Protein Science (1994), *3*:1847–1857. Cambridge University Press. Printed in the USA.

Discovering structural correlations in α-helices

TOD M. KLINGLER AND DOUGLAS L. BRUTLAG

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

(RECEIVED March 15, 1994; ACCEPTED July 22, 1994)

Abstract

We have developed a new representation for structural and functional motifs in protein sequences based on correlations between pairs of amino acids and applied it to α-helical and β-sheet sequences. Existing probabilistic methods for representing and analyzing protein sequences have traditionally assumed conditional independence of evidence. In other words, amino acids are assumed to have no effect on each other. However, analyses of protein structures have repeatedly demonstrated the importance of interactions between amino acids in conferring both structure and function. Using Bayesian networks, we are able to model the relationships between amino acids at distinct positions in a protein sequence in addition to the amino acid distributions at each position. We have also developed an automated program for discovering sequence correlations using standard statistical tests and validation techniques. In this paper, we test this program on sequences from secondary structure motifs, namely α-helices and β-sheets. In each case, the correlations our program discovers correspond well with known physical and chemical interactions between amino acids in structures. Furthermore, we show that, using different chemical alphabets for the amino acids, we discover structural relationships based on the same chemical principle used in constructing the alphabet. This new representation of 3-dimensional features in protein motifs, such as those arising from structural or functional constraints on the sequence, can be used to improve sequence analysis tools including pattern analysis and database search.

Keywords: α-helix structure; amino acid correlations; motif modeling; sequence analysis; side-chain interactions; structure analysis

Understanding the 3-dimensional structure of a protein is a necessary and critical step toward understanding the protein's function. For example, only after the structure of hemoglobin was solved was it possible to dissect the mechanisms responsible for the cooperative binding of oxygen, for the effects of pH and 2-3-diphosphoglycerate (DPG) on affinity, and for the defects causing various anemias (Stryer, 1988). Despite the increasing wealth of sequence data, the laborious and time-consuming process of empirical structure determination hampers the availability of detailed structural information. Instead, sequence analysis tools offer the best hope for quickly eliciting structural and functional information from new sequences.

Traditional methods for analyzing sequences rely on the prior analyses of known sequences and on procedures for matching sequences. These techniques encompass database search (Wilbur & Lipman, 1983), sequence classification (Klein et al., 1984; Klein & DeLisi, 1986), and analysis for motifs (Bairoch & Boeckmann, 1991; Henikoff & Henikoff, 1991), among others. Most techniques for both analysis and matching emphasize the conservation of amino acids during evolution. Specifically, one usually assumes that if 2 sequences are homologous, then the amino acids that one observes at corresponding locations in the 2 sequences are similar, where similarity refers to like physical or chemical properties.

A further assumption that is almost always made, in order to maintain computational feasibility, is one of conditional independence of amino acids in a sequence. In other words, the presence of an amino acid at one position in a protein is not affected by the presence of an amino acid at a different position. Intuitively, this assumption is troublesome because the forces that confer structure and function in a protein are mediated through specific amino acid interactions. Hydrogen bonds, electrostatic interactions, and Van der Waals forces are all interactions that require reciprocating chemical entities, and even the weakest of these side-chain–side-chain interactions play a part in determining a protein's structure (Burley & Petsko, 1988). From this, one should expect many 3-dimensional constraints between amino acids to be reflected in higher-order sequence patterns (correlations between amino acids at multiple positions). In our view, sequence differences at individual positions reflect simple evolutionary change, whereas correlated differences at multiple

Reprint requests to: Douglas L. Brutlag, Department of Biochemistry and Section on Medical Informatics, Stanford University School of Medicine, Stanford, California 94305-5307; e-mail: brutlag@cmgm.stanford.edu.

positions reflect structural constraints. This idea is further supported by comparing sequence alignments and true structural alignments (Bashford et al., 1987).

As mentioned previously, however, all of the popular sequence analysis methods, including weight matrices (Staden, 1984), consensus sequences (Bairoch & Boeckmann, 1991), profiles (Gribskov et al., 1987), blocks (Henikoff & Henikoff, 1991), and sequence alignment (Needleman & Wunsch, 1970), only model the distribution of amino acids at individual positions and not the effects imposed by amino acids at other positions. Although some secondary structure prediction algorithms (Lim, 1974; Chou & Fasman, 1978; Garnier et al., 1978; Levin et al., 1986) examine properties of neighboring residues (often within a window), none represent conditional dependencies between residues explicitly. Neural network approaches to structure prediction (Qian & Sejnowski, 1988; Stolorz et al., 1992) may model amino acid dependencies, but this has not yet been demonstrated. In this paper, we describe a method for detecting and representing higher-order sequence relationships that can be used to improve sequence classification and database search. Our starting point is a set of sequences representative of a motif (a specific structural or functional unit of a protein). From these sequences, we construct a probabilistic representation of the motif that can be used later to identify new examples of the motif. In addition, the patterns of dependencies in our representations can be used to make structural inferences.

The current state of affairs in sequence analysis is analogous to that faced by the designers of Bayesian medical diagnostic systems (Gorry & Barnett, 1968; de Dombal et al., 1972). The strict assumption of conditional independence was also made with these early systems: the probability of seeing any 1 symptom, given a specific disease, was assumed to be independent of the presence or absence of any other finding. Only with new probabilistic representations, such as Bayesian networks, have researchers been able to relax this often invalid assumption. There have been a few reports of covariation analysis in the sequence literature, including the analysis of surface loops in HIV coat proteins (Korber et al., 1993) and the analysis of tRNA sequences (Gutell et al., 1992). Both these works utilize a measure of mutual information to indicate structural interactions. However, neither addresses significance of covariation issues nor produces inference tools.

We use Bayesian networks (Pearl, 1988; Neapolitan, 1990) to represent conditional dependencies *between* positions in biological sequences in addition to the amino acid distributions at each position. Bayesain networks provide a graphical representation for sequence dependencies. They also provide a framework for storing the quantitative descriptions of both sequence distributions and correlations, simultaneously. Finally, these network models can be used for sequence classification, database search, and structure prediction using Bayesian network inference (Lauritzen & Spiegelhalter, 1988; Pearl, 1988; Neapolitan, 1990). We have written a program called MCSEQ that discovers correlations in a set of aligned sequences and produces Bayesian networks that can be used for classification and database search.

Results

α-Helix patterns

We extracted a training set of 3,157 overlapping amino acid sequences 8-residues long from 802 α-helices in the structure set described in the Materials and methods. A length of 8 residues was chosen in order to account for 2 complete turns in an α-helical conformation. Although overlapping sequences are used for training, each amino acid at a given position in one of the α-helices of the structure set is counted at most once for any amino acid node in the resulting network. Only amino acids near the helix termini are not represented in every node of the network. We found 19 of the 28 possible pairs of positions to have statistically significant correlations ($P < 0.01$) by both χ^2 and Monte Carlo tests. The network constructed from this training set using the general amino acid classification (HYDROPHOBIC, NEUTRAL, HYDROPHILIC) is shown in Figure 1. The amino acid nodes are arranged in the helical wheel pattern in order to better visualize the structural relationships. In addition, only the type nodes of this network are shown. Our program detected 2 types of correlations: (1) amino acids of like hydropathy tend to occur on the same side of a helix (Fig. 1A), and (2) amino acids of opposite hydropathy tend to occur on opposite sides of a helix (Fig. 1B). A summary of the conditional probabilities represented in each type of correlation arc is shown in Table 1. These data were calculated from 6,405 (i, $i + 2$), 5,686

