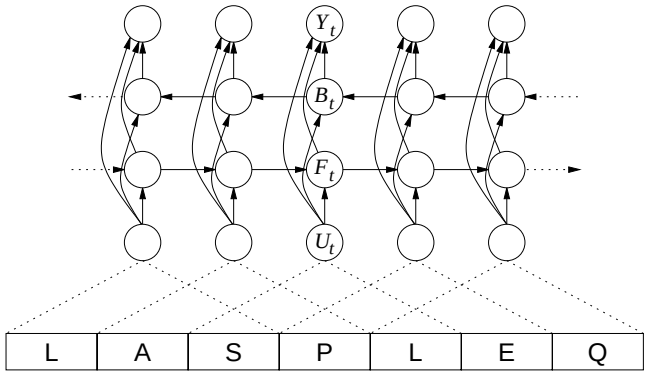

Figure 2: Bayesian network graphical model underlying BRNNs consisting of input units, output units, and both forward and backward Markov chains of hidden states.

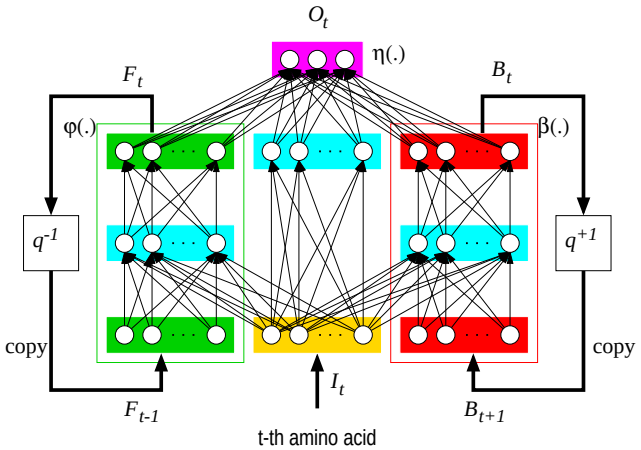

Figure 3: A BRNN architecture with a left (forward) and right (backward) context associated with two recurrent networks (wheels). Connections from input to wheels are not shown.

backward (downstream context) B_t, and the input I_t at time t. The vector $I_t \in \mathbb{R}^k$ encodes the external input at time t. In the most simple case, where the input is limited to a single amino acid, $k = 20$ by using orthogonal encoding. Larger input windows extending over several amino acids are also possible. The function η is realized by a neural network $\mathcal{N}_\eta$ (see center and top connections in Figure 3). The performance of the model can be assessed using the relative entropy between the estimated and the target distribution.

The novelty of the model is in the contextual information contained in the vectors $F_t \in \mathbb{R}^n$ and especially in $B_t \in \mathbb{R}^m$. These satisfy the recurrent bidirectional equations:

$$\begin{aligned} F_t &= \phi(F_{t-1}, I_t) \\ B_t &= \beta(B_{t+1}, I_t) \end{aligned} \qquad (2)$$

Here $\phi(\cdot)$ and $\beta(\cdot)$ are learnable non-linear state transition functions, implemented by two NNs, $\mathcal{N}_\phi$ and $\mathcal{N}_\beta$ (left and right subnetworks in Figure 3). The boundary conditions for F_t and B_t are set to 0, i.e. $F_0 = B_{T+1} = 0$ where T is the length of the protein being examined. Intuitively, we can think of F_t and B_t as "wheels" that can be rolled along the protein. To predict the class at position t, we roll the wheels in opposite directions from the N and C terminus up to position t and then combine what is read on the wheels with I_t to calculate the proper output using η. All the weights of the BRNN architecture, including the weights in the recurrent wheels, can be trained in a supervised fashion from examples by a generalized form of gradient descent or backpropagation through time, by unfolding the wheels in time, or rather space. Architectural variations can be obtained by changing the size of the input windows, the size of the window of hidden states considered to determine the output, the number of hidden layers, the number of hidden units in each layer and so forth.

