Protein Secondary Structure Modelling with Probabilistic Networks
(Extended Abstract)

Arthur L. Delcher[*]
Computer Science Dept.
Loyola College
Baltimore, MD 21210

Simon Kasif[*]
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218

Harry R. Goldberg
Mind-Brain Institute
Johns Hopkins University
Baltimore, MD 21218

William H. Hsu
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218

Abstract

In this paper we study the performance of probabilistic networks in the context of protein sequence analysis in molecular biology. Specifically, we report the results of our initial experiments applying this framework to the problem of protein secondary structure prediction. One of the main advantages of the probabilistic approach we describe here is our ability to perform detailed experiments where we can experiment with different models. We can easily perform local substitutions (mutations) and measure (probabilistically) their effect on the global structure. Window-based methods do not support such experimentation as readily. Our method is efficient both during training and during prediction, which is important in order to be able to perform many experiments with different networks. We believe that probabilistic methods are comparable to other methods in prediction quality. In addition, the predictions generated by our methods have precise quantitative semantics which is not shared by other classification methods. Specifically, all the causal and statistical independence assumptions are made explicit in our networks thereby allowing biologists to study and experiment with different causal models in a convenient manner.

Introduction

In this paper we discuss several experiments with probabilistic networks for predicting the secondary structure of proteins. We believe that the networks that we use provide a very convenient medium for scientists to experiment with different empirical models and obtain possibly important insights into problems being studied. A number of methods have been applied to this problem with various degree of success [Chou and Fasman, 1978; Garnier *et al.*, 1978; Holley and Karplus, 1989; Cost and Salzberg, 1993; Qian and Sejnowski, 1988; Maclin and Shavlik, 1992; Zhang *et al.*, 1993; Muggleton and King, 1991]. In addition to obtaining experimental results comparable to other methods, there are several theoretically and practically interesting observations that we have made in experimenting with our systems. The most important aspect of this approach is that the results obtained have a precise probabilistic semantics. Conditional independence assumptions are represented explicitly in our networks by means of causal links.

- It has been claimed in several papers that probabilistic (statistical) approaches have been outperformed by neural network methods and memory-based methods by a wide margin. We show that probabilistic methods are comparable to other methods in prediction quality. In addition, the predictions generated by our methods have precise quantitative semantics which is not shared by other classification methods. Specifically, all the causal and statistical independence assumptions are made explicit in our networks thereby allowing biologists to study causal links in a convenient manner. This generalizes correlation studies that are normally used in statistical analysis of data.
- Our methods provide very flexible tools to experiment with a variety of modelling strategies. This flexibility allows a biologist to perform many practically important statistical queries which can yield important insight into a problem.
- From the theoretical point of view we found that different ways to model the domain produce practically different results. This is an experience that AI researchers encounter repeatedly in many knowledge-representation schemes: different coding of the problem in the architecture results in dramatic differences in performance. This has been observed in production systems, neural networks, constraint networks and other representations. Our experience reinforces the thesis that while knowledge representation is a key issue in AI, a knowledge-representation

[*] Supported by NSF DARPA Grant CCR-8908092 and AFOSR Grant AFOSR-89-1151

system typically provides merely the programming language in which a problem must be expressed. The coding, analogous to an algorithm in procedural languages, is perhaps of equally great importance. However, the importance of this issue is grossly underestimated and not studied as systematically and rigorously as knowledge representation languages.

- Previous methods for protein folding were based on the window approach. That is, the learning algorithm attempted to predict the structure of the central amino acid in a "window" of k amino acids residues. It is well recognized that in the context of protein folding, very minimal mutations (amino acid substitutions) often cause significant changes in the secondary structure located far from the mutation cite. Our method is aimed at capturing this behavior.

In this paper we describe out initial experiments, for which we have chosen the simplest possible models. We first describe a causal-tree model using Pearl's belief updating. Then we describe the application of the Viterbi algorithm to this model and compare the results. We then illustrate the utilitly of probabilistic models in the context of modelling the effect of mutations on secondary structure. Finally, we describe an application of Hidden Markov Models to modelling protein segments with uniform secondary structure (*i.e.*, runs of helices, sheets or coils).