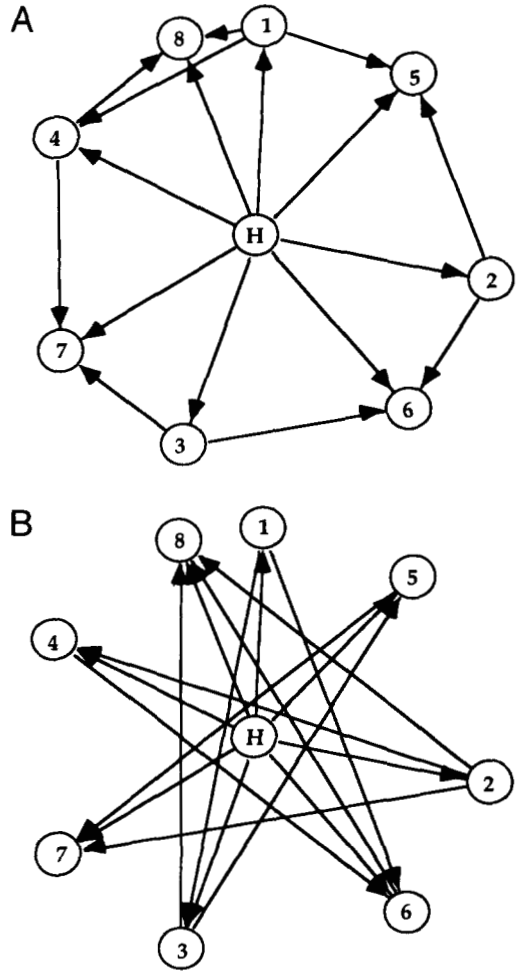


Fig. 1. A Bayesian network for amphipathic α-helices. **A:** Correlations between amino acid types that lie on the same side of a helix. **B:** Correlations between amino acid types that lie on opposite sides of a helix.

Table 1. *Hydropathicity correlations in α-helices*

Position	Both phobic[a]	Odds[b]	Phobic-philic[a]	Odds[b]	Philic-phobic[a]	Odds[b]	Both philic[a]	Odds[b]
$(i, i+2)$	342 (568)	0.60	866 (650)	1.33	903 (677)	1.33	595 (773)	0.77
$(i, i+3)$	580 (520)	1.11	473 (553)	0.86	542 (627)	0.86	772 (666)	1.16
$(i, i+4)$	569 (431)	1.32	388 (495)	0.78	397 (528)	0.75	776 (607)	1.28
$(i, i+5)$	270 (370)	0.73	512 (418)	1.22	556 (461)	1.21	439 (519)	0.85

[a] Phobic = IVLFC, philic = HNQEDKR. Observed numbers are displayed followed by expected numbers in parentheses (based on individual position frequencies for each residue).
[b] Odds values are calculated by dividing the observed by the expected occurrences.

$(i, i+3)$, 4,967 $(i, i+4)$, and 4,321 $(i, i+5)$ relative position pairs in α-helices of the structure set. In general, these arcs are a probabilistic representation of the hydrophobic periodicity of amino acids seen in amphipathic α-helices: pairs of positions that are on the same side of an α-helix, $(i, i+3)$ and $(i, i+4)$, prefer like types of residues, whereas pairs of positions that are on opposite sides of an α-helix, $(i, i+2)$ and $(i, i+5)$, tend to have residues of opposite hydropathy.

β-Sheet patterns

From the structure set described in Materials and methods, a training set of 2,349 overlapping amino acid sequences 4 residues long from 316 β-sheets was also constructed. A length of 4 residues was chosen in order to account for 2 complete pleats in the extended conformation. The network constructed from this training set, also using the general amino acid classification, is shown in Figure 2. The amino acid nodes are arranged like the C_α atoms of an extended chain. As before, only the type nodes of this network are shown. The correlation arcs found for this network, as expected, were also of 2 types: (1) amino acids of the same class tend to fall on the same side of a sheet, and (2) amino acids of opposite polarity tend to occur on opposite sides of a sheet. A summary of the conditional probabilities represented in each type of correlation arc is shown in Table 2. Again, this pattern of correlations represents the possible amphipathicity of β-sheets: position pairs $(i, i+2)$ are on the same side of a β-sheet and prefer like types of residues, whereas position pairs $(i, i+1)$ and $(i, i+3)$, which are on opposite sides of an β-sheet, tend to have residues of opposite hydropathy. Monte Carlo simulation data for this data set also fit closely to the expected χ^2 distribution, thus confirming the significance of the correlation arcs (at $P < 0.01$).

Phe-His bridge in C-termini of α-helices

In a more specific study of secondary structure, we applied our method to the carboxyl-terminal sequences of 802 α-helices described in Materials and methods. A length of 5 amino acids was used to account for just over 1 complete helical turn. From an initial correlational analysis using the general amino acid classification, which reflected many of the helical correlations noted above, we were able to successively refine our amino acid classifications until we had isolated the main determinants of a unique dependency. Our final amino acid classification was the specific one (PHE, HIS, OTHER) and yielded a single significant correlation corresponding to a recently investigated stabilizing helical interaction (Shoemaker et al., 1990). This interaction, a phenylalanine-histidine bridge, has been shown to stabilize the carboxyl-terminus of an α-helix peptide under controlled laboratory conditions. In our studies, we have found that this interaction occurs exclusively at the C-terminus of α-helices in our data set. The interaction is found in 1 orientation only, and is reflected in sequence data as well: there is a significantly higher number of Phe-Xaa-Xaa-Xaa-His sequences (7,879) than His-Xaa-Xaa-Xaa-Phe sequences (7,657) in the Swiss-Prot Release 27 sequence database (standard deviation ≈ 89). These latter observations are important for assessing the significance of the interaction as an α-helix termination signal and were not immediately obvious in the experimental studies. Figure 3 shows su-

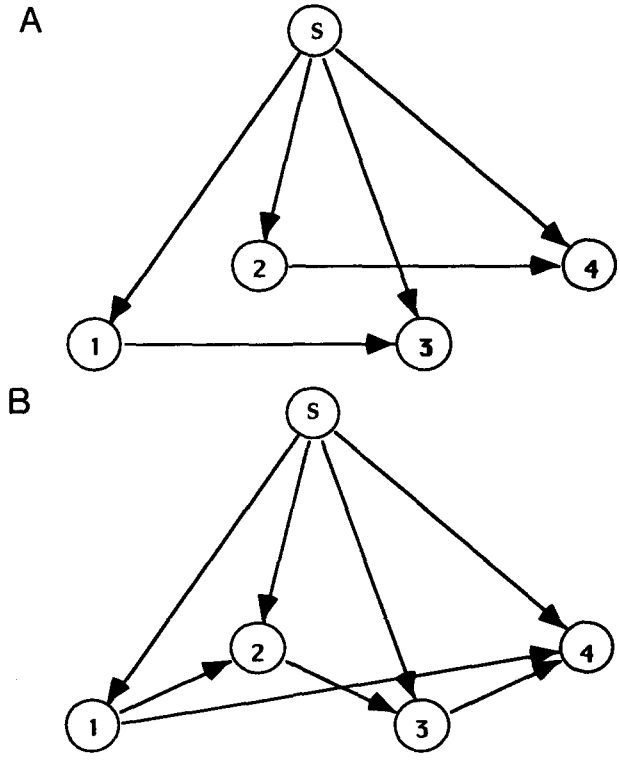


Fig. 2. A Bayesian network for amphipathic β-sheets. **A:** Correlations between amino acids that lie on the same side of a sheet. **B:** Correlations between amino acids that lie on opposite sides of a sheet.