This approach has resulted in the SSpro secondary structure
prediction server [http://promoter.ics.uci.edu/BRNN-PRED/] which has been ranked among the top predictors in the world both at the CASP 00 competition, and through an independent automatic evaluation server run by Burkhard Rost at Columbia University [http://cubic.bioc.columbia.edu/ eva/]. The new version of the server (SSpro 2.0) replaced the previous version (SSpro 1.0) in April 2001. The current 2.0 version uses more sensitive algorithms for the construction of input profiles and, at the time of this writing, achieves 78.1% correct classification, at the single amino acid level, using the hard CASP assignment for collapsing the eight output classes of the DSSP program into the three standard secondary structure classes. With an alternative easier assignment, also widely used in the literature, SSpro 2.0 achieves over 80% correct prediction (Pollastri *et al.*, 2001b). This performance exceeds the performance of simple feedforward neural networks trained on the same data by a few percentage points and tests reported in the references show that the wheels are indeed capable of extracting information over regions that extend beyond the traditional local window.

It is worth noting that such results are not achieved by simply training a machine learning system with raw data from the PDB. Considerable effort goes into preparing appropriate training and testing sets, using rigorous cleanup procedures that are essential to the success. The procedures involve steps such as removing chains that are too short, that have poor resolution, for which the DSSP program crashes. Even more important, these procedures must remove any sequence redundancy from the sets, since uneven sampling of the space of sequences, or high concentrations of similar structures, can introduce significant biases in the learning process. Redundancy reduction is achieved through all-against-all pairwise sequence alignments and elimination of worse quality homologues when similarity is detected (Hobohm *et al.*, 1992;

Abagyan & Batalov, 1997). Current large cleaned-up sets are about 1/10 the size of the PDB, with well over 1000 sequences. Considerable work is also required to produce suitable profiles (Rost & Sander, 1994; Baldi *et al.*, 1999; Jones, 1999; Pollastri *et al.*, 2001b).

Overall Strategy and Predictor Suite

While secondary structure plays an essential role in both folding and 3D structure and is directly implicated in a number of biological processes, it is still a far cry from the 3D structure. But could a machine learning system be extended to the prediction of 3D structure?

Training a large neural network to translate primary sequence information directly into 3D coordinates is likely to fail. Overfitting issues are compounded by the high degeneracy of the problem: rotating or translating the protein completely changes the coordinates, but leaves the structure invariant. Translation and rotation invariance must be built into the prediction learning system.

Thus our current strategy for 3D structure prediction is to decompose the problem into three steps (Figure 4). In the first step, from the primary sequence we predict a number of structural features. Typical structural features include: (a) secondary structure; (b) relative solvent accessibility, i.e. whether a given amino acid is on the surface of a protein or buried inside its hydrophobic core (Figure 5); (c) coordination or contact number, i.e. the number of neighboring amino acids of a given amino acid within a certain radius (Table 1); (d) the presence of disulphide bonds; and (e) the coupling of amino acids and strands within beta sheets (Baldi *et al.*, 2000b; Pollastri *et al.*, 2001a) or disulphide bonds.

model #	6Å	8Å	10Å	12Å
0	71.59	69.29	71.04	73.00
1	72.03	69.45	70.96	72.42
2	71.04	68.91	70.58	72.71
3	71.39	69.28	70.84	72.68
4	69.99	67.80	69.79	72.54
5	69.77	67.72	69.54	71.93
6	69.95	67.49	70.16	71.69
Ens	73.02	70.57	72.00	73.93
Comb	73.24	70.95	72.13	74.09

Table 1: Prediction of coordination number for four different radiuses (6, 8, 10, and 12Å. Percentage of threefold cross-validation results obtained with several BRNNs and the corresponding ensemble on the test set, for PSI-BLAST- based input profiles. Performance results expressed in percentages of correct prediction. Ens = ensemble of models in a given radius category. Comb = combination of 4 ensembles associated with different radius categories.

