A Machine Learning Strategy for Protein Analysis

Pierre Baldi and Gianluca Pollastri, *Institute for Genomics and Bioinformatics, University of California, Irvine*

To sift through and analyze the massive and increasingly available data on proteins, researchers need new computing methods. The authors use machine-learning methods in a novel, three-step strategy for protein structure prediction.

Genome and other sequencing projects are producing a deluge of DNA and protein sequence data. In current databases and sequencing projects, roughly 30 percent of proteins do not resemble any other known sequence and have no assigned structure or function. Another 20 percent are homologous to a known sequence whose structure or function (or both) is largely unknown.

Proteomics is the protein counterpart to genomics, the large-scale analysis of complete genomes. Proteomes contain a cell's total protein expression at a given time. Proteome analysis not only deals with determining protein-encoding genes' sequence and function, but is also strongly concerned with the precise biochemical state of each protein in its post-translational form—that is, the form it takes after it has been translated from its original DNA encoding.

Traditional experimental techniques for determining a protein's structure and function, such as x-ray diffraction or nuclear magnetic resonance methods, remain slow and laborious, and do not scale up to current sequencing speeds. Furthermore, using experiments to determine how proteins function is a daunting task: Protein interactions are complex, and their native operating environments are very specific, which can be difficult to replicate in the laboratory.

Researchers are developing many new high-throughput experimental techniques for proteomics applications, including mass spectrometry and protein chips. Still, given proteins' fundamental importance to biology, biotechnology, and medicine, we must continue developing computer methods that can rapidly sift through massive amounts of data and help determine the structure and function of all the proteins in a given genome.

We're applying machine-learning methods to proteomic problems, and have developed a novel strategy for completely predicting protein 3D coordinates. The strategy has three stages: structural features prediction, topology prediction, and coordinate prediction. Here, we offer an overview of the domain and our machine-learning techniques, and describe the software suite we've developed, which is available at http://promoter.ics.uci.edu/BRNN-PRED.

Proteins: An overview

Proteins are polymer chains composed of 20 simpler building blocks, or amino acids, that function as the molecular machines of living organisms. Although researchers first characterize proteins by their primary sequences—that is, the corresponding amino acid sequence—proteins typically fold into complex, 3D structures that are essential to their function. Some proteins serve as structural building blocks for cells, but most are molecular "processors" that interact with

- Each other (as in signaling networks)
- Smaller molecules (as in metabolic networks)
- Genetic DNA information (as in regulatory networks)

to form life's complex circuitry of biochemical reactions.

Protein classes

As Figure 1 shows, proteins can be partitioned into two classes:

- *Membrane proteins*, which are embedded in cell membranes and therefore live in a lipid environment
- *Globular proteins*, which are secreted from the cell or segregated to nonmembrane compartments (such as the nucleus or the cytoplasm), and therefore live in aqueous environments

1094-7167/02/$17.00 © 2002 IEEE

Membrane proteins often act as receptors, letting the cell gather information about its external environment. As such, they are often the targets of drug development efforts. In most known genomes, 20 to 30 percent of the proteins are estimated to be membrane proteins.

Because of their environmental differences, the two protein classes have different structural characteristics. Although membrane proteins might seem more constrained—for example, their known secondary structure consists of all alpha helices or, in a few cases, all beta strands—and hence simpler, they are far more difficult to crystallize. Thus, very few membrane protein structures have been resolved and are available in the Protein Data Bank (PDB).

Our own work—as well as that of the larger bioinformatics community—is focused on globular protein structure, both because there is more available data and because they represent a larger fraction of all proteins. The prediction of membrane protein structure is an important problem that remains largely unsolved.

Problem scope

To a first approximation of a cell's complex biochemistry, a gene codes for a protein, and there are approximately 40,000 genes in a typical mammalian cell. Each corresponding protein can exist in multiple copies, as well as different chemical variants (post-translational modifications), and thus a typical mammalian cell contains about one billion protein molecules.

One powerful method for rapidly sifting through protein data is *homology*, which uses dynamic programming alignment methods to look for evolutionarily related (and hence similar) sequences in the databases of known sequences. Strong sequence similarity implies similar structure and function. Homology works well when something is known about a homologue sequence's structure and function, and when the homology degree exceeds 25 percent identical residues. Thus, when they work, alignment methods are extremely valuable and the method of choice. Currently, however, they don't work in roughly half the cases, and we need other methods to fill the gap.[1]

Structural proteomics

There are several complementary computational approaches for predicting a protein's structural features and 3D structure[2,3] including

- Ab initio
- Homology modeling
- Fold recognition
- Lego
- Machine learning

Ab-initio approaches minimize the energy potential derived from physico-chemical and statistical considerations. The minimization process may or may not try to mimic the folding process itself. The main obstacles in this approach are trying to derive or approximate the right potential, and the speedup of the resulting (and formidable) optimization problem. The computational obstacles have prompted efforts such as IBM's BlueGene supercomputer and Stanford's protein folding@home distributed project.

