Discovering Side-Chain Correlation in α-Helices

Tod M. Klingler and Dougles L. Brutlag

Department of Biochemistry and Section on Medical Informatics
Stanford Univeristy School of Medicine
Stanford, California 94305-53075klingler@cmgm.stanford.edu

klingler@cmgm.stanford.edu

brutlag@cmgm.stanford.edu

Abstract

Using a new representation for interactions in protein sequences based on correlations between pairs of amino acids, we have examined α-helical segments from known protein structures for important interactions. Traditional techniques for representing protein sequences usually make an explicit assumption of conditional independence of residues in the sequences. Protein structure analyses, however, have repeatedly demonstrated the importance of amino acid interactions for structural stability. We have developed an automated program for discovering sequence correlations in sets of aligned protein sequences using standard statistical tests and for representing them with Bayesian networks. In this paper, we demonstrate the power of our discovery program and representation by analyzing pairs of residues from α-helices. The sequence correlations we find represent physical and chemical interactions among amino-acid side chains in helical structures. Furthermore, these local interactions are likely to be important for stabilizing and packing α-helices. Lastly, we have also detect correlations in side-chain comformations that indicate important structural interactions but which don't appear as sequence correlations.

Introduction

There exists a discrepancy between the common understanding of protein structure and the representations used in protein sequence analysis. Protein structural analyses continually emphasize the importance of reciprocating physical and chemical interactions among two or more residues: salt bridges, hydrogen bonds, van der Waal's interactions, size constraints and the hydrophobic effect to list the most important. Sequence analysis techniques including database search (Wilbur & Lipman, 1983), sequence classification (Klein & DeLisi, 1986; Klein, Kanehisa & DeLisi, 1984) and analysis of motifs (Bairoch & Boeckmann, 1991; Henikoff & Henikoff, 1991), among others, almost always assume conditional independence of residues in a sequence for computational efficiency. In other words, the observation of an amino acid at a specific position in a sequence has no effect on any other amino acid position. Clearly this simplifying assumption is inconsistent with the universal understanding of interactions in protein structures.

Therefore, we have been interested in developing a representation for biological sequences that can incorporate structural features conferred through dependences among amino acids. We have used Bayesian networks (Neapolitan, 1990; Pearl, 1988) to relax the conditional independence assumption by explicitly representing correlations between pairs of residues in sequences: salt bridges are correlations between charged residues; hydrogen bonds, correlations between electron donors and acceptors; size constraints, correlations between large and small side chains. We have also been able to develop a discovery program for finding these and other correlations, and an inference program for searching databases with Bayesian networks. Thus, we have made a first step in bringing critical structural information in the form of correlations into the realm of sequence analysis.

In this paper we demonstrate the discovery and representation of amino acid correlations in α-helices. α-Helices, with β-sheets, comprise most of the secondary structure of most proteins. Over the more than 30 years researchers have been trying to predict the secondary structure of proteins, we have arguably only improved the prediction accuracy from about 60% to about 70%, which is still well below the level required for good structural inference for novel sequences. With the exception of tools that represent hydrophobic patches on α-helices, practically all automated prediction tools also make explicit assumptions of conditional independence. We therefore used our new discovery and representation capabilities to look for specific interactions between pairs of residues in α-helices, particularly in the relative (i, i+4) and (i, i+3) spacings, which bring residues into proximity after one turn of an α-helix.

Materials and Methods

For this work, we have used Bayesian networks, or belief networks, to discover and represent structural interactions in protein structures (Neapolitan, 1990; Pearl, 1988). Graphically, Bayesian networks are directed, acyclic graphs with nodes representing domain variables and arcs representing the dependences between domain variables. Computationally, a dependence-arc is a table of conditional probabilities, P(B | A), where A and B take on

all values of the source node *A* and the destination node *B* for the arc, respectively. Thus, Bayesian networks are descriptions of the dependence-relationships among domain variables expressed as conditional probabilities. Alternatively, and more correctly, Bayesian networks can be thought of as explicit representations of the *independences* in a joint probability distribution over all domain variables (where independences are designated by the absence of arcs).