In the second step, we go from the primary sequence and the structural features to a topological representation of the protein that is invariant under rotation and translation. At a coarse level, this is the contact matrix between secondary structure elements essentially describing whether the centers of gravity of two secondary structure elements are close in the 3D structure or not. A database of coarse-level topological representations of proteins in the form of topology cartoons, called TOPS (Westhead *et al.*, 1998), is available at http://www3.ebi.ac.uk/tops/. With a higher resolution, this is the contact matrix between the individual amino acids of the protein chain.

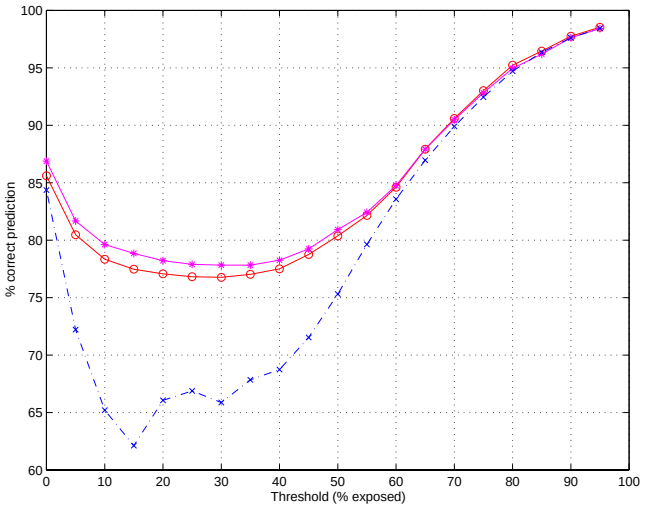

Figure 5: Performances of ACCpro for the recognition of buried/exposed aminoacids for different thresholds of relative solvent accessibility. Dashdot-crosses (blue): base-line statistical predictor (Richardson & Barlow, 1999). Solid-circles (red): test set. Solid-stars (magenta): training set. There are 20 different thresholds. For thresholds in the 10%-40%, range, where the numbers of exposed and buried residues are comparable, the ensemble of BRNNs outperforms the base-line predictor by 10% or more.

Our current approach to the problem rests on a generalization of the graphical model underlying BRNNs given in Figure 2 to process one-dimensional objects. The generalization of this architecture to two-dimensional objects, such as contact maps, is shown in Figures 6 and 7. In its basic version the Bayesian network consists of nodes regularly arranged in 6 planes: one input plane, one output plane, and 4 hidden planes.

As in the one-dimensional case, numerous variants of these ideas are possible including the use of windows in the input or output layers, the addition of connections in the hidden planes, or the use of only a subset of hidden planes rather than the full complement. Only the full complement, however, allows for the existence of a directed path from any input unit to any hidden unit. In the case of contact map prediction, relevant inputs may include the actual sequences or the pairwise statis-