In homology modeling methods, a given protein is aligned to all its known homologues. If the 3D structure of one homologue sequence is known, then a structural model can be inferred for the given protein. Fold-recognition methods take a similar approach, but thread the new sequence through all the existing folds in the protein structure databases until an optimal match is found. Differences in the techniques come not only from the alignment/threading phase, but also from the fact that homologous sequences occasionally have different structures and nonhomologous sequences have similar ones. Likewise, at the functional level, similarly structured proteins occasionally carry different functions, and proteins with similar function have different structures.

Researchers believe that natural protein fold classes 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 provide a consistent and effective solution to the structure prediction problem, albeit perhaps not as satisfactory for some as a purely ab initio approach.

In the Lego approach,[4] researchers extract a structural dictionary from the PDB database for small protein fragments consisting of sequences of nine or so amino acids. They then break a new sequence into consecutive fragments, aligning each snippet to the dictionary and deriving a rough structure that they convert (with some additional massaging) into a final prediction. At the 2000 Critical Assessment of techniques for Protein Structure Prediction competition (CASP4)—an annual, international blind comparison of structure predictors—the Lego approach produced some of the best results in 3D prediction.[5]

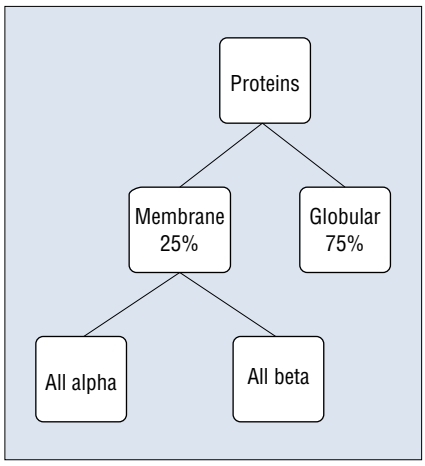

Figure 1. Proteins can be subdivided into two classes: membrane proteins and globular proteins. Membrane proteins are surrounded by membrane lipid bilayers and have peculiar structural properties. Roughly 25 percent of proteins in a typical genome are membrane proteins.

Finally, there are statistical or machine-learning approaches. Machine-learning approaches aim to extract information from data—more or less automatically—through a process of training from examples. Basically, it is a modern version of statistical model fitting. Such methods are ideally suited for domains with an abundance of data and a lack of a clear theory, which is precisely the case in bioinformatics.

All of these prediction methods are complementary, rather than exclusive, and can be combined in many ways. Our machine-learning methods, for example, rely heavily on multiple alignments and homology.

Machine-learning and secondary-structure prediction

Researchers' observations of thousands of protein structures have revealed the universal presence of three structural motifs: coils, alpha helices, and beta sheets; the latter two are characterized by periodic hydrogen-bonding patterns that researchers can detect using a PDB 3D file with a program such as Define Secondary Structure of Pro-

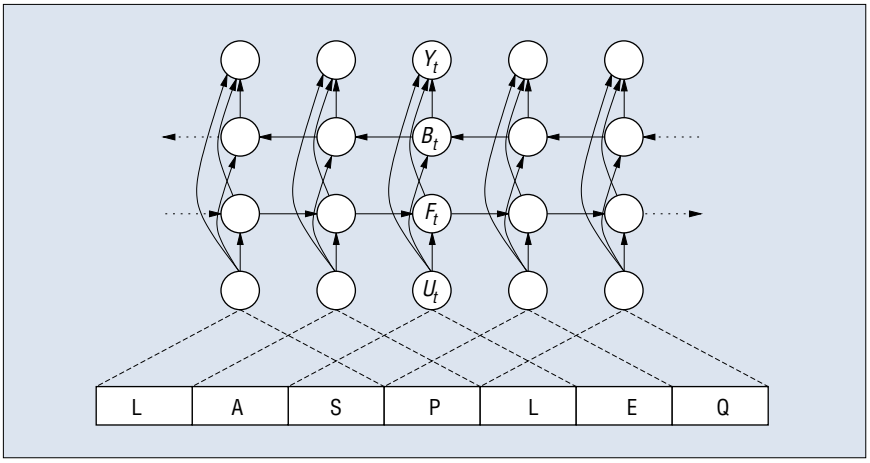

Figure 2. Bayesian network graphical model. The model underlies bidirectional, recurrent neural networks and consists of input units, output units, and both forward and backward Markov chains of hidden states.

