The Handbook of
Brain Theory
and Neural Networks

EDITED BY
Michael A. Arbib

EDITORIAL ADVISORY BOARD
George Adelman • Shun-ichi Amari • James A. Anderson
John A. Burgden • Andrew G. Barto • Françoise Fogelman-Soulie
Stephen Grossberg • John Hertz • Marc Jeannerod • B. Keith Jenkins
Mitsuo Kawato • Christof Koch • Eve Marder • James L. McClelland
Terrence J. Sejnowski • Harold Szu • Gerard Toulouse
Christoph von der Malsburg • Bernard Widrow

EDITORIAL ASSISTANT
Prudence H. Arbib

A Bradford Book
THE MIT PRESS
Cambridge, Massachusetts
London, England
Protein Structure Prediction

Burkhard Rost and Chris Sander

Introduction

What is a protein? The information for life is stored by a four-letter alphabet in the genes. Proteins perform most important tasks in organisms, such as catalysis of biochemical reactions, transport of nutrients, recognition and transmission of signals. Proteins are formed from joining amino acids into a long, stretched chain, the protein sequence. Proteins differ in the number (from 30 to 30,000) and in the arrangement of the amino acids (called residues, when joined in proteins). In water, the chain folds up to a unique three-dimensional (3D) structure. The main driving force is the need to pack residues for which a contact with water is energetically unfavorable into the interior of the molecule. This is only possible if the protein forms regular patterns of a macroscopic substructure called secondary structure (Figure 1; see Branden and Tooze, 1991).

What determines protein function and structure? The 3D structure of a protein determines its function. The 3D structure is uniquely determined by the sequence. Can the code be deciphered—i.e., can 3D structure be predicted from sequence? In principle, yes, but the computer time required to predict 3D structure from first principles is many orders of magnitude beyond today’s possibilities. However, one reason to want to know the structure is rational drug design.

Why not simply look by microscope at the 3D structure? The techniques to experimentally determine 3D structure of a protein are rather complicated. Today, the sequence is known for some 36,000 proteins, but only for 2000 has the 3D structure been determined by experiment. Large gene sequencing projects increase the sequence-structure gap further. The most accurate way to predict 3D structure from sequence is by homology modeling—i.e., search for a protein with similar sequence that has a known 3D structure and then model the 3D structure of the unknown protein in analogy to the known one. Such techniques lead to a reduction of the sequence-structure gap by some 3000 proteins.

Why can homology modeling be successful? The exchange of a few residues can already destabilize a protein. This implies that the majority of the 20⁴ possible sequences of length N form different structures. But has evolution created such an immense variety? The evolutionary pressure to conserve function and to maintain the properties of the universe of structures has the result that structure is more conserved than sequences. Evolution has produced pairs of proteins which have the same 3D structure with only 25% identical residues. For such pairs, 3D structure can be predicted rather accurately by homology.

Can the egg be unboiled? When an egg is boiled, the protein it contains unfold. Can this procedure be reversed in theory? Or, can the encrypted code of protein folding be deciphered from sequence? Current tools to predict 3D structure from sequence are rather limited (Rost and Sander, 1994b). The problem has to be simplified. One extreme simplification is to predict one-dimensional (1D) strings of secondary structure assignment (Figure 1).

How can neural networks predict protein structure? In practice, the most successful predictions are based on an analysis of common features in the data bank of known 3D structures. Artificial neural networks are well suited for pattern classification. Here, we shall attempt to show how neural networks can be used to predict protein structure. First, we give examples of
Figure 1. Hierarchy of protein structure. 1D: The amino acid sequence determines the formation of 3D structure. Here, the chain of ubiquitin (Ubq) is shown. The 3D structure can be projected onto a 1D diagram, e.g., the secondary structure (H, α-helix; E, extended; β-strand, and loop). 2D: The 3D structure can be projected onto a 2D matrix, e.g., the residue window of the matrix gives the contact between residue i and residue j (plot by Conner M. Scheu, 1999), which the mapping of sequence window to secondary structure is known as well. Sufficient testing is crucial (Rost and Sander, 1994a).