Figure 1 shows a generic network for representing structural interactions in proteins. The center node *C* in this network represents the classification of a protein sequence—helix or sheet, for example. The AA_i nodes represent the amino acids at positions *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 sequence. These arcs encode the set of conditional probabilities $P(AA_i \mid C)$ for each position.

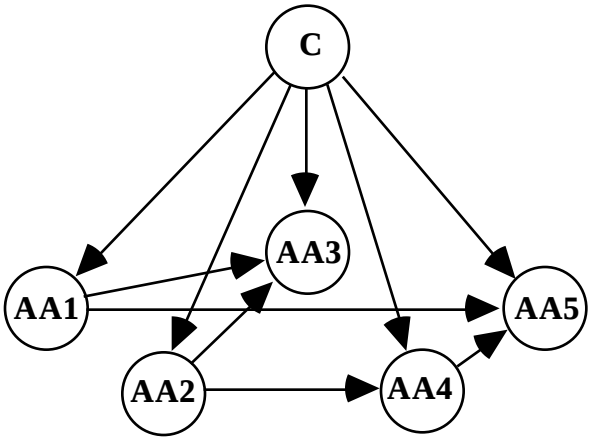


Figure 1. A Complex Bayesian Network. An arc from one amino-acid node, AA_i, to another, AA_j, represents the dependence of the amino acid at position *j* on the amino acid at position *i*, and encodes 20 x 20 = 400 probabilities $P(AA_j \mid AA_i, C)$ for each classifier value.

With Bayesian networks in this form, we can add arcs representing dependences between pairs of residues in a sequence. In this manner we can go beyond the assumption of conditional independence of sequence positions, which limits most of the existing sequence analysis and structure prediction programs. Pairwise dependences between residues in a sequence are represented in a Bayesian network with arcs from one amino-acid node to another. These arcs represent a correlation between the pairs of amino acids occurring at the two respective positions in a sequence and encode the set of conditional probabilities $P(AA_j \mid AA_i, C)$. Whereas evolutionary relationships are most commonly measured in the positional distributions of amino acids, structural relationships are best detected as correlations among residues. In this work we discover and analyze pairwise correlations between individual residues in α-helical sequences.

The discovery of positional dependences in 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 from the center node C to amino-acid nodes AA_i the null hypothesis is that amino acids 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 AA_i to AA_j correlation 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 $p < 0.001$) the corresponding arc is included in the network. Our discovery program uses a straight-forward exhaustive search of all pairs of positions in a set of sequences. When significant amino-acid correlations are found, corresponding arcs are added to the developing network.

The significance of our χ^2-tests is validated with two other statistical measures: mutual information and Monte Carlo simulations. For the latter, we iteratively test for correlations in randomized sequences constructed by independently shuffling the amino acids within each position in our original sequences. This process preserves positional amino-acid distributions in a sequence set while randomizing any pairwise correlations. Arcs remain in a motif network only if significance is maintained in the Monte Carlo analysis.

The sequences we analyze in this paper were extracted from a non-homologous set of chains from the Brookhaven Protein Data Bank (Bernstein et al., 1977). To construct this set, we first eliminated all non-protein structures, mutant structures, model structures and low resolution structures (> 2.5 Å). Next, within this set, all pairwise sequence comparisons were made using the FASTDB program of the Intelligenetics Suite of sequence analysis programs. Chains were grouped such that for every sequence in a given group, no sequence in any other group was better than 30% identical. Lastly, the chain from the structure with the best resolution was chosen from each group as the representative sequence for that group.