Protein Folding

Proteins have a central role in essentially all biological processes. They control cellular growth and development, they are responsible for cellular defense, they control reaction rates, they are responsible for propagating nerve impulses, and they serve as the conduit for cellular communication. The ability of proteins to perform these tasks, *i.e.*, the *function* of a protein, is directly related to its *structure*. The results of Christian Anfinsen's work in the late 1950's indicated that a protein's unique structure is specified by its amino-acid sequence. This work suggested that a protein's conformation could be specified if its amino acid sequence was known, thus defining the protein folding problem. Unfortunately, nobody has been able to put this theory into practice.

The biomedical importance of solving the protein folding problem cannot be overstressed. Our ability to design genes—the molecular blueprints for specifying a protein's amino acid sequence—has been refined. These genes can be implanted into a cell and this cell can serve as the vector for the production of large quantities of the protein. The protein, once isolated, potentially can be used in any one of a multitude of applications—uses ranging from supplementing the human defense system to serving as a biological switch for controlling abnormal cell growth and development. A critical aspect of this process is the ability to specify the amino acid sequence which defines the required conformation of the protein.

Traditionally, protein structure has been described at three levels. The first level defines the protein's amino acid sequence, the second considers local conformations of this sequence, *i.e.*, the formation of rod-like structures called α-helices, planar structures called β-sheets, and intervening sequences often categorized as coil. The third level of protein structure specifies the global conformation of the protein. Due to limits on our understanding of solutions to the protein folding problem, most of the emphasis on structure prediction has been at the level of secondary structure prediction.

There are fundamentally two approaches that have been taken to predict the secondary structure of proteins. The first approach is based on theoretical methods and the second is based on data derived empirically. Theoretical methods rely on our understanding of the rules governing amino acid interactions, they are mathematically sophisticated and computationally time-intensive. Conversely, empirically based techniques combine a heuristic with a probabilistic schema in determining structure. Empirical approaches have reached prediction rates approaching 70%—the apparent limit given our current base of knowledge.

The most obvious weakness of empirically based prediction schemes is their reliance on exclusively local influences. Typically, a window that can be occupied by 9-13 amino acids is passed along the protein's amino acid sequence. Based on the context of the central amino acid's sequence neighbors, it is classified as belonging to a particular structure. The window is then shifted and the amino acid which now occupies the central position of the window is classified. This is an iterative process which continues until the end of the protein is reached. In reality, the structure of an amino acid is determined by its local environment. Due to the coiled nature of a protein, this environment may be influenced by amino acids which are far from the central amino acid in sequence but not in space. Thus, a prediction scheme which considers the influence of amino acids which are, in sequence, far removed from the central amino acid of the window may improve our ability to successfully predict a protein's conformation.

Preliminaries

For the purpose of this paper, the set of proteins is assumed to be a set of sequences (strings) over an alphabet of twenty characters (different capital letters) that correspond to different amino acids. With each protein sequence of length n we associate a sequence of secondary structure descriptors of the same length. The structure descriptors take three values: h, e, c that correspond to α-helix, β-sheet and coil. That is, if we have a subsequence of $hh \ldots h$ in positions $i, i+1, \ldots, i+k$ it is assumed that the protein sequence in those positions folded as a helix. The classification problem is typically stated as follows. Given a protein

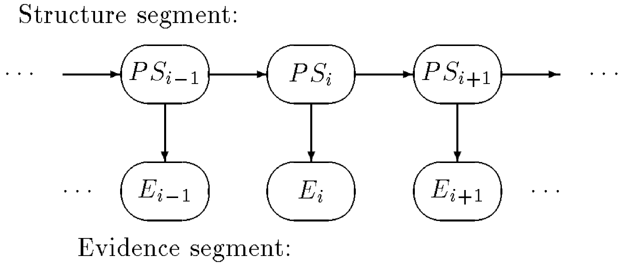

Figure 1: Causal tree model.

sequence of length n, generate a sequence of structure predictions of length n which describes the secondary structure of the protein sequence. Almost without exception all previous approaches to the problem have used the following approach. The classifier receives a window of length $2K + 1$ (typically $K < 12$) of amino acids. The classifier then predicts the secondary structure of the central amino acid (*i.e.*, the amino acid in position K) in the window.

