

# Automatic Discovery of Protein Motifs

## Ninth International ACS Biotechnology Symposium

**Douglas L. Brutlag** and **Tod M. Klingler**

Department of Biochemistry and Section on Medical Informatics  
Stanford University School of Medicine, Stanford, CA 94305-5307

We have developed a novel representation of protein motifs that permits the rapid discovery of structural features in sets of protein sequences with a common structure or function. Many popular methods for representing protein motifs (consensus sequences, weight matrices, profiles, etc.) emphasize conservation of amino acids at specific sites in the sequence. Our method looks for correlations between amino acid variations at distinct sites. Correlations between the residues represent side-chain side-chain interactions and give insight into the structural properties of the motifs. Structural correlations can be used in database search to discover other proteins bearing similar relationships. This database search is significantly more sensitive than methods depending only upon conserved residues.

### Introduction

Most methods for representing protein motifs emphasize the amino acids conserved during evolution. These conserved residues are often of critical importance in the structure or function of the protein. Dictionaries of such conserved motifs have been compiled and are extremely valuable in discovering structural and functional attributes of novel protein sequences (Bairoch, 1991).

Many protein motifs are not conserved sufficiently to be represented as a consensus sequence (Dodd and Egan, 1987). When motifs are highly variable in sequence, one usually employs probabilistic methods such as weight matrices or profiles (Staden, 1984; Gribskov, McLachlan and Eisenberg, 1987). Weight matrices give the likelihood of finding each amino acid at each position in the motif based on a large set of examples. The likelihood of finding each amino acid is calculated relative to the overall frequency of finding that residue in the proteins or in the database being examined. Like the consensus sequence method, weight matrices emphasize residues conserved in evolution.

Often, one only has a few examples of a particular protein structure, which are insufficient to determine all the likelihoods required in a weight matrix accurately. In these cases, the weight matrix generated by a few examples can be multiplied by the likelihoods of each amino acid replacing another during evolution (these

replacement matrices are known as PAM matrices, Schwartz and Dayhoff, 1979). The resulting likelihood matrix is known as a protein profile (Gribskov, McLachlan and Eisenberg, 1987). Profiles, again, give a measure of evolutionary sequence similarity rather than insights into structural attributes of the protein.

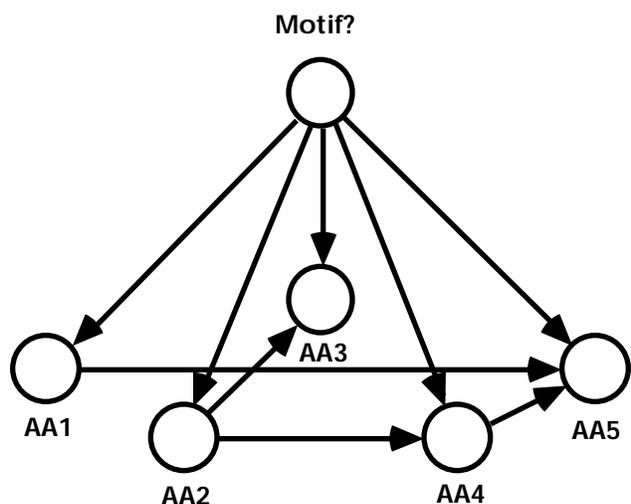
### Methods

We have undertaken a novel approach that emphasizes sequence variation within proteins of common structure or function. Along with the amino acids conserved at any one site, we have looked specifically for conserved correlations between amino acids at two distinct sites in a motif. We felt that if the structure of a motif was highly conserved independent of the sequence, then correlated changes at two distinct sites may reflect conserved amino acid side-chain interactions. If correlations between positions were conserved in evolution despite sequence variation, testing for conserved correlations as well as conserved positions would be a more discriminating method for detecting structural motifs.

In order to represent both correlated changes and conserved residues simultaneously we have used a probabilistic method known as a belief network (Neapolitan, 1990; Pearl, 1988). Belief networks are general graphic representations of Bayesian conditional probabilities. Nodes in the belief network shown in Figure 1 are of two kinds. The top node, labeled "Motif?", represents a decision node with two possible values,

## Automatic Discovery of Protein Motifs Ninth International ACS Biotechnology Symposium

yes or no. The evidence nodes in these networks represent the amino acid residues that occur in five positions along a protein motif. Each has 20 possible values. The arcs drawn between the decision node and the evidence nodes represent the conditional probabilities of each amino acid given a value for the decision node. Values for these conditional probabilities are obtained from training sets of sequences and, like weight matrices, they represent the conservation of residues at each position in the motif.



**Figure 1.** A generic belief network graphically illustrating the relationships between a decision node labeled Motif? and the evidence nodes (AA<sub>n</sub>). The decision node has two possible values and the evidence nodes have twenty possible values, one for each amino acid. The arc between any two nodes represents the conditional probabilities relating values for the two nodes. In this belief network there are conditional probabilities for each position in the motif depending on the knowledge of the motif and there are also dependencies between amino acids at different positions. These later conditional probabilities represent correlations between amino acids at those two positions observed among many examples of the protein motif.