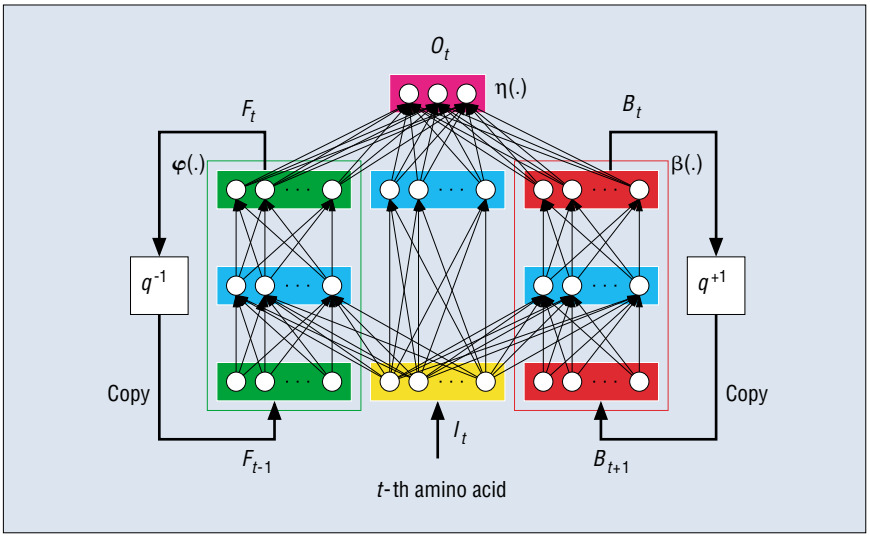

Figure 3. A bidirectional, recurrent neural-network architecture. It has a forward context (left side) and a backward context (right side) that are associated with two recurrent networks, which we think of as "wheels" that roll along the protein.

teins (DSSP).[6] For more than 15 years, researchers have used machine-learning approaches—particularly neural networks—to predict proteins' secondary structure, and have consistently had the best secondary-structure predictions.

State of the practice

As historical summaries show,[7] many researchers have built successful secondary structure predictors using feed-forward neural networks with local input windows of nine to 15 amino acids.[8,9] Over the years, performance has steadily improved by about one percent per year, thanks to increased training data and several additional techniques, including

- Output filters to clean up predictions
- Input or output profiles—associated with homologous sequence alignments—especially at the input level
- Predictor ensembles

The main weakness of these approaches likely resides in researchers' use of a local window that cannot capture long-ranged information, such as that present in beta sheets. This is partially corroborated because the beta sheet class always has the weakest performance results. Substantially increasing the input window's size, however, does not seem to improve performance. The reason is related to overfitting and the weak signal-to-noise ratio associated with long-ranged interactions; the latter play an important role, but are sparse and therefore hard to detect.

We've described our methods for trying to overcome the limitations of simple feed-forward networks elsewhere.[10–12] Basically, they consist of bidirectional, recurrent neural networks (BRNNs) capable of capturing at least partial long-ranged information without overfitting. As Figure 2 shows, we base these architectures on a probabilistic graphical model, in which inputs are transformed into outputs using both forward and backward Markov chains of hidden states.

In one sense, this is a generalization of hidden Markov models (HMMs), accomplished by adding 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. However, the models' information and learning propagation is somewhat slow owing to numerous undirected graph loops. We obtain a faster architecture by reparameterizing the graphical model using neural networks that are stationary with respect to time, thereby creating a BRNN architecture (see Figure 3).

BRNN architectures

In these general sequence-translation architectures, a translation or prediction at a given position depends on a combination of local information, provided by a standard feed-forward neural network and more distant context information. More precisely, letting t denote position within a protein sequence, the overall model outputs a probability vector O_t for each t, which represents the residue's membership probability at position t in each of the three classes. Three normalized exponential output units implement the output. The output prediction's functional form is

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

and depends on the forward (upstream) context F_t, the 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 simplest case, where 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 $\mathcal{N}_\eta$ (the center and top connections in Figure 3). We assess the model's perfor-