A Probabilistic Framework for Protein Analysis

When making decisions in the presence of uncertainty, it is well-known that Bayes rule provides an optimal decision procedure, assuming we are given all prior and conditional probabilities. There are two major difficulties with using the approach in practice. The problem of reasoning in general Bayes networks is $\mathcal{NP}$-complete, and we often do not have accurate estimates of the probabilities. However, it is known that when the structure of the network has a special form it is possible to perform a complete probabilistic analysis efficiently. In this section we show how to model probabilistic analysis of the structure of protein sequences as belief propagation in causal trees. In the full version of the paper we also describe how we dealt with problems such as undersampling and regularization. The general schema we advocate has the following form. The set of nodes in the networks are either protein-structure nodes (PS-nodes) or evidence nodes (E-nodes). Each PS-node in the network is a discrete random variable X_i that can take values which correspond to descriptors of secondary structure, *i.e.*, segments of h's, e's and c's. With each such node we associate an evidence node that again can assume any of a set of discrete values. Typically, an evidence node would correspond to an occurrence of a particular subsequence of amino acids at a particular location in the protein. With each edge in the network we will associate a matrix of conditional probabilities. The simplest possible example of a network is given in Figure 1.

We assume that all conditional dependencies are represented by a causal tree. This assumption violates some of our knowledge of the real-world problem, but provides an approximation that allows us to perform an efficient computation. For an exact definition of a causal tree see Pearl [Pearl, 1988]. Causal belief networks can be considered as a generalization of classical Markov chains that have found many useful applications in modelling.

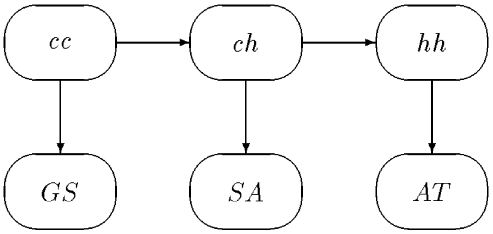

Figure 2: Example of causal tree model using pairs, showing protein segment $GSAT$ with corresponding secondary structure $cchh$

Protein Modeling Using Causal Networks

As mentioned above, the network is comprised of a set of protein-structure nodes and a set of evidence nodes. Protein-structure nodes are finite strings over the alphabet $\{h, e, c\}$. For example the string $hhhhhh$ is a string of six residues in an α-helical conformation, while $eecc$ is a string of two residues in a β-sheet conformation followed by two residues folded as a coil. Evidence nodes are nodes that contain information about a particular region of the protein. Thus, the main idea is to represent physical and statistical rules in the form of a probabilistic network. We note that the main point of this paper is advocating the framework of causal networks as an experimental tool for molecular biology applications rather than focusing on a particular network. The framework allows us flexibility to test causal theories by orienting edges in the causal network.

In our first set of experiments we converged on the following model that seems to match in performance many existing approaches. The network looks like a set of PS-nodes connected as a chain. To each such node we connect a single evidence node. In our experiments the PS-nodes are strings of length two or three over the alphabet $\{h, e, c\}$ and the evidence nodes are strings of the same length over the set of amino acids. The following example clarifies our representation. Assume we have a string of amino acids $GSAT$. We model the string as a network comprised of three evidence nodes GS, SA, AT and three PS-nodes. The network is shown in Figure 2. A correct prediction will assign the values cc, ch, and hh to the PS-nodes as shown in the figure.

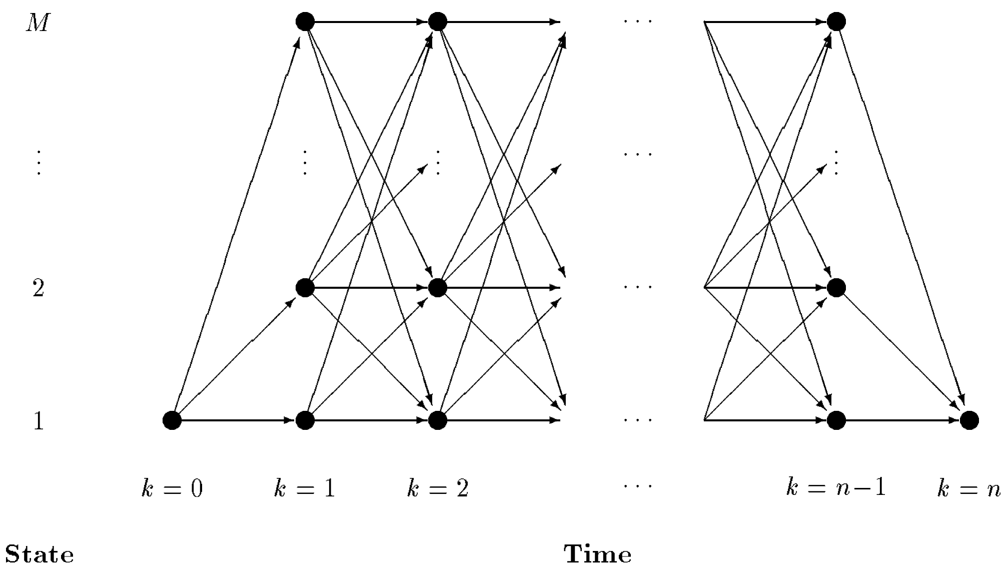

Figure 3: Modelling the Viterbi algorithm as a shortest path problem.