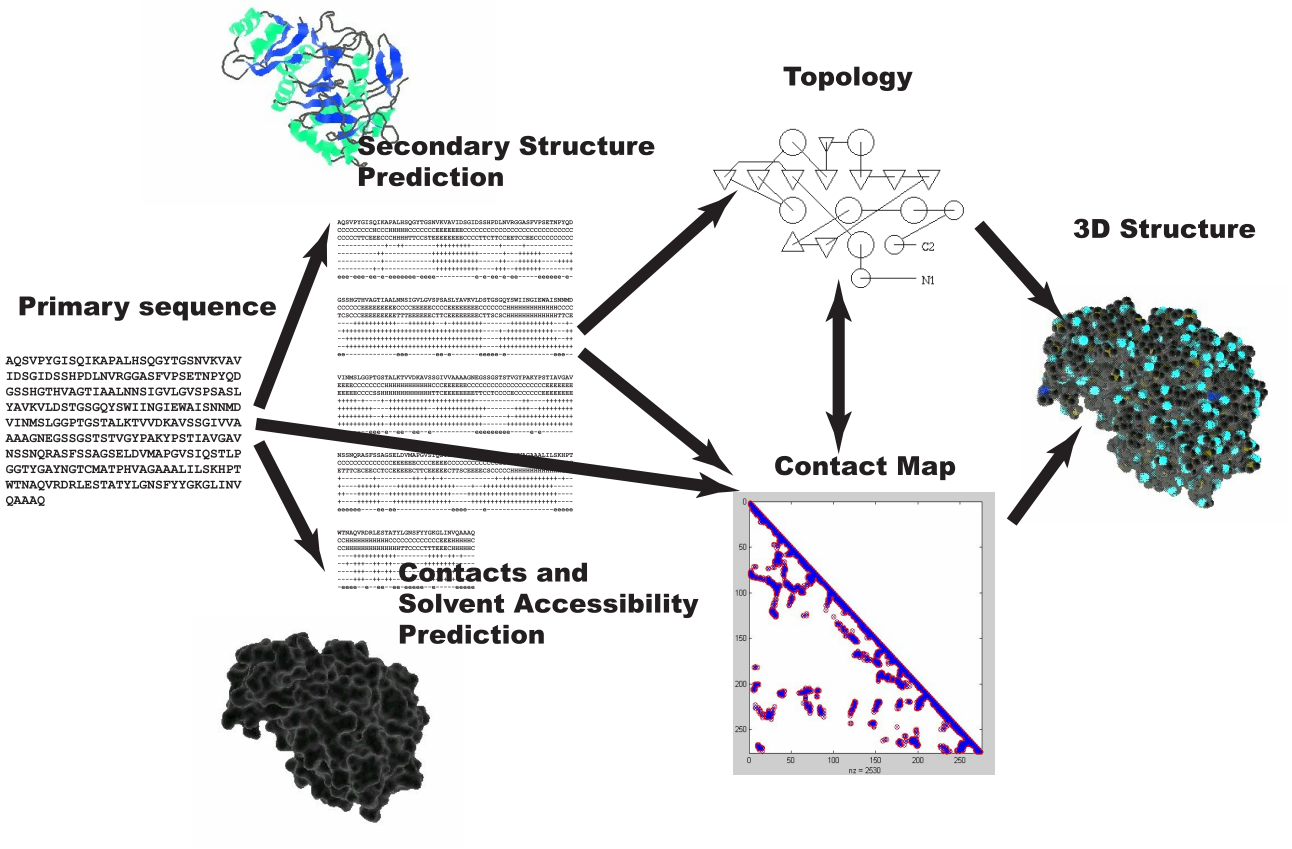

Figure 4: Overall pipeline strategy for machine learning protein structures. Example of 1SCJ (Subtilisin-Propeptide Complex) protein. The first stage corresponds to modules that predict structural features including secondary structure, contacts, and relative solvent accessibility. The second stage correspond to modules that predict 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 the atoms in the structure.

tics of the corresponding alignments to capture information about correlated mutations. Multiwise statistics could also be helpful in combination with a mechanism to control combinatorial explosion, possibly in the form of higher-order neural networks. Secondary structure and relative solvent accessibility information are also worth considering as inputs. As in the one-dimensional case, faster processing is achieved by reparameterizing the graphical models with recurrent neural networks. Note also that the graphical models introduced for the one-dimensional (Figure 2) and two-dimensional (Figures 6 and 7) cases can easily be generalized to the case of n dimensions. In three dimensions, for instance, the complete architecture requires 8 hidden planes, one for each corner of the cube. In n dimensions, the full complement requires 2^n hidden planes, one for each corner of the hypercube. While it may be possible to use the 3D version of these graphical models for protein 3D-structure prediction, here we briefly discuss an alternative approach for the last step of the strategy.