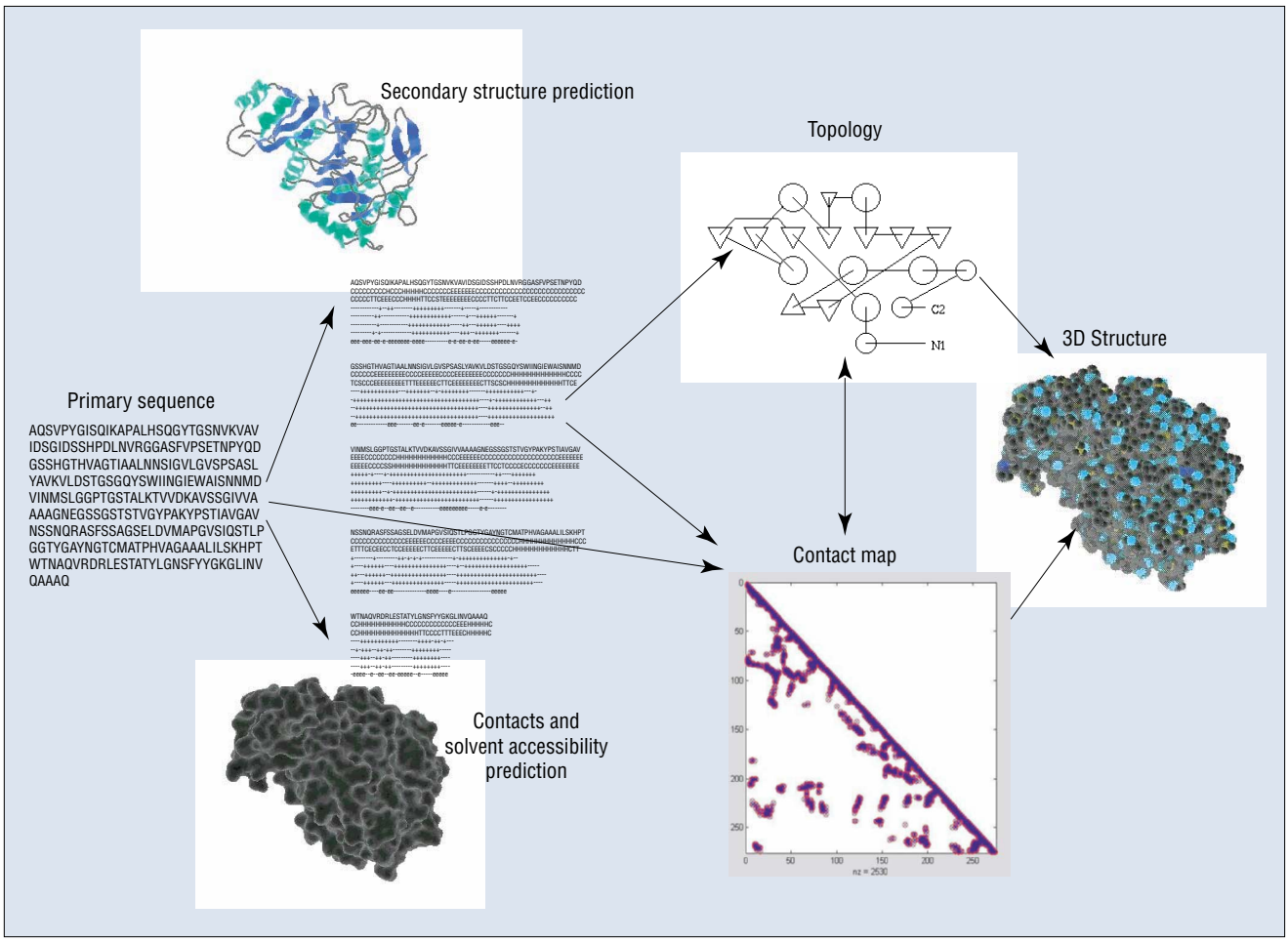

Figure 4. Overall pipeline strategy for machine-learning protein structures. The example here is the Subtilisin-Propeptide Complex (1SCJ) protein. In the first state, modules predict structural features including secondary structure, contacts, and relative solvent accessibility. In the second stage, modules predict the protein's topology, 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 contact map between the protein residues represents the high-resolution topology. In the final stage, our strategy predicts the actual 3D coordinates of all the structure's atoms.

mance using the relative entropy between the estimated and target distribution.

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

$$F_t = \phi(F_{t-1}, I_t)$$
$$B_t = \beta(B_{t+1}, I_t).$$

Here, $\phi(\cdot)$ and $\beta(\cdot)$ are learnable, nonlinear state transition functions, implemented by two neural networks, N_ϕ and N_β (the left and right subnetworks in Figure 3). The boundary conditions for F_t and B_t are set to 0—that is, $F_0 = B_{T+1} = 0$, where T is the length of the examined protein. Intuitively, we can think of F_t and B_t as "wheels" that we roll along the protein. To predict the position t class, we roll the wheels in opposite directions from the N and C terminus up to position t, and then combine what we read on the wheels with I_t to calculate the proper output using η. We train all the BRNN architecture weights, including those in the recurrent wheels, in a supervised fashion using examples extracted from the PDB and a generalized form of gradient descent or backpropagation through time—that is, by unfolding the wheels in time, or rather, space. To achieve architectural variations, we change such things as the input windows' size, the window size of hidden states that determine the output, the number of hidden layers, and the number of hidden units in each layer.

Predicting 3D structure: The pipeline strategy

While secondary structure plays an essential role in both folding and 3D structure and is directly implicated in several biological processes, it is still a far cry from the 3D structure. But could a machine-learning system be extended to predict a 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 problem's high degeneracy: 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 (see Figure 4).