Let $X_0, X_1, \ldots, X_n$ be a set of PS-nodes connected as in Figure 1. Generally, speaking the distribution for the variable X_i in the causal network as below can be computed using the following formulae. Let $e^-_{X_i} = e_i, e_{i+1}, \ldots, e_n$ denote the set of evidence nodes to the right of X_i, and let $e^+_{X_i} = e_1, e_2, \ldots, e_{i-1}$ be the set of evidence nodes to the left of X_i. By the assumption of independence explicit in the network we have

$$P(X_i | X_{i-1}, e^+_{X_i}) = P(X_i | X_{i-1})$$

Thus,

$$P(X_i | e^+_{X_i}, e^-_{X_i}) = \alpha P(e^-_{X_i} | X_i) P(X_i | e^+_{X_i})$$

where α is some normalizing constant. For length consideration we will not describe the algorithm to compute the probabilities. The reader is referred to Pearl for a detailed description [Pearl, 1988]. Pearl gives an efficient procedure to compute the belief distribution of every node in such a tree. Most importantly, this procedure operates by a simple efficient propagation mechanism that operates in linear time.

Protein Modeling Using the Viterbi Algorithm

In this section we describe an alternative model for prediction. This model has been heavily used in speech understanding systems, and indeed was suggested to us by Kai Fu Lee whose system using similar ideas achieves remarkable performance on speaker-independent continuous speech understanding.

We implemented the Viterbi algorithm and compare its performance to the method outlines above. We briefly describe the method here. We follow the discussion by Forney [Forney, 1973].

We assume a Markov process which is characterized by a finite set of state transitions. That is, we assume the process at time k can be described by a random variable X_k that assumes a discrete number of values (states) $1, \ldots, M$. The process is Markov, *i.e.*, the probability $P(X_{k+1} | X_0, \ldots X_k) = P(X_{k+1} | X_k)$. We denote the process by the sequence $X = X_0, \ldots, X_k$. We are given a set of observations $Z = Z_0, \ldots, Z_k$ such that Z_i depends only on the transition $T_i = (X_{i+1}, X_i)$. Specifically, $P(Z|X) = \prod_{k=0}^{n} (Z_i | X_i)$. The Viterbi algorithm is a solution to the maximum aposteriori estimation of X given Z. In other words we are seeking a sequence of states X for which $P(Z|X)$ is maximized.

An intuitive way to understand the problem is in graph theoretic terms. We build a n-level graph that contains nM nodes (see Figure 3). With each transition we associate an edge. Thus, any sequence of states has a corresponding path in the graph. Given the set of observations Z with any path in the graph we associate a length $L = -\ln P(X, Z)$. We are seeking a shortest path in the graph. However, since

$$\begin{aligned} P(X, Z) &= P(X) P(Z|X) \\ &= \prod_{k=0}^{n-1} P(X_{k+1} | X_k) \prod_{k=0}^{n-1} P(Z_k | X_{k+1}, X_k) \end{aligned}$$

if we define $\lambda(T_k) = -\ln P(X_{K+1} | X_K) - \ln P(Z_k | T_k)$ we obtain that $-\ln P(Z, X) = \sum_{k=0}^{n-1} \lambda_k$.

Now we can compute the shortest path through this graph by a standard application of shortest path algorithms specialized to directed acyclic graphs. For each time step i we simply maintain M paths which are the shortest path to each of the possible states we could be in at time i. To extend the path to time step $i+1$ we simply compute the lengths of all the paths extended by one time unit and maintain the shortest path to each one of the M possible states at time $i+1$.

Our experimentation with the Viterbi algorithm was completed in Spring 1992. We recently learned that David Haussler [Haussler *et al.*, 1992] and his group suggested the Viterbi algorithm framework for protein analysis as well. They experimented on a very different problem and also obtain interesting results. We document the performance of Viterbi on our problem even though, as described below, the causal-tree method outperformed Viterbi. The difference between the methods is that the Viterbi algorithm predicts the most likely complete sequence of structure elements, whereas the causal-tree method makes separate predictions about individual PS-nodes.