In contrast with the first two stages of the pipeline strategy that heavily rely on machine learning methods, the third step can be addressed using distance geometry and optimization techniques (Vendruscolo *et al.*, 1997) without learning. Various implementations are possible and any implementation must deal with chirality issues since, for instance, a protein and its mirror image yield the same contact map. Current algorithms seem to work well for relatively short proteins (up to 150 amino acids) but often fail to recover reasonable (within 5Å of root mean square deviation on backbone carbon atoms) 3D structures for longer proteins. A fundamental question that has not yet been addressed systematically in the literature is the amount of noise that can be tolerated in the predicted contact map without

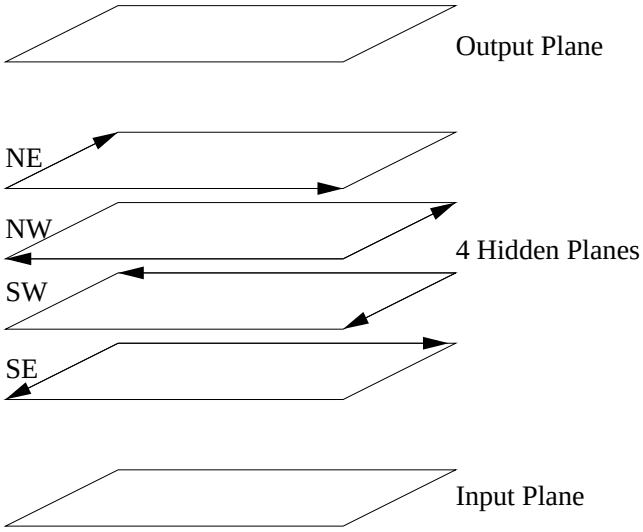

Figure 6: General layout of Bayesian network for processing two-dimensional objects such as contact maps, with units regularly arranged in one input plane, one output plane, and four planes of hidden units. All the edges of the square lattice of hidden units in each hidden plane are oriented in the direction of one of the four possible cardinal corners: NE, NW, SW, SE.

compromising the prediction of the actual coordinates. It is also important to note that in the future feedback projections could be added, if needed, for instance from the topology to the structural features.

In short, we are in the process of building a suite of structure prediction programs and servers, and combining them into a complete 3D-prediction pipeline software. At the present time, the suite contains:

- SSpro: secondary structure/server available
- SSpro8: secondary structure (in 8 categories)/server available
- ACCpro: accessibility/server available
- CONpro: contact/server available

Additional components at various stages of development include:

- DIpro: disulphide bond/under development
- BETApro: beta sheet amino acid and strand partners/partially developed
- CONTO3Dpro: from contact map to 3D coordinates/developed and available but not as a server
- 3Dpro: 3D prediction/under development

All the existing servers are available at http://promoter.ics.uci.edu/BRNN-PRED/. Users can submit a protein sequence and select the prediction categories of interest from their browser windows. Predictions are emailed to the user within a short period of time, depending on server load. At the time of this

