Artificial Intelligence Approaches to Problems in the Human Genome Project

Simon Kasif and Steven Salzberg
Department of Computer Science
The Johns Hopkins University
Baltimore, MD 21218
lastname@cs.jhu.edu

Problems of protein structure

Our research focuses on two important problems in molecular biology, both critical components of the Human Genome Project. The first problem is that of protein folding, where our emphasis is on prediction of secondary structure from primary sequence data. The second problem is identification of the *motifs*, or building blocks, from which all proteins are constructed. Below we summarize our past results and current approaches to these problems.

Protein folding

We consider two approaches to protein folding. Our past work used a memory-based (nearest-neighbor) method, and our current work uses Bayesian methods.

Memory-based methods

The disciplines of machine learning and pattern recognition have for many years been exploring methods for classification. The protein folding problem falls into the set of problems appropriate for these methods. We transform the problem into a classification problem by taking a string of amino acids and attempting to label each character with the name of the secondary structure class to which it belongs. Nearest-neighbor classification, a memory-based method that has been studied since the early 1950s (Fix & Hodges, 1952), is one of the most well-known methods that might be used for this task. More recent methods have emphasized building models in the form of rules, decision trees (Quinlan, 1986), or hyper-rectangles (Salzberg, 1991), but the nearest-neighbor method is probably the simplest algorithm for performing classification. However, when the feature values have symbolic, unordered values (e.g., the 20 amino acids in a globular protein, which have no natural inter-value "distance"), nearest-neighbor methods typically resort to much simpler metrics, such as Hamming distance. The Hamming distance between two amino acid strings is simply the number of positions for which the two strings mismatch. Simpler metrics, however, may fail to capture the complexity of the problem domains, and as a result may not perform well.

Our algorithm enhances standard nearest-neighbor by first constructing (from the training data) a distance metric that gives a numeric distance between any two amino acids, and then attaching weights to individual examples. Intuitively, the distance measure captures the correlation between each two amino acids in the context of the prediction task. A similar technique has been used in Zhang *et al.* (1992). The combination of this distance metric and the instance weights results in a robust nearest-neighbor learning algorithm that works for any domain with symbolic feature values. Our implementation performs as well as any previously-reported method for predicting protein secondary structure. Its training time is much faster than the neural net methods that have been used for this problem. In addition, it is easy to parallelize our algorithm for even greater speed-ups, as we have shown by implementing our system on a parallel machine. This method, described in detail in (Cost and Salzberg 1991), has achieved accuracy of 71.0% on a data set of 106 proteins. In an earlier study using the same data, the back propagation algorithm achieved 64.3% accuracy (Qian & Sejnowski, 1988). More recently, another study by Zhang *et al.* (1992) described on a hybrid method that achieves 66.4% accuracy on a different data set consisting of 107 proteins. All of these methods compare favorably to earlier techniques (e.g., Lim, 1974; Chou & Fasman, 1978; Garnier *et al.*, 1978).

A Bayesian approach

One problem with many of the previous approaches for predicting secondary structure is that they do not produce a qualitative description of the problem domain (e.g., rules). We have devised a simple probabilistic approach that synthesizes an evidence network from which one can extract qualitative rules. The rules of the network are precise probabilistic Bayesian constraints of the form: given this information about a protein, the probability that a specific secondary structure will occur is p. Probabilistic reasoning is a growing discipline in AI, and it has been used in speech processing, automated diagnosis, and more recently in common-

sense reasoning. We are interested in using this framework to devise probabilistic belief networks (e.g., causal trees such as those developed by Pearl (1988)) for a wide range of tasks involving scientific analysis of data. Previous work in AI on scientific discovery has been based on a different paradigm. Our current work is aimed at using the belief network framework to discover new rules about protein folding. The well-done study by Zhang *et al.* (1992) describes another probabilistic method for the same problem which is similar in spirit to our approach.

Searching for motifs

It has been conjectured by several prominent molecular biologists (John Gilbert, Hamilton Smith and others) that all proteins were formed from a common set of *motifs*, or building blocks. This conjecture is based on the hypothesis that nature typically does not tend to invent new mechanisms, but rather reuses and modifies old ones. These motifs (sequences of conserved amino-acids) are, presumably, short sequences of amino acids that can be found in proteins from widely varying life forms. However, until recently it has been impossible to gather evidence supporting this hypothesis, because of the lack of sequence data. With the growth of the protein sequence data base in recent years, it now may be possible to find motifs and perhaps to refine and verify this conjecture. Finding these building blocks could revolutionize molecular biology and protein structure research. Work has begun recently at a small number of places, including Johns Hopkins, to develop methods for finding motifs (Smith *et al.* 1990).

Most of the work on searching protein sequences has focused on pairwise sequence alignment, with some more recent algorithms dedicated to multiple sequence alignment. Optimal multiple sequence alignment problems are typically NP-hard, and solution methods typically use greedy heuristic methods. However, even these heuristic algorithms are too expensive to apply to large sets of proteins, since their complexity is usually $O(n^m)$, for m proteins of length n (Lipman *et al.* 1989. The motif problem does not require alignment of entire proteins, since motifs are only small subsequences, and thus less expensive algorithms may be devised.