Experiments

The experiments we conducted were performed to allow us to make a direct comparison with previous methods that have been applied to this problem. We followed the methodology described in [Zhang *et al.*, 1993; Maclin and Shavlik, 1992] which did a thorough cross-validated testing of various classifiers for this problem. Since it is known that two proteins that are homologous (similar in chemical structure) tend to fold similarly and therefore generate accuracies of predictions that are often overly optimistic, it is important to document the precise degree of homology between the training set and the testing set. In our experiments the set of proteins was divided into eight subsets. We perform eight experiments in which we train the network on seven subsets and then predict on the remaining subset. The accuracies are averaged over all eight experiments. This methodology is referred to as k-way cross validation.

Experimental Results

We report the accuracy of prediction on individual residues and also on predicting runs of helices and sheets. Table 1 shows the prediction accuracy of our methods using the causal network method for each one of the eight trials in our 8-way cross-validation study. In the pairs column we document the performance of the causal network described earlier using PS-nodes and E-nodes that represent protein segments of length 2. The triples column gives the results for the same network with segments of length 3. The decrease in accuracy for triples is a result of undersampling.

Table 2 shows the performance of our method in predicting the secondary structure at each amino acid position in comparison with other methods. In Table 3 we report the performance of our method on predicting runs of helices and sheets and compare those with other methods that were applied to this problem. A typical output of our experiments is shown in Figure 4

To summarize, our method yields performance comparable to other methods on predicting runs of helices and sheets. It seems to have particularly high accuracy in predicting individual helices.

Trial	**Positions**	**Correct Using:**	
		Pairs	**Triples**
1	2339	1518 (64.9%)	1469 (62.8%)
2	2624	1567 (59.7%)	1518 (57.9%)
3	2488	1479 (59.5%)	1435 (57.7%)
4	2537	1666 (65.7%)	1604 (63.2%)
5	2352	1437 (61.1%)	1392 (59.2%)
6	2450	1510 (61.6%)	1470 (60.0%)
7	2392	1489 (62.3%)	1447 (60.5%)
8	2621	1656 (63.2%)	1601 (61.1%)
All	19803	12322 (62.2%)	11936 (60.3%)

Table 1: Causal tree results for 8-way cross-validation using segments of length 2 and length 3.

Method	**Total**	**Helix**	**Sheet**	**Coil**
Chou-Fasman	57.3%	31.7%	36.9%	76.1%
ANN	61.8%	43.6%	18.6%	86.3%
w/ state	61.7%	39.2%	24.2%	86.0%
FSKBANN	63.4%	45.9%	35.1%	81.9%
w/o state	62.2%	42.4%	26.3%	84.6%
Viterbi	58.5%	48.3%	47.0%	69.3%
Chain-Pairs	62.2%	55.9%	51.7%	67.4%
Chain-Triples	60.3%	53.0%	45.5%	70.8%

Table 2: Overall prediction accuracies for various prediction methods. Comparative method results from [Maclin and Shavlik, 1992].

Towards Automated Site-Specific Muta-genesis

An experiment which is commonly is done in biology laboratories is a procedure where a particular site in a protein is changed (*i.e.*, a single amino-acid mutation) and then it is tested whether the protein settles into a different conformation. In many cases, with overwhelming probability the protein does not change its secondary structure outside the mutated region. One experiment that is easy to do using our method is the following procedure. We assume the structure of a protein is known anywhere outside a window of length l, $l = 1, 2, 3, \ldots$ and try predict the structure inside the unknown window. Table 4 shows the results of such an experiment.