Table 2. *Hydropathicity correlations for β-strands*

Position	Both phobic[a]	Odds[b]	Phobic–philic[a]	Odds[b]	Philic–phobic[a]	Odds[b]	Both philic[a]	Odds[b]
$(i, i+1)$	628 (728)	0.86	464 (382)	1.21	462 (384)	1.20	168 (201)	0.84
$(i, i+2)$	567 (520)	1.09	274 (287)	0.95	279 (280)	1.00	168 (155)	1.08
$(i, i+3)$	313 (352)	0.89	236 (205)	1.15	258 (217)	1.19	110 (127)	0.87

[a] Phobic = IVLFC, philic = HNQEDKR. Observed numbers are displayed followed by expected numbers in parentheses (based on individual position frequencies for each residue).
[b] Odds values are calculated by dividing the observed by the expected occurrences.

perimposed interacting Phe–His pairs and the network for this interaction. Note that the AA_i nodes in this network are shown, and that this network differs from the previous examples in the use of the interaction-specific amino acid classification.

Amino acid correlations in α-helices

Finally, we analyzed specific positions in α-helical sequences for ungrouped amino acid correlations. First, training sets composed of 7,124 $(i, i+1)$ pairs, 6,405 $(i, i+2)$ pairs, 5,686 $(i, i+3)$ pairs, 4,967 $(i, i+4)$ pairs, and 4,321 $(i, i+5)$ pairs were generated from the structure set described in Materials and methods. Because of the large number of helical segments in these training sets, we can examine all 400 (20 amino acids × 20 amino acids) unique residue pairs for significant correlations. Thus, for each of the 4 position pairs, correlations between each amino acid in each position were evaluated. For example, the contingency table for the most significant correlation found among the 5 training sets is:

i vs. $i+4$	Aspartate	Not Asp
Lysine	33 (11.8)	250 (271)
Not Lys	172 (193)	4,456 (4,435)

$\chi^2 = 42.2$, $P < 10^{-5}$, odds = 2.79

For this interaction, the pair lysine-aspartate was found 33 times in the structure set separated by 3 intervening amino acids. Based on the frequencies of these 2 amino acids in their respective positions, the expected number for each pair was calculated and shown in parentheses for each entry. For the pair lysine-aspartate, fewer than 12 occurrences were expected, giving a χ^2 value of 42.2 ($P < 10^{-5}$). For each of the 5 training sets, all possible pairs of amino acids were evaluated for over- and underrepresentation. A χ^2 test was run only if there was enough data to do so accurately (i.e., there was an expected number of at least 5 in all bins). For all sets of pairs, between 250 and 300 of the possible 400 pairs were evaluated with the χ^2 test under this criterion. Because multiple tests were run on data derived from the same population, we used the Bonferroni inequality (Snedecor & Cochran, 1989) to calculate an individual test significance threshold of $P < 0.0002$, giving an overall a posteriori significance of better than 0.05 (if we used a significance of 0.05 for each test, then by chance alone, 15 of 300 residue pairs would be judged significant). Table 3 lists the most significant pairs for the $(i, i+4)$ and $(i, i+3)$ sequence interactions with a dotted line drawn at $P < 0.0002$ significance.

Several possible explanations for amino acid correlations in α-helices are possible. First, they may reflect the amphipathic patterns in helices. We have already detected amphipathic correlations by grouping amino acids, so we might expect individual amino acids to show similar position preferences. However, the statistical significance of the individual amino acid pairs (Table 3) exceeds the significance of the amphipathic bias (Table 1). Hence, we examined the structural conformations of these pairs

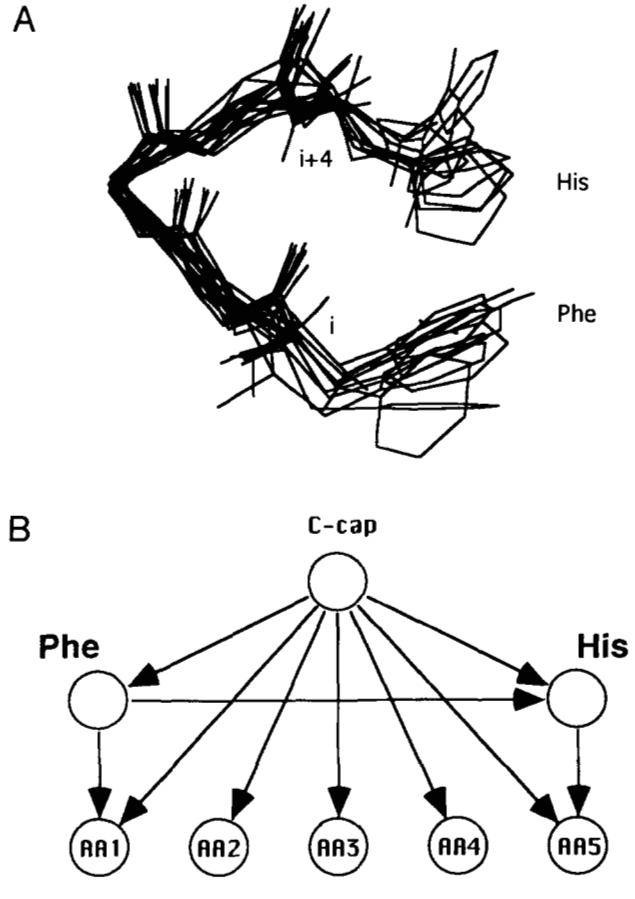


Fig. 3. Interacting Phe–His structures (**A**) and corresponding probabilistic network (**B**). A Bayesian network for the Phe–His bridge for the correlation between amino acids at the C-termini of α-helices. Amino acid 5 (AA_5) is the C-terminal residue of the helix.

Table 3. *(i, i + 4) and (i, i + 3) sequence interactions*

A. (*i*, *i* + 4) sequence correlations

Pair	Observed[a]	Expected[b]	χ^{2c}	Odds[d]	More/less[e]
KD	33	11.8	42.1	2.79	+
KE	42	20	27.6	2.10	+
LL	97	62.1	25.0	1.56	+
EK	55	30.4	23.4	1.81	+
FM	17	6.15	20.6	2.76	+
IL	60	37.9	15.8	1.58	+
QE	32	17.3	14.1	1.85	+
KL	16	36.1	13.6	0.44	−
SA	47	29.3	13.0	1.61	+
GA	43	27.8	10.1	1.55	+
PF	13	5.68	10.1	2.29	+

B. (*i*, *i* + 3) sequence correlations

Pair	Observed[a]	Expected[b]	χ^{2c}	Odds[d]	More/less[e]
DR	36	18.6	18.4	1.94	+
LI	56	37.2	11.3	1.50	+
VA	73	51.9	10.6	1.41	+

[a] The number of observed pairs in the data.
[b] The number of expected pairs based on the individual frequencies of the residues in each position (i.e., K in position *i* and D in position *i* + 4 in the first row), respectively.
[c] The χ^2 value for the corresponding 2 × 2 contingency table (all have significances $P < 0.05$; the top 5 in **A** have significances $P < 10^{-4}$).
[d] The odds obtained by dividing the observed number of occurrences by the expected number.
[e] A "+" if the sequence pair is overrepresented (odds > 1) and a "−" if the sequence pair is underrepresented.

to test for specific side-chain–side-chain interactions. We hypothesized that overrepresented pairs reflect specific side-chain–side-chain interactions. Pairs of side chains can interact in an α-helix when they are in the (*i*, *i* + 4), (*i*, *i* + 3), and (*i*, *i* + 1) arrangements. However, to form an interaction, the amino acid side chains are constrained to a subset of their possible rotamer conformations (McGregor et al., 1987; Ponder & Richards, 1987; Creamer & Rose, 1992; Pickett & Sternberg, 1993), particularly at the χ_1 angle (the dihedral angle for the bond between C_α and C_β). In α-helices, side-chain contacts between positions *i* and *i* + 4 are most likely when the χ_1 angle at position *i* is *trans* and the χ_1 at position *i* + 4 is *gauche*+, whereas side-chain contacts between positions *i* and *i* + 3 are most likely when the χ_1 angle at position *i* is *gauche*+ and the χ_1 at position *i* + 3 is *gauche*+. Therefore, we compared the rotamer frequencies at χ_1 for the amino acids involved in highly overrepresented pairs and the rotamer frequencies for those amino acids anywhere in a helix.