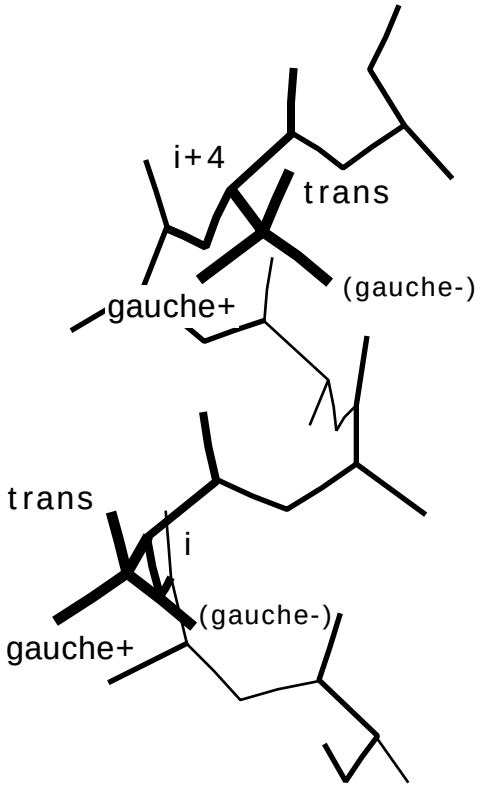
This algorithm gave a structure set of 167 chains. We used the Iditis program from Oxford Molecular (Thornton & Gardner, 1989), a program for querying the a relational database form of the PDB, to extract the residue pairs at specific relative spacings in α-helical sequences assigned by the extended DSSP method (Kabsch & Sander, 1983). We analyzed all 4967 (*i*, *i*+4) pairs and all 5686 (*i*, *i*+3) pairs from the structure set described for significant correlations between amino acid pairs in each of the relative spacings. These two relative spacings in helical

segments allow side chain contacts across adjacent loops in an α-helix.

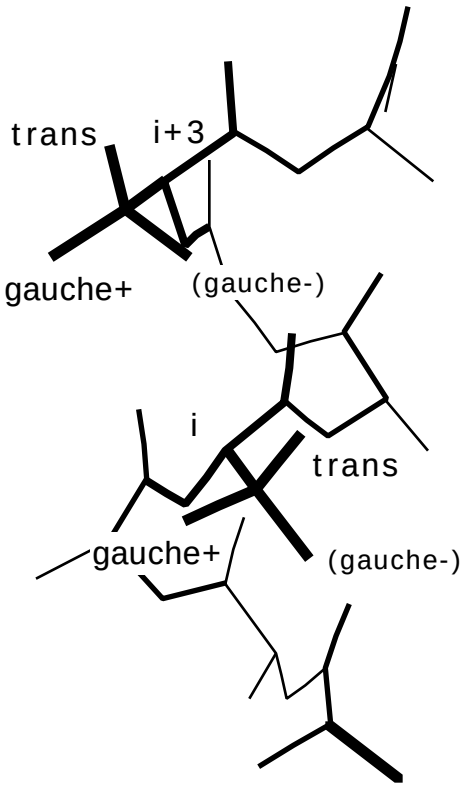
Furthermore, we examined the structural conformations of the significantly correlated amino acid pairs for specific side chain-side chain interactions. We hypothesized that over-represented pairs reflect specific side chain-side chain interactions. Although pairs of side chains can interact in an α-helix when they are in the (*i*, *i*+4), (*i*, *i*+3) and (*i*, *i*+1) arrangements, to form an interaction the amino-acid side chains are constrained to a subset of their possible rotamer conformations, particularly at the χ_1 angle (the dihedral angle about the C_α–C_β bond). Therefore, we compared the rotamer frequencies at χ_1 for the amino acids involved in over-represented amino-acid pairs and the rotamer frequencies for those amino acids anywhere in α -helix.

The rotamer frequencies are obtained by partitioning all side-chain rotamers into distinct classes based on χ_1. Side-chain dihedral angles χ, range from -180° to 180°, with classes defined as follows for angles between tetrahedral atoms: *trans* ($\chi > 120°$ and $\chi \leq -120°$), *gauche*+ ($-120° < \chi \leq 0°$) and *gauche*– ($0° < \chi \leq 120°$). The preferred χ_1 angles for contacting (*i*, *i*+4) and (*i*, *i*+3) pairs are shown in Figure 2.