The results above are conservative estimates of the accuracy of prediction for this type of an experiment

```
* * * Protein #2:

      Predicted       --- Pair Weights --        --- Trip Weights --     Counts
Real  Pair Trip         h       e       c          h       e       c      Pair Trip

E c   c    c        0.0000  0.0000  1.0000     0.0000  0.0000  1.0000     38    4
N c   c    c        0.0034  0.0051  0.9915     0.0059  0.0046  0.9895     64    6
L c   c    c        0.2347  0.1267  0.6386     0.0461  0.0101  0.9439    113    8
K c   c    c        0.2689  0.2714  0.4596     0.0724  0.0478  0.8797     93    6
L c   c    c        0.2817  0.3354  0.3828     0.1056  0.1474  0.7471     85    5
G c   c    c        0.2582  0.3165  0.4253     0.1280  0.1910  0.6809     58    4
F e   e    e        0.2662  0.5115  0.2223     0.1675  0.4686  0.3638     46    4
L e   e    e        0.2607  0.5691  0.1702     0.2066  0.4808  0.3125    105   11
V e   e    e        0.2589  0.5434  0.1976     0.2159  0.4034  0.3807     89    2
K c   c    c        0.2283  0.3629  0.4088     0.2156  0.3189  0.4656     18    2
Q c   c    c        0.2168  0.0927  0.6904     0.2106  0.1199  0.6695     35    3
P c   c    c        0.2250  0.0149  0.7601     0.2085  0.0455  0.7460     65    5
E c   c    c        0.4105  0.0140  0.5754     0.2076  0.0113  0.7811     68    3
E c   h X  c        0.5121  0.0114  0.4765     0.2267  0.0014  0.7719     35    0
P c   h X  c        0.5789  0.0100  0.4111     0.2772  0.0000  0.7228     11    0
W c   h X  h X      0.7637  0.0139  0.2225     0.9953  0.0000  0.0047      7    0
F h   h    h        0.8217  0.0173  0.1610     0.9999  0.0000  0.0001     20    1
Q h   h    h        0.8430  0.0156  0.1414     0.9999  0.0000  0.0001     38    4
T h   h    h        0.8353  0.0100  0.1547     0.9972  0.0000  0.0028     47    1
E h   h    h        0.9240  0.0141  0.0619     0.9961  0.0000  0.0039     19    1
W h   h    h        0.9402  0.0247  0.0351     0.9994  0.0000  0.0006     21    0
K h   h    h        0.9341  0.0254  0.0405     0.9999  0.0000  0.0001     36    1
F h   h    h        0.8908  0.0237  0.0855     0.9944  0.0000  0.0056     43    1
A h   h    h        0.8450  0.0127  0.1423     0.9425  0.0000  0.0575     86    4
D h   h    h        0.7480  0.0033  0.2487     0.9548  0.0000  0.0452     60    6
K h   h    h        0.5636  0.0027  0.4337     0.9528  0.0001  0.0472    132   11
A h   c X  h        0.4406  0.0032  0.5562     0.9373  0.0001  0.0626    134    8
G h   c X  c X      0.1805  0.0032  0.8163     0.0077  0.0001  0.9922    103    8
K h   c X  c X      0.1967  0.0337  0.7696     0.0148  0.0014  0.9838     60    9
D h   c X  c X      0.2192  0.0450  0.7357     0.0155  0.0050  0.9795     71    4
```

Figure 4: Sample output from prediction experiment. The first column shows the actual amino-acid sequence with corresponding correct secondary structure. The next two columns show the predicted value for length-2 and length-3 segments respectively, with an 'X' indicating and incorrect prediction. The next six columns give the belief values for each of the three possible secondary structure types for each of the two segment lengths. Finally the rightmost two columns are the number of examples of the same amino-acid segment encountered in training. These are used to estimate evidential probabilities in the model.

Description	Chain-Pair	FSKBANN	ANN	Chou-Fasman
Average length of predicted helix run	9.4	8.52	7.79	8.00
Average length of actual helix run	10.3	–	–	–
Percentage of actual helix runs overlapped by predicted helix runs	66%	67%	70%	56%
Percentage of predicted helix runs that overlap actual helix runs	62%	66%	61%	64%
Average length of predicted sheet run	3.8	3.80	2.83	6.02
Average length of actual sheet run	5.0	–	–	–
Percentage of actual sheet runs overlapped by predicted sheet runs	56%	54%	35%	46%
Percentage of predicted sheet runs that overlap actual sheet runs	60%	63%	63%	56%

Table 3: Precision of run (segment) predictions. Comparative method results from [Maclin and Shavlik, 1992].

Length of Predicted Segment	Amino-Acid Positions Predicted Correctly
1	90.38 %
2	87.29 %
3	85.18 %
4	82.99 %
6	79.32 %
8	76.49 %
12	72.39 %
16	69.85 %
20	68.08 %
24	66.94 %

Table 4: Accuracy of prediction of a subsegment of amino acids, given the correct secondary structure information for the remainder of the protein. Results are averaged over all possible segments of the given length in all proteins.

and can be easily improved. We are currently checking whether, the high accuracy of prediction is just a result of momentum effects and the prediction accuracy for transitions from coil-regions to helices and sheets remains low.