Arcs between two amino acid nodes represent the probability of amino acid correlations between two positions in the training set. We calculate probabilities of amino acid correlation using  $\chi^2$  statistics and check them using a

Monte Carlo simulation. Comparisons between all pairs of nodes are determined and those arcs whose probability lies below 0.01 are retained in the belief network. Normally arcs between two positions would represent 400 different possible correlations (as there are 20 different amino acids at each position) and thus would require over 2,000 examples of a motif to determine the each correlation significantly. With limited numbers of examples, we generally can not compare individual amino acid residues at each position. Instead, we classify amino acids into a few structural or functional types and look for correlations between types of amino acids. Our initial amino acid classes include nonpolar (ILMVAGPFWY), polar but uncharged (CNQST), basic (HKR) and acidic (DE) classes. With four classes we can readily detect correlations between types of amino acids with as few as 80 examples.

Once correlations between positions are discovered, belief networks like the one shown in Figure 1 can be used to search for motifs as well. Given values for all the amino acids in the chain, one can infer the posterior probability that a sequence is the motif in question (Lauritzen and Spiegelhalter, 1988). Applying this probabilistic inference procedure to each position in a protein sequence database allows one to perform search.

### Results : The Helix-Turn-Helix Motif

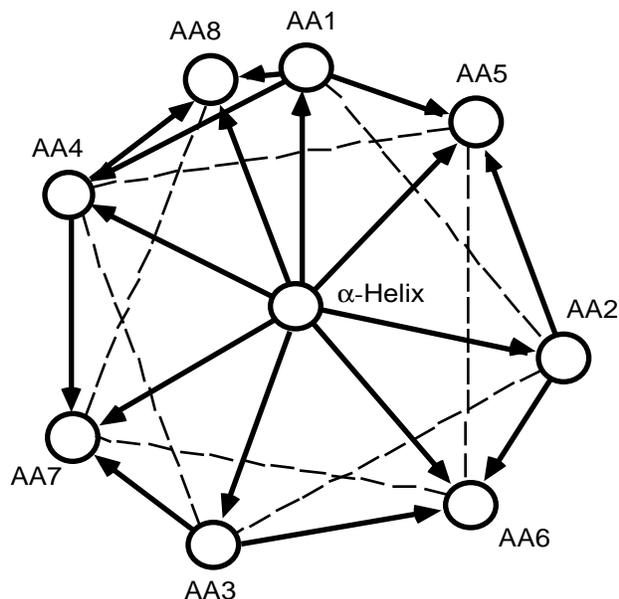
We have examined 80 prokaryotic helix-turn-helix containing proteins for correlations among the 22 positions comprising the helix-turn-helix motif. We discovered seven pairs of positions none of which were highly conserved, but which were strongly correlated with each other. When a database search was conducted for sequences giving a high score with either a weight matrix (finds conserved residues) or a belief network (finds conserved residues and correlations simultaneously), the number of false positives was reduced at all threshold levels. Especially significant were the relative scores for sequences clearly unrelated to a helix-turn-helix motif. The relative scores for such sequences for the belief network were 40 to 100 fold lower than for the weight matrix. This increased discrimination makes the belief network more useful in database search.

### Amino Acid Interactions in $\alpha$ -Helices

We examined the protein sequences of 234  $\alpha$ -helical segments taken from a unique subset

## Automatic Discovery of Protein Motifs Ninth International ACS Biotechnology Symposium

of the Brookhaven Structure Library. This subset contained only the highest resolution member of each protein superfamily. We initially looked for sequence correlations in eight amino acid segments (two complete  $\alpha$ -helical turns) contained within these segments.



**Figure 2.** This belief network displays some of the highly significant correlations observed between the amino acid side-chains in  $\alpha$ -helical segments in the Brookhaven Database. The network is depicted as a helical wheel. The dotted lines show the path of the peptide main chain. The solid arrows around the circumference show the strong correlations observed between amino acid side chains that are adjacent in space. Finally the solid spokes represent the dependence of the amino acid composition on the helical nature of the sequence.

Figure 2 shows half of the highly significant correlations that were observed within  $\alpha$ -helical segments. This belief network is displayed as a helical wheel, with the dotted line showing the path of the protein peptide backbone and the solid lines showing the correlated positions. The arcs represented in Figure 2 indicated a strong dependence of amino acids at positions  $i+3$  and  $i+4$  on the amino acid at position  $i$ . The correlation indicated that if amino acid  $i$  were hydrophobic, then the amino acids at  $i+3$  and  $i+4$  were also more likely to be hydrophobic and more unlikely to be hydrophilic. This is the well-known

hydrophobic patch that characterizes many  $\alpha$ -helices.

The other half of the significant correlations not displayed here, shows a dependence of amino acids at  $i+2$  and  $i+5$  as well. The residues at these positions have hydrophobicity opposite that at position  $i$ . This observation is consistent with the amphipathic nature of most of the  $\alpha$ -helices in the Brookhaven database.