For each of the highly significant sequence correlations, an analysis of χ_1 angles was performed to ascertain whether a structural interaction was responsible for the sequence correlation. The distribution of χ_1 angles for the bond between a tetrahedral C_α and a tetrahedrl C_β in proteins 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 hindrance with backbone carbonyls: the respective χ_1 frequences are 38.5%, 54.7% and 6.8%. Furthermore, there are characteristic χ_1 conformations for each of the amino acids because of the specific steric properties of individual side chains. For example, valine, isoleucine and threonine side chains (with branched C_β's) are even more strongly constrained to *trans* and *gauche*+ than are the other amino acids. In



A. i and i+4 χ_1 angles.



B. i and i+3 χ_1 angles.

Figure 2. χ_1 Angles in (*i*, *i*+3) and (*i*, *i*+4) α-Helical Pairs. The predominant χ_1 orientations for (*i*, *i*+4) interactions are *trans* and *gauche*+, respectively. The predominant χ_1 angles for (*i*, *i*+3) interactions are *gauche*+ and *gauche*+, repectively.

our χ_1 analysis, expected side-chain χ_1 angles are calculated from the amino acid-specific distributions. 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+*, while 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+* (or rarely *trans* and *gauche–*, respectively).

Lastly, we searched for amino acid pairs in the (*i*, *i*+4) and (*i*, *i*+3) relative orientations that had significantly skewed χ_1 angle distributions. We performed this analysis in order to detect any amino-acid pairs that make structural contacts in α-helices but don't show significant sequence correlations.

In all the analyses described we discover dependences among variables that *may* indicate structural interactions: these dependences are not limited to *contacting* preferences. Our method is general in the sense that "negative" correlations are also detected indicating interactions that are avoided. For each of the analyses performed, correlations with χ^2-values above 10.0 were reported (overall $p<0.05$ for all tests using the Bonferroni approximation).

Results

Table 1 lists the most significant ($p < 0.005$) pairs for the (*i*, *i*+4) and (*i*, *i*+3) sequence interactions. Each row of the table represents a test performed on a 2x2 contingency table in which each of the two dimensions of the table represent one of the amino acids in each pair. Significant correlations can be due to observing many more sequence pairs than expected, or many fewer pairs than expected. There are a range of explanations, varying in scale, for the reported amino-acid correlations in α-helices. Most generally, some may reflect the amphipathic patterns seen in α-helices, which places side chains of similar hydropathy in proximity. Indeed, most of the pairs in Table 1 involve amino acids of like hydropathy; the only under-represented pair (KL) consists of residues of opposite hydropathy. However, helix amphipathicity, which involves more than just pairs of residues in α-helices, wouldn't be expected to give skewed χ_1 angles. And since Table 2 will show preferred χ_1 angles for almost all of the pairs in Table 1, we hypothesize that over-represented pairs reflect specific side chain-side chain interactions. It is somewhat surprising that contacting χ_1 angle preferences are seen even for pairs of hydrophobic residues, which might not be expected to participate in specific pairwise interactions.

Table 1. *(i, i+4) and (i, i+3) Sequence Interactions.*

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

Pair	observed[a]	expected[b]	χ^2[c]	Odds[d]
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^2[c]	Odds[d]
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] Observed pairs in structure data.

[b] Expected pairs in structure data.

[c] χ^2 value for 2x2 contingency table.

[d] Odds: observed number divided by expected number.

There also appears to be evidence for a size effect in helices. Two of the pairs in Table 1, namely SA and GA, involve the smallest side-chains. Considering these and the other, larger pairs, we hypothesize that forming knobs and sockets on the sides of α-helices, in order to facilitate helical packing, is an important feature of helical structure that involves coordination among multiple residues.

Table 2 lists the pertinent χ_1 angles for the correlated residue pairs of Table 1 (the missing pairs involve side chains without C_γ atoms). Each row contains two tests for similarity of χ_1 frequencies (in the *trans*, *gauche+* and *gauche–* conformations) for each residue in the pair compared to the frequencies for those residues anywhere in a helix (each test has two degrees of freedom). All but the last pair in both Tables 2A and 2B show significantly skewed χ_1 angles indicating structural interactions involving almost all the pairs of amino-acid discovered by sequence alone. And more precisely, all of the significantly different χ_1 angles listed are skewed towards their respective preferred contacting angle (see Figure 2).