Using the EM algorithm

We now briefly mention one more set of experiments that can be performed with a probabilistic model of the type discussed in this paper. (For further details, see the complete version of the paper.) The idea is very simple and is strongly influenced by the methodology used in speech recognition. Our goal in this experiment is to create a simple probabilistic model that recognizes runs of helices. We use the framework of the Viterbi algorithm described above. We previously defined the notion of the most likely path in the probabilistic network given all the evidence. This path can be described as a sequence of nodes (states) in the network, *i.e.*, given a particular sequence of amino acids, we want to find a sequence of states which has the highest probability of being followed given the evidence. Alternatively, we can regard the network as a probabilistic finite state machine that generates amino acid outputs as transitions are made.

In this experiment we would like to create the most likely model that recognizes/generates sequences of helices. Intuitively (and oversimplifying somewhat), we would like to find a network for which the probabilities of traversing a path from initial to final state given helical sequences of amino acids are greater than the probabilities for non-helical sequences. Figure 5 show the network that we used.

Initially we assigned equal probabilities to every transition from a given node, and for each transition we set the probabilities of outputting amino acids to the relative frequencies of those amino acids in the training data. We then use the Baum-Welch method (or EM, expectation-modification) [Rabiner, 1989] to adjust the probabilities in the network to increase its probabilitly of recognizing the input sequences.

We constructed three networks (for helix, sheet and coil runs) and trained them to recognize their respective runs. All the networks were of the form shown in the figure, but were of different lengths, corresponding to the average length of the respective type of run.

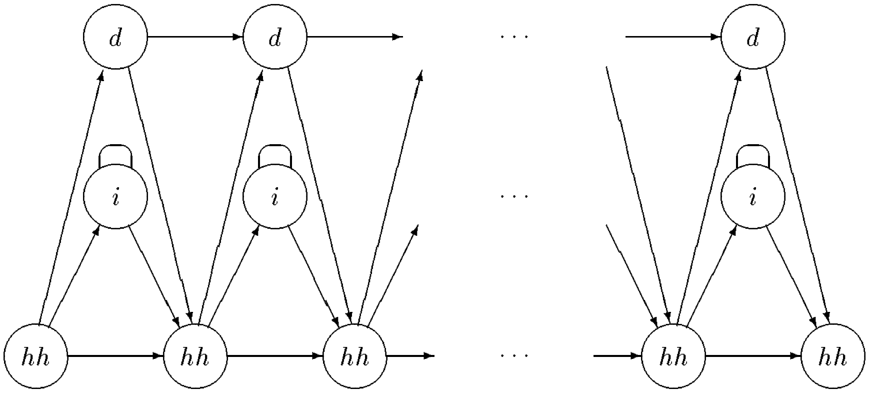

Figure 5: Hidden Markov Model used to recognize a seqence of helices. With each edge out of a node, there is an associated probability of taking that transition together with a set of probabilities of generating each of the 20 possible amino acids while making that transition. Nodes labelled i allow for the insertion of extra amino acids for long chains. Nodes labelled d represent deletion of amino acids thereby permitting the model to generate short chains. Edges to d-nodes generate no amino acid.

	Network Trained to Recognize:		
Input Type	Helices	Sheets	Coils
Helix	469 (91.1%)	34 (6.6%)	12 (2.3%)
Sheet	231 (28.2%)	344 (42.0%)	244 (29.8%)
Coil	433 (33.0%)	114 (8.7%)	766 (58.3%)

Table 5: Relative frequencies with which HMM networks had highest probabilities of generating sequences of particular type.

The helix network had 9 nodes on the bottom level, the sheet network 4, the coil network 6. We then tested the networks by giving each one the same run sequence and computing its probability (using the Viterbi algorithm) of generating that sequence. Table 5 shows the relative frequency with which each of the 3 networks had the highest probability of generating each type of input sequence. The fact that helices are predicted far more accurately than sheets is in part attributable to the fact that the helix network is much larger.

By way of comparison, we used the causal tree model of Figure 2 to predict the same segments, it predicted only about 20% of helix-run sequences correctly. This is not surprising when we consider that most of the sequence examples were coils, which strongly biased the model to predict coils.