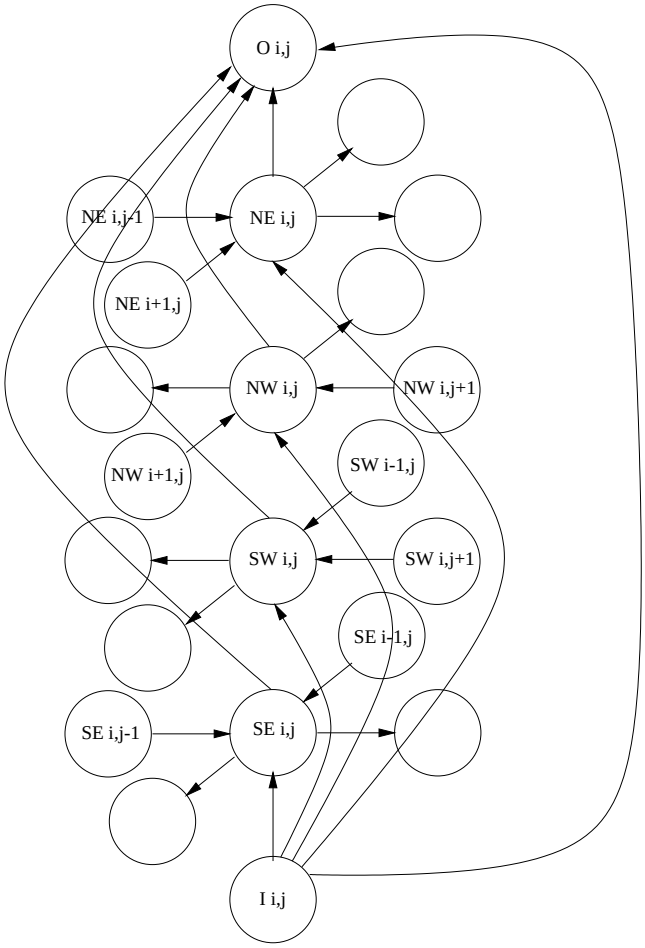

Figure 7: Details of connections within one column of Figure 6. The input unit is connected to the four hidden units, one in each hidden plane. The input unit and the hidden units are connected to the output unit. . $I_{i,j}$ is the vector of inputs at position (i, j). $O_{i,j}$ is the corresponding ouput. Connections of each hidden unit to its lattice neighbors within the same plane are also shown.

writing, the servers are averaging of the order of 100 queries per day.

It should be noticed that statistical correlations between secondary structure and contact number or accessibility are quite low and therefore it makes sense to develop separate predictors. BRNNs are used in all the machine learning architectures. Current performance on the accessibility is 77.51% (at 15% threshold). Performances for different accessibility thresholds are shown in Figure 5, against the base-line predictor which always outputs the most numerous category (Richardson & Barlow, 1999). Performance on contact prediction is 73.24% (at 6Å) or 74.09% (at 12Å) (Pollastri *et al.*, 2001a) (Table 1). In both instances, these results are better than any previously reported results,

often by several percentage points.

Thus, in summary, machine learning methods for the prediction of secondary structure and other protein structural attributes continue to improve at an average annual rate of about 1%, and are reaching good levels of performance, close to roughly 80% for secondary structure. The improvements originate both from data expansion and new algorithmic developments. We have developed new machine learning architectures and constructed a suite of structural predictors that could be integrated or combined with other methods to predict full 3D structures.

Machine Learning Functional Proteomics

From the get go, as invariably the case with biological problems, it should be clear that the boundaries of the notion of protein structure or function are somewhat fuzzy. Therefore perfect prediction in all cases cannot be expected. Furthermore, at the structural level, some protein do not fold spontaneously but may require other protein (chaperones) for proper folding. Some proteins may exist in different structural conformations, and conformations could depend on external variables such as solvent acidity. In many cases, several distinct protein chains aggregate to form so-called quaternary structures that cannot be predicted from single chains. Whether the limit horizon in the case, for instance, of secondary structure prediction is 85% or 95% is not known and for now prediction efforts should continue unabated.

The situation is even more complex when we consider the notion of protein function, which also strongly depends on the surrounding molecular context and inherently covers a number of different topics and questions, including: (1) molecular function (enzymatic catalysis, membrane transport) and analysis of conformation and active sites; (2) cellular function (inter/intracellular communication, structural, movement); (3) physiological function (organ development); (4) phenotypical function (visible effects); (5) disfunction (effect of a protein that is absent or mutated); (6) transcriptional and posttranscriptional modifications (RNA editing); (7) post-translational modifications (phosphorylation, glycosylation); (8) cellular localization (nucleus, cytoplasm, membrane, secretion). [The examples in parenthesis are not meant to be exhaustive of course and each one of them could be further expanded (e.g enzymatic catalysis could yield subcategories such as substrate, cofactors, and products).]