Our current approach uses clustering methods to attempt to find significant clusters of short sequences, in which the average intra-cluster distance is minimized, while the inter-cluster distance is maximized. To test our approach, we have gathered a large database containing over 2000 proteins, with the help of collaborators from the Human Genome Database (Prof. Ken Fasman) and in the Department of Molecular Biology (Prof. Hamilton Smith). Our current algorithms have quadratic complexity, and the size of this database presents a computationally expensive problem: the database has over 600,000 strings that are considered to be candidate motifs, where each such string is of length k. (We vary k in our experiments, but the number of strings is basically the same for $10 < k < 50$.) Thus a quadratic algorithm requires roughly 3.6×10^{11} iterations. In order to avoid this expensive calculation, we have chosen heuristic methods such as the following: choose a small number of proteins (e.g., 100), find good candidate motifs, and then search the entire database to determine whether or not those motifs occur frequently across all proteins. We have also devised a simple data structure which is a variant of discrimination net (trie). The data structure compresses the data using a labelled tree. Each path in the tree is a prefix of some string in the database. Thus, all strings that share a common prefix are sons of the same node. This saves us from comparing the same prefixes over and over. The data structure will allow us to achieve substantial constant speed-up for clustering.

Critical issues for discussion and presentations

Scientific analysis of data is a very important application area for AI research. Traditional methods of data analysis, e.g., regression analysis, typically do not function adaptively and require substantial user guidance. They do not generate hierarchical concept descriptions (e.g, decision trees or rules). They do not generate qualitative domain descriptions. They do not generate experiments to validate partially constructed models.

We believe the ultimate data analysis system using AI techniques will integrate a variety of data analysis tools and will do all of the above. It should have a wide range of analysis tools (including statistical methods) at its disposal. It will adaptively choose various methods and automatically modify its behavior based on partial findings. It will be able to generate simulations automatically and verify models constructed based on a given data set. When the model does not fit the data, the system will try to explain the source of error, conduct additional experiments, and choose a different model by modifying system parameters. When it needs user guidance, it will produce a simple low-dimensional view of the model and the multi-dimensional data, which will allow the user to guide the system in designing the next set of experiments.

Our preliminary work on the motif problem fits this general paradigm. Our system comes up with a set of candidate motifs (a model) and tries to validate the model on the full data set. When the model does not fit the data (which happens frequently), the system samples the data again and produces a different model. We are in the process of constructing other analysis tools (in addition to clustering). It will be most interesting to have the system integrate results from different data analysis programs and modify its behavior based on the partial results.

In the conference, we would like to discuss the general paradigm of data analysis based on AI techniques. Along these lines, we can present our work on secondary structure prediction, described above, and our new Bayesian approach. We can also present our work-in-progress on the motif problem and the preliminary results we have.

Recent related publications by the authors

Cost, S. and S. Salzberg (1992). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, in press.

Cost, S. and S. Salzberg (1990). Exemplar-based Learning to Predict Protein Folding. *Proceedings of the 1990 Symposium on Computer Applications in Medical Care* (pp. 114–118), Washington, D.C., November 1990.

Heath, D., S. Kasif, and S. Salzberg (1992). Learning oblique decision trees. Technical Report JHU-92/05, Department of Computer Science, Johns Hopkins University, April 1992.

Kasif, S., A. Delcher, S. Salzberg, and B. Hsu (1992). Secondary Prediction with Probabilistic Belief Methods, JHU Technical Report (in preparation).

Salzberg, S. (1992). Predicting Protein Secondary Structure with a Nearest-Neighbor Algorithm. *Journal of Molecular Biology*, to appear, 1992.

Salzberg, S. (1991) Distance Metrics for Instance-Based Learning. *Methodologies for Intelligent Systems: 6th International Symposium, ISMIS '91*, Z. Ras and M. Zemankova (eds.), pp. 399-408. New York: Springer-Verlag, 1991.

Salzberg, S. (1991). A Nearest Hyperrectangle Learning Method. *Machine Learning, 6*, 251-276.

Salzberg, S., A. Delcher, D. Heath, and S. Kasif (1991). Learning with a Helpful Teacher. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 705-711). Sydney, Australia, August 1991.

Other references

Chou, P. & Fasman, G. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advanced Enzymology* **47**, 45-148.

Fix, E.F. & Hodges, J.L. (1952). Discriminatory analysis: Nonparametric discrimination: Small sample performance. Project 21-49-004, Report No. 11, USAF School of Aviation Medicine, Randolph Field, Texas, 1952, 280-322.

Garnier, J., D. Osguthorpe, & B. Robson (1978). Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Lim, V. (1974). Algorithms for prediction of α-helical and *beta*-structural regions in globular proteins. *J. Mol. Biol.* **88**, 873-894.

Lipman, D., S. Altschul, and J. Kececioglu (1989). A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences, 86*, 4412-4415.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Los Altos, CA: Morgan Kaufmann.

Qian, N. & T. Sejnowski (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865-884.

Smith, H., T. Annau, and S. Chandrasegaran (1990). Finding sequence motifs in groups of functionally related proteins. *Proceedings of the National Academy of Sciences, 87*, 826-830.

Zhang, X., J. Mesirov, & D. Waltz (1992). A hybrid system for protein secondary structure prediction. *J. Mol. Biol.*, in press.