The rotamer frequencies are obtained by partitioning all side-chain rotamers into distinct classes based on χ_1. The side-chain dihedral angle χ_1, ranges from −180° to 180°, with classes defined as follows: *trans* ($\chi > 120°$ and $\chi \leq -120°$), *gauche*+ ($-120° < \chi \leq 0°$), and *gauche*− ($0° < \chi \leq 120°$). The preferred χ_1 angles for (*i*, *i* + 3), and (*i*, *i* + 4) pairs are shown in Figure 4. Continuing the example from above, 23 of the 33 lysines at position *i* in (*i*, *i* + 4) lysine-aspartate pairs have *trans* χ_1 angles compared to 16 expected (based on all helical lysine χ_1 angles). Likewise, 31 of the 33 aspartates at position *i* + 4 in (*i*, *i* + 4) lysine-aspartate pairs have *gauche*+ χ_1 angles compared to 25 expected. Both these discrepancies are significant at the $P < 0.05$ level.

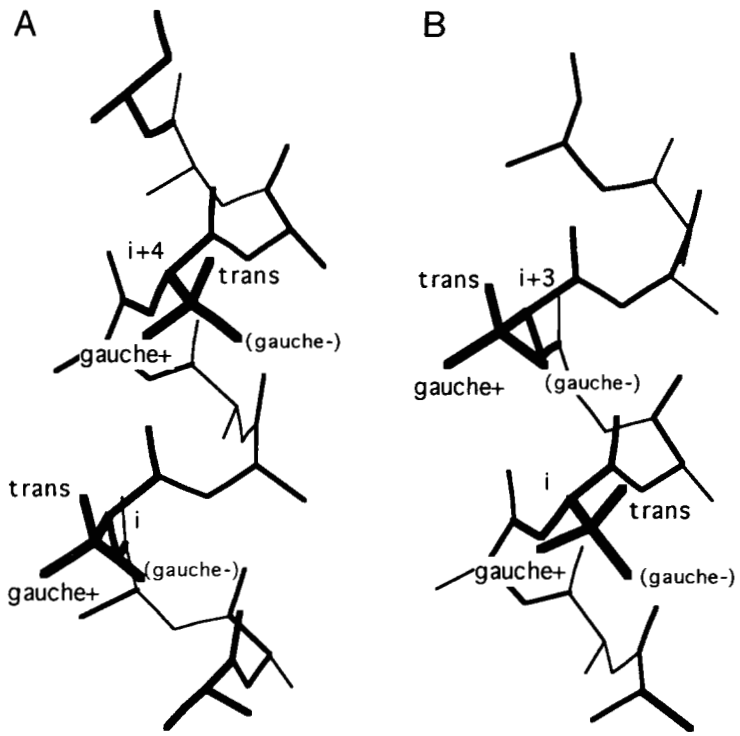
For each of the most significant sequence correlations of the previous analysis where each member has C_β atoms, an analysis of χ_1 angles was done to determine if structural interactions were responsible for the sequence correlations. The distribution of χ_1 for all side chains in our structure set is 32.1% *trans*, 50.4% *gauche*+, and 15.2% *gauche*−. In α-helices, the *gauche*− conformation at χ_1 is rare because of steric constraints: 38.5% *trans*, 54.7% *gauche*+, and 6.8% *gauche*−. Further, there are characteristic, preferred χ_1 conformations for each of the amino acids, again because of the specific steric properties of each side chain. For example, valine, isoleucine, and threonine side chains, with their branched C_β's, are constrained to *trans* and *gauche*+ more frequently than for the other amino acids. In the analysis below, expected side-chain χ_1 angles are calculated from these amino acid-specific distributions. Table 4 lists the pertinent χ_1 angles for the most strongly correlated residue pairs.

Discussion

Using the technique of Bayesian networks, we have represented, in a general way, pairwise dependencies in protein sequences. Traditional methods for sequence analysis model evolutionary variation at individual positions in a protein. Our system adds second-order structural information, represented as correlations between pairs of amino acids, and thereby increases the informational power of motif representations. As a first approximation, we believe that many of the higher-order interactions seen in biological sequences can be decomposed into pairwise correlations. If this assumption is true, then we can claim a high degree of representational power and generality. In other words, our system can discover and represent most important structural interactions if such interactions can adequately be described with, at most, second-order relationships, and we have large enough data sets. Restricting our system to first- and second-order information also simplifies the computation required both for discovering motif patterns and for evaluating query sequences.

Our work takes advantage of the benefits afforded by the Bayesian network framework—we represent the sequence relationships in motifs, as mediated through structural interactions, both graphically and quantitatively. The graphical representation of motifs (in Figs. 1, 2, 3) gives a qualitative, spatial, and intuitive understanding of the relationships in protein sequences. The benefit of presenting motifs in this form is especially evident with the secondary structure motifs, where the arrangement of the networks mimic the actual structures, and the meaning of the arcs is easily summarized. The quantitative representation of motifs is contained in the set of conditional probability tables associated with each node in a network. These numbers are readily interpreted as probabilities or likelihoods, and can be evaluated for significance and strength with standard techniques.

In this paper, we have presented 3 sets of protein sequences, each of different complexity, for which we have built networks. The number of sequences in the data sets for these applications is quite large, occasionally approaching 10,000 examples. Thus, we have the greatest confidence in our probability estimates and



i and *i*+4 χ_1 angles. *i* and *i*+3 χ_1 angles.

Fig. 4. χ_1 Angle diagrams for $(i, i + 3)$ and $(i, i + 4)$ interactions. The preferred χ_1 angles for $(i, i + 4)$ contacts are *trans/gauche+*. The preferred χ_1 angles for $(i, i + 3)$ contacts are *gauche+/gauche+* (χ_1 angles in the *trans/gauche−* orientations, respectively, could form a contact except that *gauche−* χ_1 angles are very rare in α-helices because of steric hindrances).