Step 1: Structure prediction

We first predict several of the primary

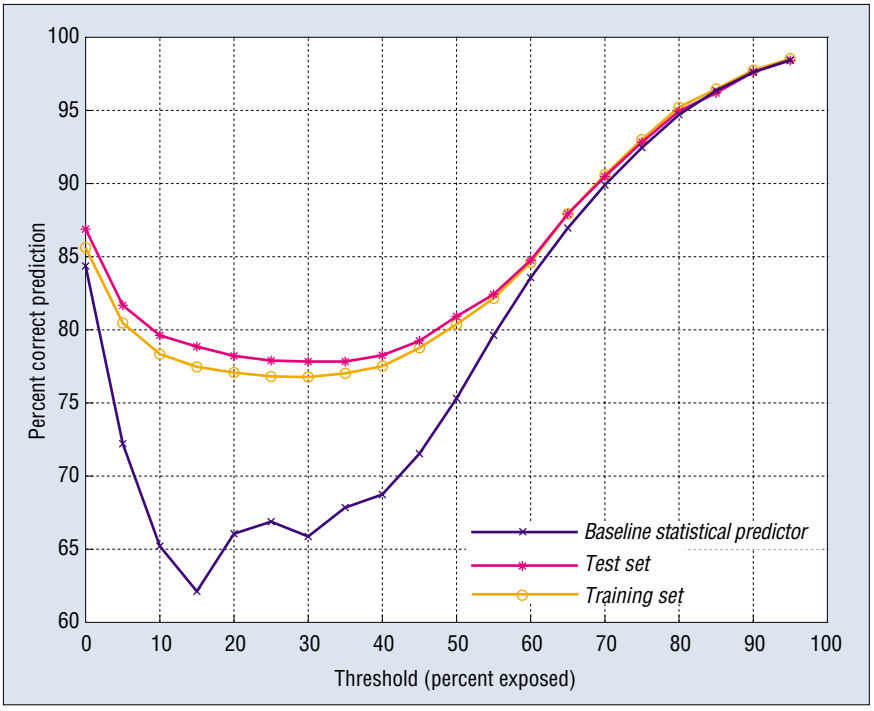

Figure 5. Performances of ACCpro for the recognition of buried and exposed amino acids for 20 different thresholds of relative solvent accessibility. For thresholds in the 10- to 40-percent range, given a comparable number of exposed and buried residues, the BRNNs ensemble outperforms the baseline predictor by 10 percent or more.

sequence's structural features. Typical structural features include

- Secondary structure
- Relative solvent accessibility (whether a given amino acid is on a protein's surface or buried inside its hydrophobic core; see Figure 5)
- Coordination or contact number (the number of a given amino acid's neighboring amino acids within a certain radius; see Table 1)
- Disulphide bonds
- Amino acids coupled by beta sheet strands[13,14] or by disulphide bonds

Step 2: Topology prediction

Next, we move from the protein's primary sequence and structural features to a topological representation, which is invariant under rotation and translation. At a coarse level, this is the contact matrix between secondary structure elements, which in its simplest form describes whether the gravity centers of two secondary structure elements in the 3D structure are close. A database of coarse-level topological representations, in cartoon form, called TOPS,[15] is available at www3.ebi.ac.uk/tops. With a higher resolution, this is the contact matrix between the protein chain's individual amino acids.

Our current approach to the problem rests on a generalization of BRNNs' underlying graphical model, which processes 1D objects (see Figure 2). Figures 6 and 7 show this architecture's generalization to 2D objects, such as contact maps. In its basic version, the Bayesian network consists of nodes regularly arranged in six planes: one input plane, one output plane, and four hidden planes.

As in the 1D case, numerous variants of these ideas are possible, including

- Using input or output layer windows
- Adding connections in the hidden planes
- Using only a subset of hidden planes, rather than the full complement

Only the full complement, however, allows a directed path from any input unit to any hidden unit. With contact map prediction, relevant inputs can include the actual sequences and the corresponding profiles, or the corresponding alignment's pairwise statistics to capture information about correlated mutations. Higher-order statistics could also be helpful if combined with a mechanism to control combinatorial explosion (possibly using higher-order neural networks). We also consider secondary structure and relative solvent accessibility information as inputs. As in the one-dimensional case, we achieve faster processing by reparameterizing the graphical models with recurrent neural networks.

The graphical models introduced for Figure 2's 1D case and Figure 6's 2D case can easily be generalized to the case of n dimensions. In 3D, for example, the complete architecture requires eight 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 might be possible to use the graphical models' 3D version for protein 3D-structure prediction, here we briefly discuss an alternative approach for the strategy's last step.

Table 1. Correct prediction percentages for coordinating numbers at four distance cutoffs using seven different BRNNs and their combinations.

Model	Radiuses			
	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
Model ensemble	73.02	70.57	72.00	73.93
All four ensembles	73.24	70.95	72.13	74.09

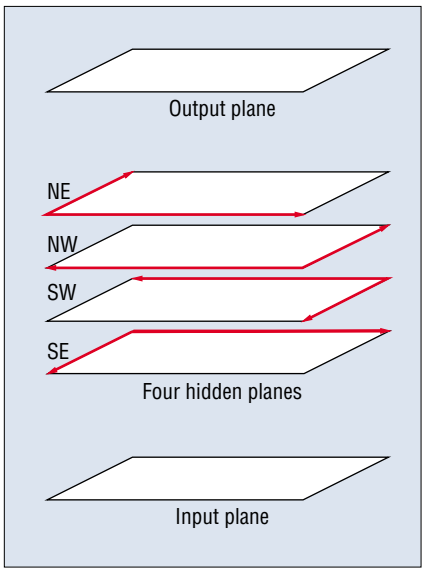

Figure 6. General layout of a Bayesian network for processing 2D objects. Units are regularly arranged in one input plane, one output plane, and four planes of hidden units. The square lattice edges in each hidden plane are oriented toward one of the four possible cardinal corners: NE, NW, SW, SE.