Table 2. *Side Chains Interactions between or Sequence Pairs in Table 1.*

A. (*i*, *i*+4) side-chain interactions.

Pair	Num	***i* trans obs/exp**[a]	Sig.[c]	***i*+4 *gauche*+ obs/exp**[d]	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	--

B. (*i*, *i*+3) side-chain interactions.

Pair	Number	***i gauche*+ obs/exp**[a]	Sig.[b]	***i*+3 *gauche*+ obs/exp**[c]	Sig.[d]
DR	36	27 / 27.3	--	9 / 15.6	<0.025
LI	56	36 /33	--	45 / 47.5	--

[a] Observed and expected number of first residue χ_1 angles in the predominant interaction orientation.

[b] Significance of the first residue χ_1 angle (by the χ^2 statistic).

[c] Observed and expected number of second residue χ_1 angles in the predominant interaction orientation.

[d] Significance of the second residue χ_1 (by the χ^2 statistic).

Table 3 lists the most significant amino acid pairs with skewed χ_1 angles indicating structural interactions alone. This analysis detects residue interactions without regard for sequence correlations (as was done previously) and finds pairwise interations that aren't manifested in the sequence. Again, each row contains two tests for similar distributions, one for each amino acid in the pair. The table is sorted by the sum of the χ^2 values for each amino acid in the pairs. This represents a single test with four degrees of freedom for similarity of χ_1 frequencies (in *trans*, *gauche*+ and *gauche*– conformations) for both amino acids in each pair compared to their frequencies anywhere in a helix .

In the Table 3A, the amino acid pairs DR, QD, KE, QE and LI are in preferred contacting orientations for both residues in the pairs, while pairs ND, DE and KR appear to avoid contacting orientations at both residues in the pairs. The remaining three entries in this table have mixed preferences. The entries in Table 3B are more difficult to understand.

Table 3. *Side Chains Interactions in a-helix Pairs.*

A. (*i*, *i*+4) side-chain interactions.

Pair	***i trans* obs/exp**[a]	χ^2[b]	***i*+4 *gauche*+ obs/exp**[c]	χ^2[b]	Sum of χ^2[d]
ND	2 / 1.3	1.4	1 / 4.9	18.9	20.3
DE	2 / 4.0	1.4	6 / 14.6	18.2	19.6
DR	9 / 3.3	12.3	13 / 9.0	4.4	16.7
FL	21 / 20.8	1.4	10 / 18.9	13.6	15.0
QD	15 / 8.1	11	17 / 14.6	2.9	13.9
DH	3 / 1.1	4.6	0 / 3.4	9.2	13.8
KE	28 / 23.1	6.2	35 / 26.7	7.1	13.3
KR	4 / 9.9	8	4 / 9.0	4.9	12.9
QE	21 / 14.5	5.4	25 / 20.4	7.5	12.9
LR	9 / 17.9	10.7	19 / 16.9	1.4	12.1
TR	4 / 1.0	9.8	7 / 7.5	1.1	10.9
LI	34 / 24.2	9.1	43 / 40.3	1.5	10.6

B. (*i*, *i*+3) side-chain interactions.

Pair	***i trans* obs/exp**[a]	χ^2[b]	***i*+4 *gauche*+ obs/exp**[c]	χ^2[b]	Sum of χ^2[d]
KT	9 / 5.8	4.8	3 / 7.3	8.4	13.2
QD	5 / 6.4	1.3	7 / 11.8	10.8	12.1
DQ	2 / 1.7	10.6	5 / 5.7	1.3	11.9
LS	23 / 14.4	10.5	12 / 14.6	0.9	11.4
RQ	18 / 12.2	6.3	9 / 13.1	4.5	10.8
TM	1 / 1.0	0.4	3 / 7.1	10.4	10.8
WL	3 / 7.6	6.8	5 / 8.0	4	10.8

[a] Observed and expected number of first residue χ_1 angles in the predominant interaction orientation.

[b] χ^2 value for the corresponding χ_1 angle distributions.

[c] Observed and expected number of second residue χ_1 angles in the predominant interaction orientation.

[d] Sum of the χ^2 values for the χ_1 angles.