Table 4. *Structural interactions between side chains for sequence pairs in Table 3*

A. $(i, i + 4)$ structural interactions

Pair	Number	*i trans* obs.[a]	*i trans* exp.[b]	Sig.[c]	*i* + 4 *gauche+* obs.[d]	*i* + 4 *gauche+* exp.[e]	Sig.[f]
KD	33	23	16.1	<0.01	31	25.0	<0.05
KE	42	28	20.5	<0.02	35	19.2	<0.005
LL	97	61	39.3	<0.005	71	57.2	<0.025
EK	55	27	19.0	<0.05	24	25.1	–
FM	17	14	10.3	<0.05	15	11.4	<0.05
IL	60	9	5.8	–	45	35.4	<0.05
QE	32	21	12.6	<0.005	25	18.8	<0.01
KL	16	8	7.8	–	13	9.4	–
LI	46	34	27.1	<0.05	43	39.1	–

B. $(i, i + 3)$ structural interactions

Pair	Number	*i gauche+* obs.[a]	*i gauche+* exp.[b]	Sig.[c]	*i* + 3 *gauche+* obs.[d]	*i* + 3 *gauche+* exp.[e]	Sig.[f]
DR	36	27	27.3	–	9	15.6	<0.025
LI	56	36	33.0	–	45	47.5	–
TN	17	16	12.2	<0.05	12	13.0	–
DL	22	14	16.7	–	11	13.0	–
LD	9	6	3.6	–	7	6.8	–

[a] The observed number of first residue χ_1 angles in the predominant orientation for an interaction between the residue pair.
[b] The expected number of first residue χ_1 angles in the same orientation (i.e., from the χ_1 frequencies for the first residue anywhere in a helix).
[c] The significance of the first residue χ_1 angles calculated by the χ^2 statistic.
[d] The observed number of second residue χ_1 angles in the predominant orientation for an interaction between the residue pair.
[e] The expected number of second residue χ_1 angles in the same orientation (i.e., from the χ_1 frequencies for the second residue anywhere in a helix).
[f] The significance of the second residue χ_1 angles calculated by the χ^2 statistic.

statistical tests in this domain. Data set sizes can be a problem when applying this techniques to more specific motifs, where the number of known examples often falls below 100.

The first application is the discovery of the hydropathy relationships in α-helix and β-sheet data. Figure 5 shows the helical wheel representation of an α-helix, which diagrams the arrangement of amino acids in a helix when viewed on end. The nodes in our network for the α-helix are arranged in this manner in order to understand the discovered correlations. The correlations shown in Figure 1A represent interactions between an amino acid at a position i and the amino acids at positions $i + 3$ and $i + 4$. These amino acids are adjacent to each other in a helix, as seen in the helical wheel diagram. The dependence between these positions shows a preference for similar types of amino acids. Analogously, Figure 1B shows interactions between an amino acid at a position i and the amino acids at positions $i + 2$ and $i + 5$. These amino acid pairs are on opposite sides of the helix and show a preference for dissimilar amino acid types. Specifically, if a hydrophobic residue is present at position i, then there is a greater chance of seeing hydrophobic residues at $i + 3$ and $i + 4$ than one would expect based on the frequencies of amino acids in helices. There is also a greater chance of seeing hydrophilic residues at $i + 2$ and $i + 5$. These correlations correspond to the hydrophobic periodicity that characterizes many amphipathic α-helices and are summarized in Table 1. In fact, even the relative over- or underrepresentation, as measured by the odds for hydrophobic–hydrophobic and hydrophilic–hydrophilic pairs, mirrors the hydropathic moment vector for the 2 positions (see Fig. 6). In other words, the more coincident 2 residues are on one side of an α-helix, the more likely they are to be of the *same* hydropathy. Conversely, the closer to a 180° separation 2 residues are, the more likely they are to be of *opposite* hydropathies. We believe the α-helix network validates our approach because it automatically rediscovers an important principle of protein structure.

Similarly, the β-sheet correlations shown in Figure 2 are of 2 types depending on the relative positions of amino acid pairs. Amino acids positioned on the same side of a sheet (i and $i + 2$) are more likely to be of the same type, whereas amino acids positioned on opposite sides of a sheet (i and $i + 1$, or i and $i + 3$) are more likely to be of different hydropathy. The difference in the spacing of correlations found in α-helices and β-sheets can have a large effect on the prediction of secondary structure using Bayesian networks. In this case, the pattern of correlations can be more distinctive than amino acid distributions alone (because the frequencies of amino acids found in α-helices and β-sheets is quite similar). In addition, the strengths of the β-sheet correlations again appear to parallel the magnitude of the hydropathic moment calculated for each pair of positions (Eisenberg et al., 1984) in a β-strand.

The second, and related, application is the C-terminus sequences from α-helices. The data set size for this application is smaller than the previous example, but in the range of other sequence data sets we've worked with successfully. This application demonstrates the versatility of our system in moving from abstract amino acid classifications to detailed interaction-specific classifications. With the initial amino acid classes, we largely saw general helix dependencies, whereas later refinement brought out a unique interaction. This unique interaction, the Phe–His bridge, has been shown to stabilize model α-helical peptides in solution (Shoemaker et al., 1990). Further analysis of the entire Brookhaven database confirmed that the pattern Phe-Xaa-Xaa-Xaa-His is seen at the C-terminus of a helix in 10 of the 11 occurrences

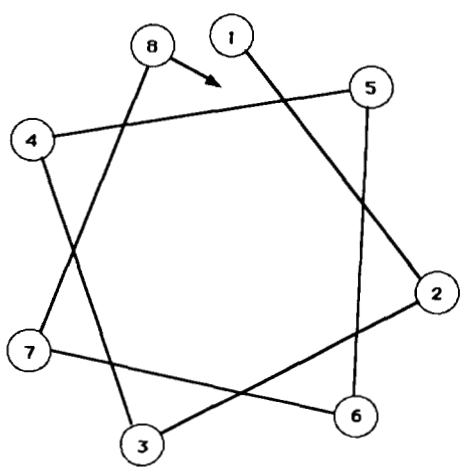


Fig. 5. The helical wheel. Arrangement of amino acids in an α-helix as viewed from the N-terminus. Network for α-helical segments in Figure 5 has amino acid nodes arranged in this manner.

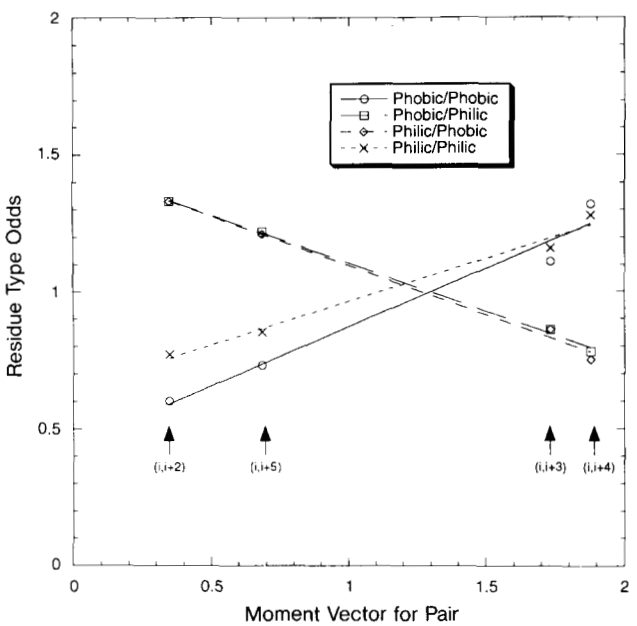


Fig. 6. Graph of residue pair odds from Table 1 against the hydropathic moment vector calculated for each pair of positions. As a measure for how coincident 2 residues are on 1 side of an α-helix, we calculated the hydropathic moment vector for the position pairs (i, $i + 2$), (i, $i + 3$), (i, $i + 4$), and (i, $i + 5$). The hydropathic moment is the magnitude of the sum of the unit vectors pointing out from the helical axis to the C_α atoms in the positions pair. This measure is based on the hydrophobic moment calculation (Eisenberg et al., 1984). The hydropathic moments for the 4 pairs listed above are 0.347, 1.732, 1.879, and 0.684, respectively, indicating that (i, $i + 4$) are the closest in an α-helical turn and (i, $i + 2$) are the most distant. This agrees with what one would expect from the helical wheel diagram (Fig. 5).

of the pattern in helices (Fig. 3A). Conversely, the pattern His-Xaa-Xaa-Xaa-Phe in helices is positioned only 1 of 9 times near the C-terminus. Although this interaction is again validated by empirical evidence, this application shows that the detection of amino acid correlations can indicate novel structural interactions.