Prediction of Secondary Structure

Presence of the Protein to the Network

The usual goal of secondary structure prediction methods is to classify a pattern of adjacent residues as either H (α-helix), E (extended) or L (for loop = all others). Sequences are translated into patterns by shifting a window of adjacent residues through the protein and looking at the secondary structure for the central residue (Figure 2). Networks used for secondary structure prediction are multi-layer feedforward networks (Figure 2). The network error is given by the difference between actual network output (uniquely determined by the choice of connections) and desired output (looked up from data bank). Training of learning means changing the connections such that the error decreases for the given examples (gradient descent; see, e.g., BACKPROPAGATION: Basics and New Developments and Learning as Hill-Climbing in Weight Space). If training is successful, the patterns are correctly classified. But how can new patterns be classified correctly? The hope is that the network extracts general rules by the classification of the training patterns. The generalization ability is checked by another set of test samples for which the mapping of sequence window to secondary structure is known as well. Sufficient testing is crucial (Rost and Sander, 1994a).

Prediction Performance of Simple Neural Networks

Networks of the type described reach values for three-state overall prediction accuracy of around 60%. This is comparable to the performance of non-network methods. In the past, the first application of neural networks to the prediction of secondary structure (Quast et al., 1990), more than 20 groups have followed (Hirt et al., 1992; Rost and Sander, 1994a). Prediction accuracy was not improved significantly without using biological expertise, as we shall see in the next section.

Using Evolutionary Information of Multiple Sequences

Some residues can be replaced by others without changing the structure. But not every amino acid can be replaced by any other. On the contrary, the residue substituting patterns are very specific for a certain 3D structure. Can this information be used to improve the prediction accuracy for neural networks? Indeed, using evolutionary information as derived from a database of proteins with homologous 3D structure improves the performance accuracy by about 10 percentage points to > 72% (Rost and Sander, 1994a). The basic procedure is as follows: First, sequences of proteins which are similar enough in se-
sequence to know that they are similar as well in 3D structure are aligned (optimal fit). Second, for each residue position, it is counted how often any of the 20 amino acids occur in the alignment at that position. Third, the counts are used as input to a network.

Prediction of Structural or Functional Protein Class

According to the relative content in secondary structure, proteins can be classified into structural classes. From the predicted secondary structure, the relative content of secondary structure of a protein can be calculated to predict structural class (Roost and Sandor, 1994). An alternative approach is to directly predict the secondary structure content of a protein by a neural network that uses as input a vector of 20 components giving the frequency with which each amino acid occurs in a particular protein (Mukal and Kim, 1992).

A different task is to predict that two proteins are similar in terms of function or 3D structure. Methods have been based (1) on multiple feedforward networks (Fratzke and Argue, 1992), using proteins of similar sequences as input; (2) on simple feedforward networks, using different amino acid features as input (Wu et al., 1992); and (3) on Kohonen maps (see Self-Organizing Feature Maps Kohonen Maps), using residue pair frequencies as input (Ferrara and Ferrara, 1992). Whereas feedforward networks are useful to learn a classification into known features (secondary structure, structural class), the Kohonen maps have been applied to render a general classification scheme (e.g., A and B are similar, and A is more similar to C than B). Such a classification is in general a priori not evident (and in itself provides a controversial research area, attempting to answer questions like "Are we more similar to an orangutan than to a pig?").

Prediction of Other Structural or Functional Features

Most applications of neural networks use a similar sliding window input as described above. Approaches address the predictions of surface exposure, disulfide bonds, and function-specific sequence motifs.

Surface exposure. A simple feature of 3D structure is also of interest for molecular biology, to the extent to which a residue is exposed to the solvent. de Hoogbroek, Muzet, and Kim (1990) used a network to classify amino acid residues as either buried or exposed. The results were evaluated on too small a data set, yielding some 70% accuracy in two states (buried/exposed).

Disulfide bonds. Disulfide bonds between cysteine residues (one of the 20 amino acids is cysteine) are often of functional and structural importance. Mukal, de Hoogbroek, and Kim (1990) used a single-layer feedforward network to predict the existence or absence of disulfide bonds.