Discussion

In this paper we have reported several experiments with probabilistic networks as a framework for the problem of protein secondary structure prediction. One of the main advantages of the probabilistic approach we described here is our ability to perform detailed experiments where we can experiment with different probabilistic models. We can easily perform local substitutions (mutations) and measure (probabilistically) their effect on the global structure. Window-based methods do not support such experimentation as readily. Our method is efficient both during training and during prediction, which is important in order to be able to perform many experiments with different networks.

Our initial experiments have been done on the simplest possible models where we ignore many known dependencies. For example, it is known that in α-helices hydrogen bonds are formed between every i^{th} and $(i+4)^{\text{th}}$ residue in a chain. This can be incorporated in our model without losing efficiency. We also can improve our method by incorporating additional correlations among particular amino acids as in [Gibrat *et al.*, 1987]. We achieve prediction accuracy similar to many other methods such as neural networks. We are confident that with sufficient fine tuning we can improve our results to equal the best methods. Typically, the current best prediction methods involve complex hybrid methods that compute a weighted vote among several methods using a combiner that learns the weights. *E.g.*, the hybrid method described by

[Zhang *et al.*, 1993] combines neural networks, a statistical method and memory-based reasoning in a single system and achieves an overall accuracy of 66.4%.

We also have used a more sophisticated model influenced by the techniques used in speech recognition. Our networks are trained to recognize sequences of α-helix/*beta*-sheet/coil runs. Thus, the helix network is designed to generate sequences of amino-acids that are likely to generate runs of helices. A similar approach was used in the paper by [Haussler *et al.*, 1992] to recognize globins. We reported some preliminary results in using such networks for predicting secondary structure. The network that was trained to generate runs of helices did relatively well on identifying such runs during testing on new sequences of helices.

Bayesian classification is a well-studied area and has been applied frequently to many domains such as pattern recognition, speech understanding and others. Statistical methods also have been used for protein-structure prediction. What characterizes our approach is its simplicity and the explicit modeling of causal links (conditional independence assumptions). We believe that for scientific data analysis it is particularly important to develop tools that clearly display such assumptions. We showed that probabilistic networks provide a very convenient medium for scientists to experiment with different empirical models which may yield important insights into problems.

To summarize, scientific analysis of data is an important potential application of Artificial Intelligence (AI) research. We believe that the ultimate data analysis system using AI techniques will have a wide range of tools at its disposal and will adaptively choose various methods. It will be able to generate simulations automatically and verify the model it constructed with the data generated during these simulations. When the model does not fit the observed results the system will try to explain the source of error, conduct additional experiments, and choose a different model by modifying system parameters. If it needs user assistance, it will produce a simple low-dimensional view of the constructed model and the data. This will allow the user to guide the system toward constructing a new model and/or generating the next set of experiments. We believe that flexibility, efficiency and direct representation of causality in probabilistic networks are important desirable features that make them very strong candidates as a framework for biological modelling systems.

References

Chou, P. and Fasman, G. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Advanced Enzymology* 47:45–148.

Cost, S. and Salzberg, S. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* 10(1):57–78.

Forney, G. D. 1973. The Viterbi algorithm. *Proceedings of the IEEE* 61(3):268–278.

Garnier, J.; Osguthorpe, D.; ; and Robson, B. 1978. Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* 120:97–120.

Gibrat, J.-F.; Garnier, J.; and Robson, B. 1987. Further developments of protein secondary structure predicition using information theory. *Journal of Molecular Biology* 198:425–443.

Haussler, D.; Krogh, A.; Mian, S.; and Sjolander, K. 1992. Protein modeling using hidden markov models. Technical Report UCSC-CRL-92-23, University of California, Santa Cruz.

Holley, L. and Karplus, M. 1989. Protein secondary structure prediction with a neural network. In *Proceedings of the National Academy of Sciences USA*, volume 86. 152–156.

Maclin, R. and Shavlik, J. 1992. Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings Tenth National Conference on Artificial Intelligence*. 165–170.

Muggleton, S. and King, R. 1991. Predicting protein secondary structure using inductive logic programming. Technical report, Turing Institute, University of Glasgow, Scotland.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Qian, N. and Sejnowski, T. 1988. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* 202:865–884.

Rabiner, Lawrence R. 1989. A tutorial on hidden markow models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Zhang, X.; Mesirov, J.; and Waltz, D. 1993. A hybrid system for protein secondary structure prediction. *Molecular Biology (to appear)*.