The third application is the specific side-chain–side-chain interactions. Many of the sequence correlations in Table 3 represent specific side-chain–side-chain interactions that have been shown to stabilize α-helices (Marqusee et al., 1989; Shoemaker et al., 1990; Armstrong & Baldwin, 1993; Huyghues-Despointes & Baldwin, personal communication; Padmanabhan & Baldwin, personal communication). Others are novel and may reflect unexamined side-chain interactions important in local protein stability. In order to separate side-chain interactions from nonspecific hydropathy-based correlations, we examined the side-chain conformations for the significant sequence correlations.

Almost all of the highly significant $(i, i + 4)$ sequence correlations (Table 3) correspond to specific side-chain conformation (Table 4). For example, not only is the lysine-aspartate pair overrepresented at $(i, i + 4)$ in α-helices, but their χ_1 angles are significantly skewed toward the orientation preferred for side-chain interaction (the χ_1 for lysines at position i are more often *trans* than expected, and the χ_1 for aspartates at position $i + 4$ are more often *gauche*+ than expected).

The interaction of sequence pairs KD, KE, and EK are electrostatic, as one would expect. The interaction of sequence pair QE is likely a hydrogen bond. The interaction of sequence pairs LL, IL, and LI is hydrophobic in nature. One should note that all of these interactions are stronger (more significant) than the more general helical amphipathic patterns discussed above. In contrast, because the sequence pair KL shows no side-chain χ_1 preferences and is not as significant, its underrepresentation in helices may result from a hydrophobic effect alone (i.e., a strong hydrophobic residue and a strong hydrophilic residue are underrepresented in positions that would place them next to each other on a helix).

Only 2 of the 5 significant $(i, i + 3)$ sequence correlations show contact preference (as determined by χ_1 analysis): the salt bridge, DR, and the potential hydrogen bond, TN. The remaining 3 sequence correlations correspond to nonspecific interactions such as hydrophobic interactions or amphipathicity. Lastly, except in a few cases, it is difficult to analyze the conformations of underrepresented sequence pairs because the number of examples is small. If the underrepresentation is the result of an unfavorable interaction, one might expect that the side chains would point away from each other. Alternatively, as with KL in $(i, i + 4)$ above, nonspecific amphipathic patterns might be responsible, giving no side-chain conformation preferences.

Most of these interactions have been shown to contribute to a proteins' stability (Stryer, 1988). Even more specifically, some have been shown to stabilize α-helices (Marqusee et al., 1989; Shoemaker et al., 1990; Armstrong & Baldwin, 1993; Huyghues-Despointes & Baldwin, personal communication; Padmanabhan & Baldwin, personal communication). However, one of our highly significant correlations, namely the phenylalanine-methionine pair in $(i, i + 4)$, shows contact preference in χ_1 angles and has little mention in the literature (Reid et al., 1985; Burley & Petsko, 1988) and no experimental confirmation. When the 17 $(i, i + 4)$ phenylalanine-methionine pairs are superimposed (Fig. 7), one sees a regularity in side-chain interaction. We propose that this side-chain interaction, the sulfur-aromatic, may play a role in stabilizing proteins, particularly α-helices. Using Iditis, we have found 65 examples of contacting (one atom from each side chain within 5 Å of each other and regardless of sequence positions) phenylalanine-methionines in the unique structure set described in the Materials and methods section.

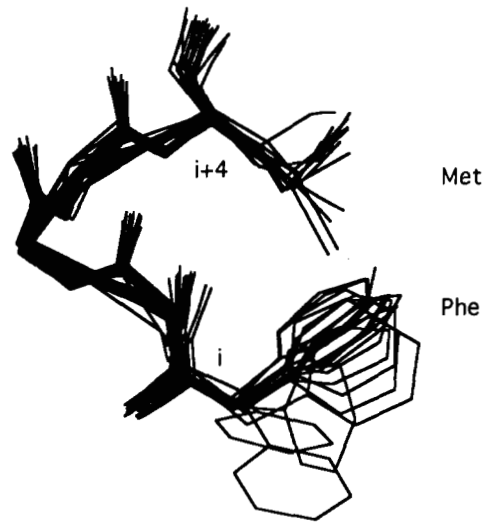


Fig. 7. Superimposed phenylalanine-methionine interactions. Seventeen examples of phenylalanine-methionine pairs in $(i, i + 4)$ sequence positions superimposed with backbone atom coordinates only.

In this paper, we have described correlations using discrete values only. In an immediate extension to this work, we are examining correlations in protein sequences based on numerical parameters. For example, one can run regression analyses on hydrophobicity values, side-chain volumes, etc., to generate quantitative relationships. We are currently exploring the secondary structure prediction problem by constructing networks for many different secondary structure units, scoring test sequences, and developing methods for comparing their scores.

Materials and methods

The method described in this paper was developed as a generalization of the traditional weight matrix method for representing sequence motifs and for searching biological sequence databases for related functional or structural motifs (Staden, 1984). A weight matrix is a table of likelihoods calculated by dividing the positional amino acid frequencies in a set of related aligned sequences (the "motif") by the amino acid frequencies in a representative protein database. The logarithms of these weights, or log likelihoods, are then used to evaluate a query protein sequence by summing the corresponding values for each amino acid in the query. An arbitrary threshold on these sums is used to classify the query sequence (as an example of the motif or not). Summing the logarithms of the weights is equivalent to multiplying likelihoods, so the final weight matrix score for a query sequence can be transformed to the likelihood that the sequence belongs to the training set class. The conditional-independence assumption is evident because the amino acid weights for each position in a motif are calculated separately and without regard for the distributions at any other position.

Bayesian networks

In this work, we employ Bayesian networks, or belief networks (Pearl, 1988; Neapolitan, 1990) to discover and represent spatial constraints (i.e., dependencies) imposed by a protein's structure. Bayesian networks are directed, acyclic graphs in which nodes represent variables and arcs represent the dependencies between variables. The dependencies are quantified in terms of conditional probabilities: if an arc exists from node A to node B, then there is a conditional probability function $P(B|A)$ over all possible values of A and B. Bayesian networks are explicit descriptions of the known relationships (dependencies as well as independences) among variables.

A Bayesian network with the simple topology shown in Figure 8 models amino acid distributions at independent positions. The center node C in this type of network represents the classification of a protein sequence and is usually binary (i.e., has values *example-of-motif* or *not-example-of-motif*). An AA_i node represents the amino acid at position i of the sequence. An arc from the center node C to an AA_i node represents the positional distribution of amino acids at position i in the training set (and amino acid frequencies in a randomly chosen sequence), and these are calculated from the frequencies of occurrence. These arcs encode the set of conditional probabilities $P(AA_i|C)$ for each position.

Whereas traditional approaches are limited by the assumption of conditional independence of sequence position, the network discussed above can be easily augmented to represent higher-order interactions. All pairwise constraints, or interactions, between residues at remote sites can be represented in a Bayesian network by simply adding arcs (and the appropriate conditional probability tables) of the form shown in Figure 9. An arc from one amino acid node to another represents a correlation between the pairs of amino acids occurring at the 2 respective positions in a set of sequences. For example, an electrostatic interaction in a specific motif would be represented as an arc with probabilities of the form $P(AA_i = \text{Asp or Glu}|AA_j = \text{Lys or Arg}, C = \text{motif})$.