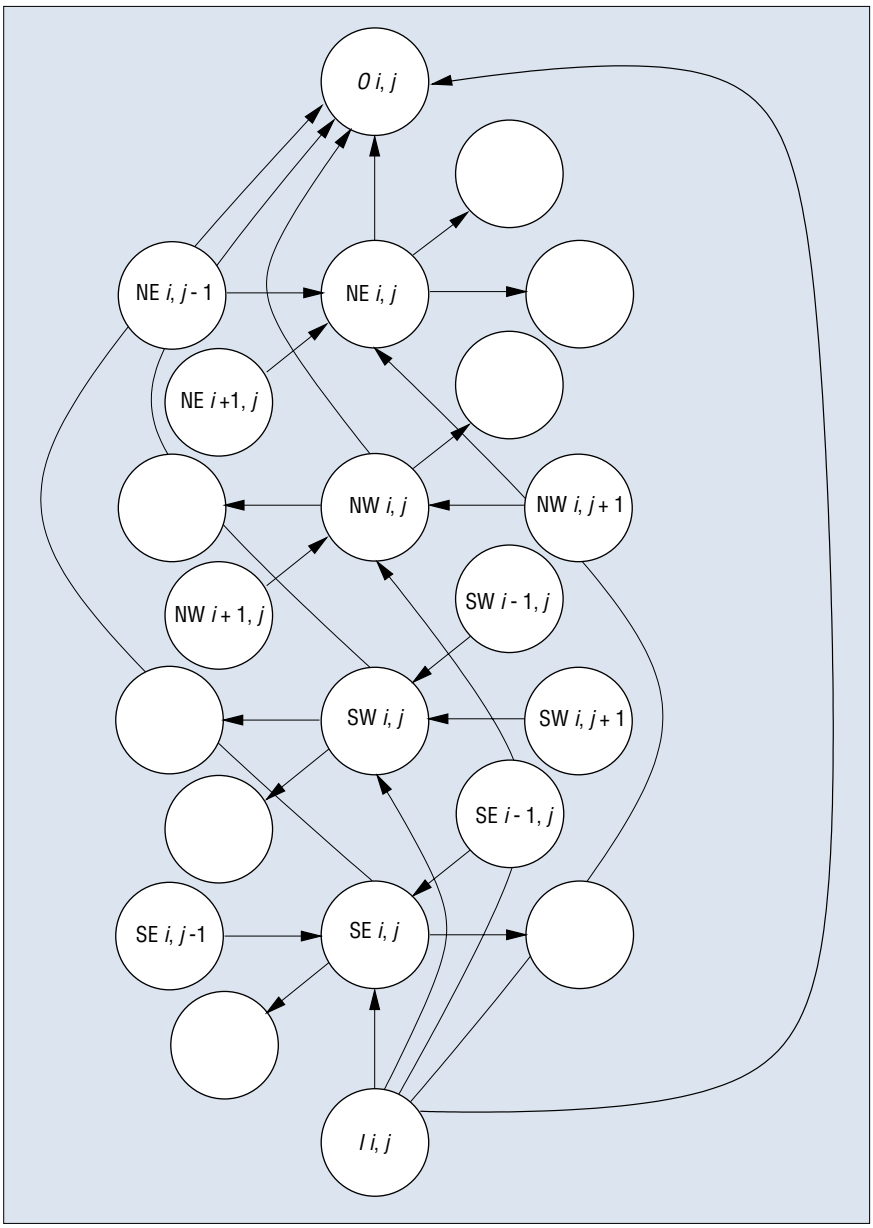

Figure 7. Connection details within one of Figure 6's columns. The input unit is connected to the four hidden units, one in each hidden plane. The input unit and the hidden units are connected to the output unit. $I_{i,j}$ is the input vector at position (i,j). $O_{i,j}$ is the corresponding output. The figure also shows each hidden unit's connection to its lattice neighbors within the same plane.

Step 3: 3D-structure prediction

In contrast with our strategy's first two stages, which heavily rely on machine-learning methods, we can address the third stage using distance geometry and optimization techniques[16] without learning. Various implementations are possible; any must deal with chirality issues, because, for example, 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 for longer proteins, they often fail to recover reasonable 3D structures (within 5Å of root mean square deviation on backbone carbon atoms).

A fundamental question that the literature has yet to systematically address is the amount of noise that the predicted contact map can tolerate without compromising coordinate prediction. Also, if necessary in the future, we could add feedback projections, such as from the topology to the structural features.

Project status and results

We are using our pipeline strategy to build a suite of structure prediction programs and servers, and combining them into a complete 3D-prediction pipeline software package.

The state of the package

Our suite now contains

- SSpro, a secondary-structure predictor (three categories)
- SSpro8, a secondary-structure predictor (eight categories)
- ACCpro, an accessibility predictor
- CONpro, a contact predictor

Each of these has an available server at http://promoter.ics.uci.edu/BRNN-PRED. Users can submit a protein sequence and select the prediction categories from their browser windows. We email predictions to users as soon as possible, depending on server load. Currently, the servers are averaging around 250 queries a day.

We also have several additional compo-

nents in development, including

- Dipro, for predicting disulphide bonds
- BETApro, for predicting beta sheet amino acid and strand partners
- COMApro, for predicting contact maps
- COMATO3Dpro, for predicting 3D coordinates from contact maps
- 3Dpro, for 3D prediction

We've completed development of COMApro, COMATO3Dpro, and 3Dpro, and expect them to be online for the 2002 CASP5 experiment.