Discussion

Many of the correlations listed in Tables 1, 2 and 3 represent specific side chain-side chain interactions that have been shown experimentally to stabilize α-helices (Armstrong & Baldwin, 1993; Burley & Petsko, 1988; Huyghues-Despointes & Baldwin, 1994; Marqusee, Robbins & Baldwin, 1989; Padmanabhan & Baldwin, 1994; Shoemaker et al., 1990). Some are novel and may

indicate relatively unknown side-chain interactions important for local protein stability. Figure 3 shows superimposed structures of the top 8 (*i*, *i*+4) interactions and the top (*i*, *i*+3) interaction in table 1.

All but one of the highly significant (*i*, *i*+4) sequence correlations (Table 1A) correspond to specific side-chain

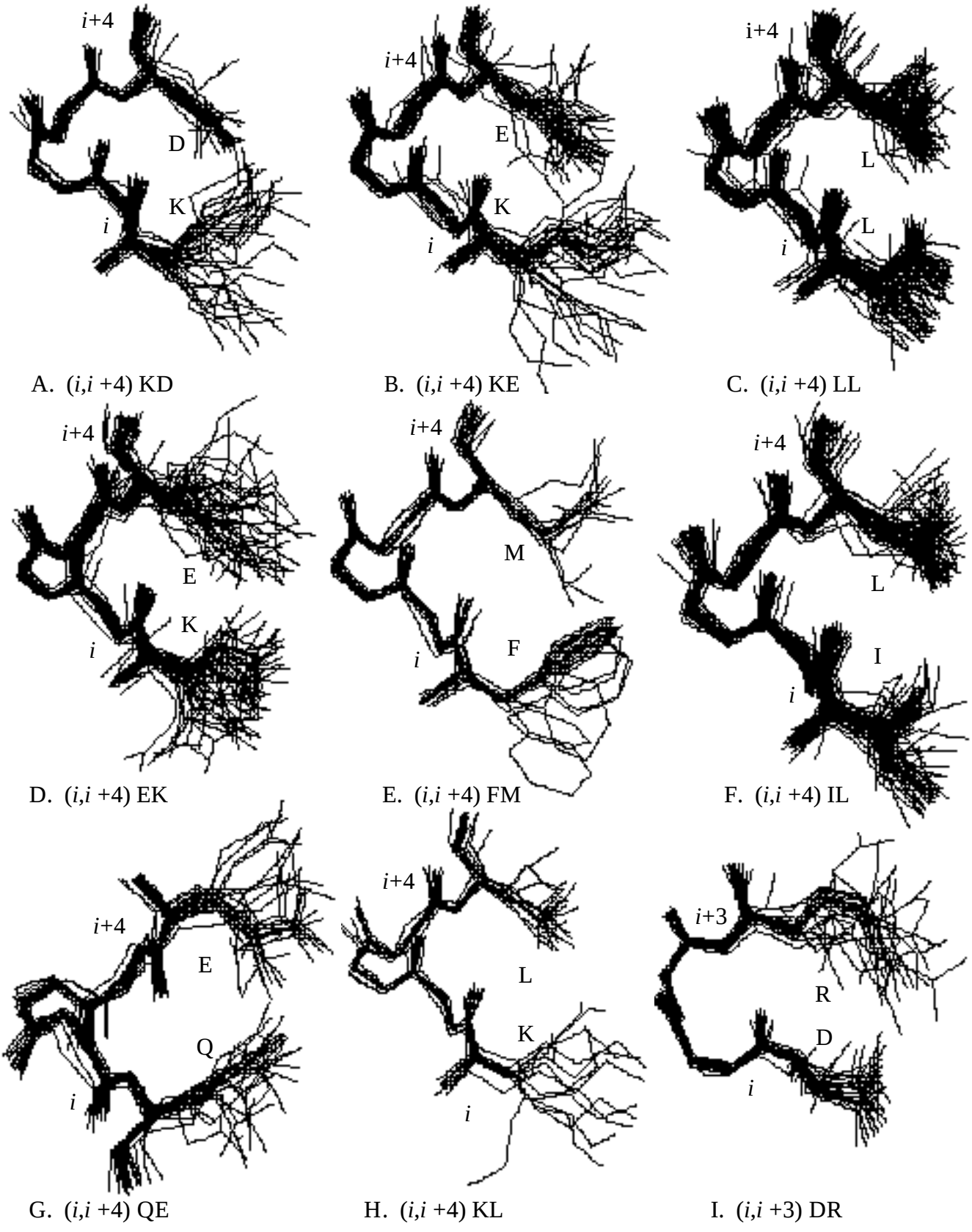


Figure 3. Superimposed (*i*, *i*+4) and (*i*, *i*+3) Interactions

conformations (Table 2A) indicated by preferred contacting χ_1 angles. Electrostatic interactions can be attributed to sequence pairs KD, KE and EK while the sequence pair QE is likely a hydrogen-bond. The interaction of sequence pairs LL, IL and LI is hydrophobic in nature. In contrast, since the sequence pair KL is under-represented and shows no side-chain χ_1 preference, it may be due to the hydrophobic effect alone (i.e. a strong hydrophobic residue and a strong hydrophilic residue are under-represented in positions that would place them in proximity). All but one (the KL pair) of the structures in Figure 3, which are superimposed by backbone atoms only, show side-chain clustering indicative of contacts for a majority of the structures compising each panel.

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 and no experimental confirmation. When the 17 (i, i+4) phenylalanine-methionine pairs are superimposed (Figure 3E) 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.

The presence of the significant pairs SA and GA raises the interesting idea that there are size-based interactions in helices. Whereas most of the other significant pairs are relatively large (and large enough to form an interaction), the two pairs with small or no side chains suggest that some helices require coordinated gaps, possibly for packing against other parts of a protein. This effect may be an extended version of the knobs-and-holes model for helix-helix interactions (Lesk, 1991).