In building up a network, we begin with no arcs and add them when dependencies between 2 nodes are detected statistically.

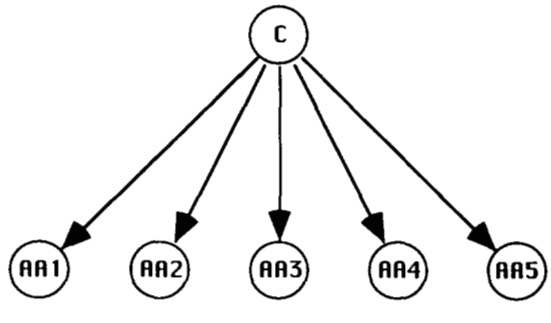


Fig. 8. A simple Bayesian network. The central node C is a classifier node, representing a variable with 2 settings: (1) a sequence is an example of *motif*, and (2) a sequence is not an example of *motif*. A leaf node AA_i, is an evidence node, and represents an amino acid variable for a single position in the motif. An arc from C to an AA_i node represents the distribution of amino acids occurring at that position in the motif. It encodes a conditional probability table containing probabilities $P(AA_i|C)$ for all combinations of both variables (i.e., $2 \times 20 = 40$ probabilities).

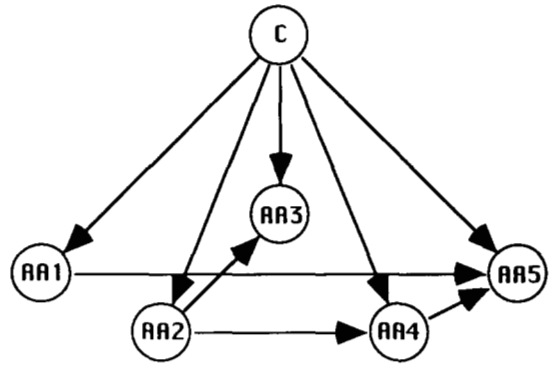


Fig. 9. A complex Bayesian network. An arc from one amino acid node, AA_i, to another, AA_j, represents the dependence of the amino acid at position j on the amino acid at position i and encodes $20 \times 20 = 400$ probabilities $P(AA_j|AA_iC)$ for each classifier value.

First, an arc from the center node C to an amino acid node AA_i is added if the amino acid distribution at position i of the input sequence deviates from the background distribution of amino acids in the sequence database (using the χ^2 test for comparing 2 distributions). This condition exists for every position in the networks described in this paper. If this condition fails, the position is not useful for discrimination by itself (but still may provide correlation information). When searching for dependencies between the amino acids at 2 positions in a motif, as described below, we are often limited by the number of canonical sequences we analyze (the training set size). Although the sequence databases are growing rapidly, training sets for specific motifs rarely contain more than several hundred examples, which is too small to detect significant correlations between amino acid distributions at pairs of positions. Analyzing the full amino acid dependencies at 2 positions in a motif requires the evaluation of a contingency table with 400 (20×20) bins. Adequately populating such a table requires a minimum of several thousand sequences—many more than are normally available. Additionally, if the amino acid distributions for a motif are skewed (which is common, especially for conserved positions), empty bins appear in the resulting contingency table.

However, by classifying the amino acids into chemical or functional classes, we can more easily detect significant correlations between *types* of amino acids because we look for dependencies between variables with a fewer number settings. For example, if an amino acid classification groups the 20 amino acids into 4 classes, we need only tabulate the occurrences of the 16 possible pairs of amino acid types at 2 positions. In this way, more accurate conditional probabilities can be generated from smaller training sets. We do, however, sacrifice some detail when we represent dependencies between more abstract variables. A Bayesian network representing dependencies between classes of amino acids is shown in Figure 10.

Amino acid alphabets

Because amino acid correlations are assumed to result from physical and chemical interactions, we can construct different amino acid classifications based on the properties of their side chains. A natural classification based on general physical and

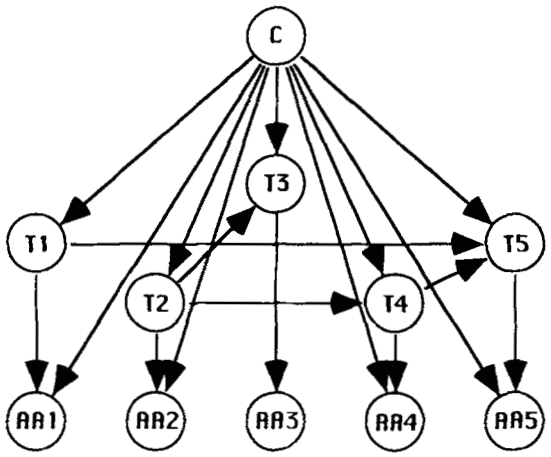


Fig. 10. A complex Bayesian network with amino acid classes. A T_i node represents an amino acid class for a single position in the motif. An arc from C to a T_i node represents the distribution of amino acid types occurring at that position in the motif. It encodes a conditional probability table containing the probabilities $P(T_i|C)$ for all combinations of both variables. An arc from a T_i node to its corresponding AA_i node completes the definitional cycle: if AA_i is known, then T_i is determined with probability 1 from an application of Bayes' Rule with the conditional probabilities $P(AA_i|C)$, $P(T_i|C)$, and $P(A_i|T_i)$. An arc from one amino acid type node, T_i, to another, T_j, represents a correlation of the amino acid type at position j on the amino acid type at position i, and encodes the probabilities $P(T_j|T_iC)$ for each type value.

chemical properties is: HYDROPHOBIC (Ile, Val, Leu, Phe, Cys); NEUTRAL (Met, Ala, Gly, Thr, Ser, Trp, Tyr, Pro); and HYDROPHILIC (His, Asn, Gln, Asp, Glu, Lys, Arg). These groups are defined using a standard hydropathy scale (Kyte & Doolittle, 1982) as follows: HYDROPHOBIC (hydropathy index > 2.0); HYDROPHILIC (hydropathy index < -2.0); and NEUTRAL ($2.0 \geq$ hydropathy index ≥ -2.0). A special-purpose classification that we use to detect a specific interaction between imino groups and aromatic rings is: PHE/TYR (Phe, Tyr); HIS (His); and OTHER (Ala, Cys, Asp, Glu, Gly, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp). These 2 examples of amino acid classifications represent 2 extremes: a general classification that is used to detect common interactions mediated through hydrophobic centers and electrostatic bridges, and a specific classification that detects a specific and unusual interaction.

Discovering sequence correlations

The discovery of positional dependencies for our Bayesian networks is accomplished with χ^2 statistical tests. Given the generic topology described above, arcs (and nodes) are included after rejecting null hypotheses about pairs of nodes. For arcs between the center node C and amino acid nodes AA_i (and T_i), the null hypothesis is that amino acids (and types) are distributed as in the sequence database. With all well-defined motifs we have examined, this hypothesis is rejected at high significance ($P < 0.001$) for every position in the motif. For T_i to T_j interpositional arcs, the null hypothesis is that the positions are uncorrelated, or conditionally independent. When the null hypothesis is rejected at some arbitrary significance level (usually $P < 0.01$), the corresponding arc is included in the network.