Performance

The SSpro secondary-structure-prediction server was ranked among the world's top predictors both at CASP4 and by Columbia University's Burkhard Rost, who ran an independent Automatic evaluation of structure-prediction servers (http://cubic.bioc.columbia.edu/eval).

We released SSpro 2.0 in April 2001 and expect to release SSpro 3.0 in time for the CASP5 experiment. The new version uses more sensitive algorithms for constructing input profiles and currently achieves 78.1 percent correct classification at the single amino acid level, using the hard CASP assignment for collapsing DSSP's eight output classes into the three standard secondary-structure classes. Given an alternative, easier assignment (but also widely used in the literature), SSpro 2.0's correct prediction rate is more than 80 percent.[12] This performance exceeds that of simple feed-forward neural networks trained on the same data by a few percentage points. Tests also show that the wheels are indeed capable of extracting information over regions that extend beyond the traditional local window.[10]

We did not achieve such results by simply training a machine-learning system on raw PDB data. We put considerable effort into preparing appropriate training and testing sets, using rigorous cleanup procedures that are essential to success. The procedures involve removing chains that, for example, are too short, have poor resolution, or cause DSSP to crash. Even more important, these procedures must remove any sequence redundancy from the sets, since uneven space-sequence sampling or a high concentration of similar structures can introduce significant learning-process biases. We achieve redundancy reduction by using all-against-all pairwise sequence alignments and eliminating the lower-quality homologues when similarity is detected.[12] Currently, large cleaned-up sets are about one-fifth the size of the PDB, with well over 3,000 sequences. We also put considerable work into producing suitable profiles.[8–10,12]

Statistical correlations between secondary structure and contact number or accessibility are quite low, and it therefore makes sense to develop separate predictors. We use BRNNs in all the corresponding machine-learning architectures. Current accessibility performance is 77.51 percent (at 15 percent threshold). Figure 5 shows performances for different accessibility thresholds, against the baseline predictor, which outputs the most numerous categories.[17] Contact prediction performance is 73.24 percent (at 6Å) or 74.09 percent (at 12Å;[14] see Table 1). In both cases, the results are better than any previously reported, often by several percentage points.

> The SSpro secondary-structure-prediction server was ranked among the world's top predictors both at the CASP4 competition and by Columbia University's Burkhard Rost.

In general, machine-learning methods for predicting proteins' secondary structure and other attributes continue to improve, at an average annual rate of about 1 percent. They are also reaching good performance levels—close to 80 percent for secondary structure. Such improvements originate both from data expansion and new algorithmic developments.

Outstanding issues

As is invariably the case with biological problems, the notions of protein structure and function have fuzzy boundaries. Therefore, we can't expect perfect prediction in all cases. At the structural level, some proteins do not fold spontaneously and require other proteins (chaperones) for proper folding. Furthermore, some proteins exist in different structural conformations, and conformations can 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. We don't yet know whether the limit horizon of secondary structure prediction, for example, is 85 or 95 percent. For now, prediction efforts should continue unabated.

At the protein function level, the situation is even more complex. Function strongly depends on the surrounding molecular context and inherently covers many different topics and questions, including

- Molecular function (such as enzymatic catalysis and membrane transport) and conformation and active site analysis
- Cellular function (such as inter/intracellular communication, structure, and movement)
- Physiological function (such as organ development)
- Phenotypical function (such as visible effects)
- Disfunction (such as the effect of absent or mutated protein)
- Transcriptional and posttranscriptional modifications (such as RNA editing)
- Posttranslational modifications (such as phosphorylation and glycosylation)
- Cellular localization (such as nuclear, cytoplasmic, membrane, or secreted)

Here again, machine-learning methods, together with other experimental and computational approaches, can make valuable contributions. Consider, for example, posttranslational modifications. Once translated from their original DNA sequence, proteins often undergo numerous 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 example, are an important family of proteins involved in phosphorylation that use this process as a mean of transmitting information along many different cellular pathways.

Many other types of posttranslational modifications exist, such as fatty-acid additions and signal peptide cleavage in the N-terminus of secretory proteins translocated across a membrane. In fact, there are several hundred kinds of posttranslational modifications. Genomic data does not explicitly present knowledge of such posttranslational sites, but it can provide important clues to function or localization, and we

can recover it from the primary sequence.

With the growth of databases and available training examples, we can train neural networks, HMMs, and other machine-learning systems to detect signal peptides, glycosylation sites, phosphorylation sites, and so on,[7] or to recognize specific protein classes, such as membrane proteins (see, for example, the servers at www.cbs.dtu.dk or the HMM programs at www.netid.com or http://hmmer.wustl.edu).

Again, with sufficient resources, a small team of researchers can create an entire suite of such programs and regularly update them with larger training sets. With offline training, such a suite can rapidly sift through large volumes of data.

Although machine-learning methods today cannot by themselves entirely describe a new protein's function, they can provide valuable information regarding numerous functional attributes. In turn, you can couple such a suite with other information, including that from homology, structure, DNA microarrays and other high-throughput technologies, and literature searches.