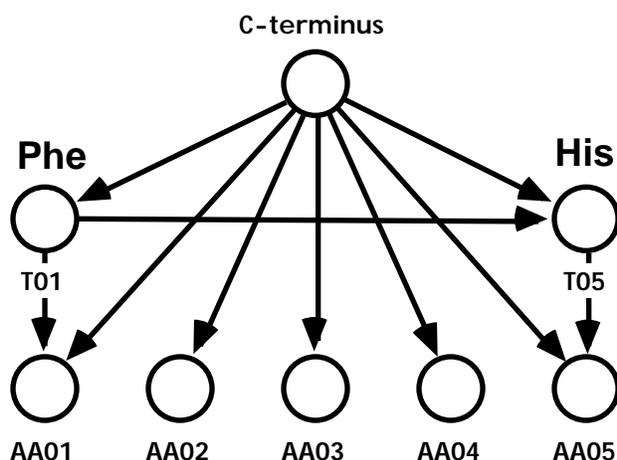
Similar networks have been made for  $\beta$ -strand sequences taken from the Brookhaven Structure database. They also suggest the amphipathic nature of  $\beta$ -strands.

While these observations do not provide any novel insights into secondary structure, they show that the probabilistic representation can rediscover well-known principles of protein structure. Moreover, the quantitative relationships contained in the belief networks should allow us to discriminate regions of secondary structure with more precision than previously possible.

### Phe-His Correlation at C-Termini of $\alpha$ -Helices

Finally, we searched for correlations between the amino acids found near the ends of  $\alpha$ -helices. We examined the first and last 5 amino acid residues from both the N- and C-terminal regions as determined by the DSSP program of Kabsch and Sander (1983). Most of the correlations we observed were of the types mentioned above for  $\alpha$ -helical segments in general. However, among the 234 C-terminal ends of  $\alpha$ -helices, we discovered a small but statistically significant correlation between an aromatic amino acid four residues before the terminus and a basic amino acid at the terminal position. When helices displaying this arrangement of amino acids at their C-terminus were examined, we discovered five occurrences of the amino acid pattern Phe-Xaa-Xaa-Xaa-His among the 234 helical segments examined (Figure 3). When we searched for the amino acid pattern above in all the sequences in the Brookhaven database it occurred at the C-terminal ends of  $\alpha$ -helices ( $\pm$  one residue) a total of twelve times. This pattern occurred in no other helical region.

**Automatic Discovery of Protein Motifs**  
**Ninth International ACS Biotechnology Symposium**



**Figure 3.** This belief network shows the observed correlation between amino acids within five residues from the C-terminus of  $\alpha$ -helices. Of the 234  $\alpha$ -helices examined in a high resolution unique subset of the Brookhaven Structural database, twelve displayed the pattern Phe-Xaa-Xaa-Xaa-His. This sequence pattern was found at no other position in  $\alpha$ -helical segments in the database.

The occurrence of the Phe-Xaa-Xaa-Xaa-His sequence has been known to stabilize  $\alpha$ -helical segments (Shoemaker *et al.*, 1990). The nature of the interaction appears to be one of an ring hydrogen of histidine interacts with the aromatic ring of phenylalanine. This interaction requires the histidine side-chain to bend towards the preceding phenylalanine. Our results suggest that this interaction may terminate  $\alpha$ -helical segments.

This final result shows that the search for correlations in protein sequences can lead to the discovery of novel side-chain side-chain interactions in the structures of proteins. The application of probabilistic inference using belief networks that represent general secondary structure elements such as  $\alpha$ -helices,  $\beta$ -strands, and C-terminal segments of  $\alpha$ -helices should lead to more accurate predictions of protein secondary structure.

### References

- Bairoch, A. *Nucleic Acids Res.*, **1991**, *19*, 2241-2245.
- Dodd, I. B.; Egan, J. B. *J. Mol. Biol.*, **1987**, *194*, 557-564.
- Gribskov, M.; McLachlan, A. D.; Eisenberg, D. *Proc. Natl. Acad. Sci. USA*, **1987**, *84*, 4355-4358.
- Kabsch, W.; Sander, C. *Biopolymers*, **1983**, *22*, 2577-637.
- Lauritzen, S. L.; Spiegelhalter, D. J. *Journal of the Royal Statistical Society*, **1988**, *50 B*, 157-224.
- Neapolitan, R. E. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*; John Wiley and Sons: New York, NY, **1990**;
- Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann Publishers, Inc.: San Mateo, CA, **1988**;
- Schwartz, R. M.; Dayhoff, M. O. *Atlas of Protein Structure*, **1979**, *5*, 353-358.
- Shoemaker, K. R.; Fairman, R.; Schultz, D. A.; Robertson, A. D.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Biopolymers*, **1990**, *29*, 1-11.
- Staden, R. *Nucleic Acids Research*, **1984**, *12*, 505-519.