Here again machine learning methods, together with other experimental and computational approaches, can make valuable contributions. Consider, for instance, the case of post-translational modifications. Proteins, after they have been translated from their original DNA sequence, often undergo a large number of modifications that alter their activities. For example, certain amino acids can be linked covalently (or noncovalently) to carbohydrates, representing so-called glycosylation sites. Other amino acids are subjected to phosphorylation, where phosphate groups are added to the polypeptide chain. Kinases, for instance, are an important family of proteins involved in phosphorylation which use this process as a mean of transmitting information along many different pathways in the cell. Many other types of post-translational modifications exist, such as addition of fatty acids and the cleavage of signal peptides in the N-terminus of secretory proteins translocated across a membrane. In fact, the total number of different kinds of posttranslational modifications is in the hundreds. Knowledge of such post-translational sites is not explicitely present in genomic data but can provide important clues to function or localization and can be recovered from the primary sequence.

With the growth of data bases and available training examples, neural networks, HMMs (Hidden Markov Models) and other machine learning systems can be trained to detect, for instance, signal peptides, glycosylation sites, and phosphorylation sites (Baldi & Brunak, 2001), or to recognize specific classes of proteins, such as membrane proteins. Several servers of this kind can be found at www.cbs.dtu.dk. In many cases, the best prediction algorithm available today for any of these properties is indeed a machine learning algorithm. Again with sufficient resources an entire suite of such programs can be created and regularly updated with larger training sets. Training being done off-line, such a suite is capable of rapidly sifting through large amounts of data. While by itself it cannot answer entirely the question of the function of a new protein, it can provide valuable information regarding a large number of functional attributes. In turn, such a suite can be coupled with other information ranging from homology, to structure, to DNA microarrays and other high-throughput technologies, to literature searches.

Protein structure/function prediction is one of the central problems in bioinformatics. It is the hinge and bottleneck between sequencing efforts and drug design. Its solution should result in new enabling technologies in several areas of medicine and biotechnology. While the taxonomy of protein structure/function is complex, it can be broken down into a large but manageable number of aspects and categories. For each one of them, increasing amounts of data are rapidly becoming available in publicly repositories and data bases. This creates significant opportunities for intelligent system approaches to complement currently useful but insufficient methods, such as homology searches. Unlike conventional experimental methods, the resulting programs can rapidly sift through large amounts of data and are readily applicable to new sequences, whether naturally occurring or synthetic.

Acknowledgements

The work of PB and GP is supported by a Laurel Wilkening Faculty Innovation award and a Sun Microsystems award to PB at UCI.

References

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

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

Baldi, P. & Brunak, S. (2001). Bioinformatics: the machine learning approach. MIT Press, Cambridge, MA. Second edition.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. & Soda, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. & Soda, G. (2000a). Bidirectional dynamics for protein secondary structure prediction. In Sun, R. & Giles, C. L., (eds.) *Sequence Learning: Paradigms, Algorithms, and Applications*. Springer Verlag, New York, pp. 99–120.

Baldi, P., Pollastri, G., Andersen, C. A. F. & Brunak, S. (2000b). Matching protein β-sheet partners by feedforward and recurrent neural networks. In *Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA*. AAAI Press, Menlo Park, CA, pp. 25–36.

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

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Jones, D. T. (2000). Protein structure prediction in the postgenomic era. *Curr. Opin. Struct. Biol.*, **10**, 371–379.

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

Lesk, A. M., Conte, L. L. & Hubbard, T. J. P. (2001). Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. *Proteins*. Submitted.

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

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

Richardson, C. J. & Barlow, D. J. (1999). The bottom line for prediction of residue solvent accessibility. *Protein Engineering*, **12**, 1051–1054.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Sanchez, R. & Sali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *PNAS*, **954**, 13597–13602.

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

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

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