Predicting proteins' structure and function is a central problem in bioinformatics. It is the hinge and bottleneck between sequencing efforts and drug design. Solving this problem should result in new enabling technologies in medicine and biotechnology. Although the protein structure and function taxonomy is complex, we can break it down into manageable aspects and categories. For each of them, researchers are rapidly producing increasing amounts of data and making them publicly available in repositories and databases. This creates significant opportunities for intelligent system approaches to complement 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 natural or synthetic sequences. ■

Acknowledgments

Our work is supported by grants from the National Institute of Health, a Laurel Wilkening Faculty Innovation award, and a Sun Microsystems award to Pierre Baldi.

References

1. D. Baker and A. Sali, "Protein Structure Prediction and Structural Genomics," *Science*, vol. 294, no. 5540, 2001, pp. 93–96.
2. R. Sanchez and A. Sali, "Large-Scale Protein Structure Modeling of the Saccharomyces Cerevisiae Genome," *Proc. Nat'l Academy of Science,* vol. 954, no. 23, 10 Nov. 1998, pp. 13597–13602.
3. D.T. Jones, "Protein Structure Prediction in the Postgenomic Era," *Current Opinion in Structural Biology*, vol. 10, no. 3, June 2000, pp. 371–379.
4. K.T. Simons, C. Strauss, and D. Baker, "Prospects for Ab Initio Protein Structural Genomics," *J. Molecular Biology*, vol. 306, no. 5, 2001, pp. 1191–1199.
5. A.M. Lesk, L.L. Conte, and T.J.P. Hubbard, "Assessment of Novel Fold Targets in CASP4: Predictions of Three-Dimensional Structures, Secondary Structures, and Interresidue Contacts," *Proteins*, vol. 45, no. 5, 2001, pp. 98–118.
6. W. Kabsch and C. Sander, "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features," *Biopolymers*, vol. 22, no. 12, Dec. 1983, pp. 2577–2637.
7. P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed., MIT Press, Cambridge, Mass., 2001.
8. B. Rost and C. Sander, "Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure," *Proteins*, vol. 19, no. 1, May 1994, pp. 55–72.
9. D.T. Jones, "Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices," *J. Molecular Biology*, vol. 292, no. 2, 1999, pp. 195–202.
10. P. Baldi et al., "Exploiting the Past and the Future in Protein Secondary Structure Prediction," *Bioinformatics*, vol. 15, no. 11, 1999, pp. 937–946.
11. P. Baldi et al., "Bidirectional Dynamics for Protein Secondary Structure Prediction," in R. Sun and C.L. Giles, eds., *Sequence Learning: Paradigms, Algorithms, and Applications*, Springer Verlag, New York, 2000, pp. 99–120.
12. G. Pollastri et al., "Improving the Prediction of Protein Secondary Strucure in Three and Eight Classes using Recurrent Neural Networks and Profiles," *Proteins*, vol. 47, 2002, pp. 228–235.
13. P. Baldi et al., "Matching Protein-Sheet Partners by Feed-Forward and Recurrent Neural Networks," *Proc. 2000 Conf. on Intelligent Systems for Molecular Biology (ISMB 00)*, AAAI Press, Menlo Park, Calif., 2000, pp. 25–36.
14. G. Pollastri et al., "Prediction of Coordination Number and Relative Solvent Accessibility in Proteins," *Proteins*, vol. 47, 2002, pp. 142–153.
15. D.R. Westhead, D.C. Hatton, and J.M. Thornton, "An Atlas of Protein Topology Cartoons Available on the World Wide Web," *Trends in Biochemical Sciences*, vol. 23, no. 1, 1998, pp. 35–36.
16. M. Vendruscolo, E. Kussell, and E. Domany, "Recovery of Protein Structure from Contact Maps," *Folding and Design*, vol. 2, no. 5, 1997, pp. 295–306.
17. J. Richardson and D.J. Barlow, "The Bottom Line for Prediction of Residue Solvent Accessibility," *Protein Engineering*, vol. 12, no. 12, 1999, pp. 1051–1054.

The Authors

Pierre Baldi is a professor in the Department of Information and Computer Science and the department of Biological Chemistry at the University of California, Irvine, where he is director of the Institute for Genomics and Bioinformatics. His main research interests are in bioinformatics/computational biology and machine learning. He is the author of more than 100 scientific articles and several books, including *Bioinformatics: The Machine Learning Approach* (MIT Press), *The Shattered-Self—The End of Natural Evolution* (MIT Press), and *DNA Microarrays and Gene Regulation* (Cambridge University Press). He received an MS in mathematics and psychology from the University of Paris and a PhD from Caltech. Contact him at pfbaldi@ics.uci.edu; www.ics.uci.edu/~pfbaldi.

Gianluca Pollastri is a PhD candidate in computer science at the University of California, Irvine, where he is a member of the Institute for Genomics and Bioinformatics. His research focuses on machine-learning methods and protein structure prediction. He received an MSc in telecommunication engineering from the University of Florence, Italy. Contact him at gpollast@ics.uci.edu.