Function-specific sequence motifs. Often, function depends on a rather short (5 to 10 residues) sequence motif (unique pattern of adjacent amino acids). Residues that are associated with particular functions were also subject to neural network predictions. Examples are: (1) sequence motifs that reveal binding of energy storage molecules (Hurst and Sternberg, 1992); (2) sequence motifs specific for particular proteins—e.g., the immunoglobulins (Beggio and Poulos, in Hurst and Sternberg, 1992); and (3) signal peptide motifs found in sequences (Lodat et al., 1991).

Aiming at Prediction in 3D

Distance Constraints

The projection of the 3D structure onto a two-dimensional distance matrix (Figure 1) could be an important step on the way to predicting 3D structure. This enterprise was undertaken by Bohl et al. (1990), who used a neural network to predict residues which are closer than 8 angstroms (= 8 x 10^{-10} m) to any of the 30 residues adjacent in sequence. The predicted fragments of the distance matrix were used for a simple steepest descent energy minimization procedure. The training set comprised 13 proteins. The method was tested on only one protein that has sufficient sequence identity to proteins used for training. (The prediction was worse than the one that could have been obtained by homology modeling.)

Spin-Glass Models for Proteins

The putative analogies of the energy landscapes of spin glasses and proteins led to a multitude of models attempting to describe protein folding with the formalism known from spin-glass theory (Elber, 1993). Such models have been used for attempts to predict 3D structure (Goldstein, Lubensky-Cholet, and Weinig, 1992). The principal idea is to define an effective energy functional from a database of known 3D structures that is flexible enough to encode description of a large class of structures and simple enough to surmount the multiple minima problem of conventional energy minimization calculations for proteins. The analogy to spin-glass theory consists in constructing an energy function based on pairwise interactions (between
Discussion

Neural networks can be used for predicting structural features of proteins. There were at least 50 articles on the application of neural networks for protein structure prediction until 1993. One message of the literature is convincingly: neural networks can be used to predict secondary structure, structural class, family relations, surface exposure, functional motifs, distance matrices, and even the 3D structure of proteins.

Neural network methods are seldom superior to non-network approaches. The second message of the literature is that networks are superior to alternative techniques, but this answer is not convincing. The general problem is a lack of rigor in evaluating results. A common example is the allowance of significant sequence identity between test and training set. Any evaluation that allows for sequence identity has to be compared to homology modeling. And in this comparison, all prediction methods are clearly inferior. The conclusion is that neural network applications have almost never yielded significant improvements over current techniques (Hirst and Sternberg, 1992). An exception is a network that uses evolutionary information to predict secondary structure (Ross and Sander, 1994a). So far, this is the only example for a neural network prediction of protein structure being clearly superior to alternative techniques.

Neural network predictions have not been made sufficiently available to biochemists. Unfortunately, the tendency to oversell the performance accuracy of network predictions has not contributed much to their acceptance by biochemists. Another problem is that almost none of the network methods is publicly available to those researchers who need predictions.

Neural network techniques will continue to be useful for the prediction of protein structure. First, the problems of predicting protein structure is far from solved. For a sequence of unknown 3D structure for which no homology to a known fold can be detected, the best one can achieve today is a more or less reliable prediction of secondary structure, surface exposure, or functional class. Second, the constantly growing data banks provide an increasing body of information about protein structure. Changes are that methods based on data bank analysis will be the first to practically solve the prediction of protein 3D structure. Third, neural networks might be well suited for appropriately incorporating the increased information. Using evolutionary information will be one way to improve predictions by networks. Neural network applications can become increasingly important for the research of tomorrow's molecular biology, provided that testing is done with care and that methods become available to potential users.

Read More: Applications of Neural Networks
Background: I.2. Dynamics and Adaptation in Neural Networks

References


Pursuit Eye Movements

Richard J. Krauzlis

Introduction

When viewing objects, monkeys and humans use a combination of saccade and pursuit eye movements to keep the retinal image of the object of regard within the high-acyt region near the fovea. While these movements mix seamlessly in normal behavior, their properties and origins are quite distinct. Saccades are ballistic movements that quickly direct the eyes toward a visual target, thereby translating the image of the target from an eccentric retinal location to the fovea. In contrast, pursuit is a continuous movement that slowly rotates the eyes to compensate for any motion of the visual target, minimizing the drift of the target's image across the retina that might otherwise compromise visual acuity. While other mammalian species can generate smooth optokinetic eye movements—which track the motion of the entire visual surround—only primates