Table 5. *Unique, high-resolution Brookhaven database chains used in discovering sequence correlations*

1abk	1gpb	1phy	2csc	2trx B	4mdh B
1acp	1hfi	1prc C	2er7 E	2ts1	4mt2
1ada	1hip	1prc H	2fcr	2tsc B	4pfk
1ain	1hoe	1prc M	2gn5	2wrp R	4ptp
1ak3 B	1hrh B	1r69	2had	2yhx	4sgb I
1ald	1hsa D	1rbp	2hbg	3adk	4xia B
1aps	1hsa E	1rhd	2hmz D	3b5c	5abp
1atp E	1hsc	1rnb A	2ila	3bcl	5acn
1bbp D	1hsp	1rnh	2lh7	3cbh	5cpa
1c5a	1hyp	1rop A	2liv	3chy	5hvp B
1cbp	1ifb	1rsl C	2ltn D	3cla	5p21
1cc5	1lac	1sn3	2mcm	3dpa	5pti
1cd4	1lpe	1tfd	2ovo	3gap B	5rub B
1col B	1lz1	1tie	2pab B	3gf1	5rxn
1cox	1mad H	1tpk C	2pia	3grs	5tim B
1cpc K	1mad L	1ubq	2por	3il8	6ldh
1cpc L	1mbd	1utg	2rhe	3pgk	6tmn E
1cro O	1mbp	1wsy A	2rig	3rub S	7rsa
1cse E	1mle	1wsy B	2rsp B	3sdp B	8adh
1cse I	1msb B	1ycc	2sar B	3tgl	8atc C
1dhr	1nn2	256b B	2sc2 A	451c	8atc D
1eco	1nrd	2aaa	2sc2 B	4bp2	8cat B
1fkf	1nxb	2abd	2sdh B	4cpv	8dfr
1fnr	1ova D	2aza B	2sga	4enl	9icd
1fxd	1paz	2ca2	2sic I	4fgf	9pap
1gcr	1pec	2ccy B	2sns	4fxn	9rnt
1gd1 R	1pgx	2cdv	2sod Y	4i1b	9wga B
1gp1 B	1phh	2cpp	2stv	4lzm	

When T_i and T_j are correlated, arcs from the center node to each of these are included, as well as the arcs from T_i to AA_i and T_j to AA_j. As stated previously, these latter arcs are deterministic, functioning to define amino acid classes. Currently, our discovery program uses a straightforward exhaustive search of all pairs of positions. When significant amino acid type correlations are found, corresponding arcs are added to the developing network.

The significance of our χ^2 tests, especially with small or skewed data sets, is validated using Monte Carlo simulations. We iteratively construct simulated data sets by independently shuffling the amino acids found at each position in our original sequence alignments. This process preserves positional amino acid distributions while randomizing any pairwise correlations. Running the χ^2 tests on these simulated data sets gives an empirical estimate of how often significant pairwise correlations are detected due to chance alone. Arcs remain in a motif network only if significance is maintained in the Monte Carlo analysis.

Constructing sequence data sets

The sequences we analyze in this paper were extracted from a nonhomologous set of chains from the Brookhaven Protein Data Bank (Bernstein et al., 1977). To construct this set, we first eliminated all nonprotein structures, mutant structures, model structures, and low-resolution structures (>2.5 Å). Next, all pairwise sequence comparisons were made with the chains from

the remaining structures using the FASTDB program in the Intelligenetics Suite of sequence analysis programs. Chains were grouped such that, for every sequence in a specific group, there is at least one other sequence of greater than 30% identity in the same group (or, for every sequence in a given group, no sequence from a different group was better than 30% identical). Lastly, the chain with the best resolution was chosen from each group as the representative sequence for that group.

This procedure gave a high-resolution, nonhomologous, nonmutant, nonmodel structure set of 167 chains (Table 5). We used the Iditis program from Oxford Molecular (Thornton & Gardner, 1989), a program for querying the PDB in relational form, to extract sequences of specific secondary structure assigned by the extended DSSP method (Kabsch & Sander, 1983) implemented in Iditis.

Acknowledgments

This work is supported in part by the CAMIS grant from the National Library of Medicine LM05305 and in part by a seed grant from the Stanford Office of Technology Licensing. Tod M. Klingler is a predoctoral trainee of the National Library of Medicine.

References

Armstrong KM, Baldwin RL. 1993. Charged histidine affects alpha-helix stability at all positions in the helix by interacting with the backbone charges. *Proc Natl Acad Sci USA 90*:11337–11340.

Bairoch A, Boeckmann B. 1991. The SWISS-PROT Protein Sequence Data Bank. *Nucleic Acids Res 19*:2247–2249.

Bashford D, Chothia C, Lesk AM. 1987. Determinants of a protein fold: Unique features of the globin amino acid sequences. *J Mol Biol 196*: 199–216.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Burley SK, Petsko GA. 1988. Weakly polar interactions in proteins. *Adv Protein Chem 39*:125–189.

Chou PY, Fasman GD. 1978. Empirical predictions of protein conformation. *Annu Rev Biochem 47*:251–276.

Creamer TP, Rose GD. 1992. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci USA 89*:5937–5941.

de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Harrocks JC. 1972. Computer-aided diagnosis of acute abdominal pain. *Br Med J 2*:9–13.

Eisenberg D, Weiss RM, Terwilliger TC. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci USA 81*:140–144.

Garnier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting secondary structure of globular proteins. *J Mol Biol 120*:97–120.

Gorry GA, Barnett GO. 1968. Experience with a model of sequential diagnosis. *Comp Biomed Res 1*:490–507.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.

Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. 1992. Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res 20*:5785–5795.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res 19*:6565–6572.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Klein P, DeLisi C. 1986. Prediction of protein structural class from the amino acid sequence. *Biopolymers 25*:1659–1672.

Klein P, Kanehisa M, DeLisi C. 1984. Prediction of protein function from sequence properties. *Biochim Biophys Acta 787*:221–226.

Korber BT, Farber RM, Wolpert DH, Lapedes AS. 1993. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelop protein: An information theoretic analysis. *Proc Natl Acad Sci USA 90*:7176–7180.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol 157*:105–132.

Lauritzen SL, Spiegelhalter DJ. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *J R Stat Soc 50*:157–224.

Levin JM, Robson B, Garnier J. 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett 205*:303–308.

Lim VI. 1974. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J Mol Biol 88*:873–894.

Marqusee S, Robbins VH, Baldwin RL. 1989. Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci USA 86*:5286–5290.

McGregor MJ, Islam SA, Sternberg MJE. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol 198*:295–310.

Neapolitan RE. 1990. *Probabilistic reasoning in expert systems: Theory and algorithms*. New York: Wiley and Sons.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

Pearl J. 1988. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, California: Morgan Kaufmann Publishers, Inc.

Pickett SD, Sternberg MJE. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J Mol Biol 231*:825–839.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Qian N, Sejnowski TJ. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 202*:865–884.

Reid KSC, Lindley PF, Thornton JM. 1985. Sulphur-aromatic interactions in proteins. *FEBS Lett 190*:209–213.

Shoemaker KR, Fairman R, Schultz DA, Robertson AD, York EJ, Stewart JM, Baldwin RL. 1990. Side-chain interactions in the C-peptide helix: Phe 8^-–His 12^+. *Biopolymers 29*:1–11.

Snedecor GW, Cochran WG. 1989. *Statistical methods*. Ames, Iowa: Iowa State University Press.

Staden R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res 12*:505–519.

Stolorz P, Lapedes A, Xia Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J Mol Biol 225*:363–377.

Stryer L. 1988. *Oxygen-transporting proteins: Myoglobin and hemoglobin*. New York: W.H. Freeman and Company. pp 143–173.

Thornton JM, Gardner SP. 1989. Protein motifs and data-base searching. *Trends Biochem Sci 14*:300–304.

Wilbur WJ, Lipman DJ. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc Natl Acad Sci USA 80*:726–730.