Table 3 confirms many of the sequence pairs found in the sequence analyses and adds a few new pairs. KE, QE and LI are common to both, while DR and QD are not detected as significantly correlated at (i, i+4). DR is a likely salt bridge and QD can form a stabilizing hydrogen bond (Huyghues-Despointes & Baldwin, 1994). Also seen in Table 3A are three interesting pairs that avoid contacts: ND, DE and KR. The latter two are probably due to charge repulsion. Both involve side chains with like charges that would be destabilizing were they to lie near each other. The (i, i+3) pairs in Table 3B are difficult to analyze because of the complexity of the arrangements, but many of the pairs fit into the categories already mentioned (hydrogen bonding or hydrophobic interactions).

All of the correlations described are modeled as conditional probabilities in arcs between pairs of amino acid nodes in a Bayesian network representing an α-helix. We can also include correlations of lower significance while leaving the remaining pairs in their default, independent state. We are currently developing these networks in order to measure the improvement in secondary structure prediction one can get from representing structural interactions.

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 pre-doctoral trainee of the National Library of Medicine.

References

Armstrong, K. M. and Baldwin, R. L. 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(23): 11337-40.

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. J., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. 1977. The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.*, 112: 535-542.

Burley, S. K. and Petsko, G. A. 1988. Weakly Polar Interactions in Proteins. *Adv. Prot. Chem.*, 39: 125-189.

Henikoff, S. and Henikoff, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucl. Acids. Res.*, 19(23): 6565-6572.

Huyghues-Despointes, B. and Baldwin, R. L. 1994. Helical stabilization by hydrogen-bonding sidechains. Personal Communication.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12): 2577-637.

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

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

Lesk, A. M. 1991. *Protein Architecture-A Practical Approach*. Oxford: IRL Press at Oxford University Press.

Marqusee, S., Robbins, V. H. and Baldwin, R. L. 1989. Unusually stable helix formation in short alanine-based peptides. *Proc. Natl. Acad. Sci. USA*, 86(14): 5286-90.

Neapolitan, R. E. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. New York, NY: Wiley and Sons.

Padmanabhan, P. and Baldwin, R. L. 1994. Helical stabilization by hydrophobic sidechains. Personal Communication.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Shoemaker, K. R., Fairman, R., Schultz, D. A., Robertson, A. D., York, E. J., Stewart, J. M. and Baldwin, R. L. 1990. Side-chain Interactions in the C-peptide helix: Phe 8^{-}—His 12^{+}. *Biopolymers*, 29: 1-11.

Thornton, J. M. and Gardner, S. P. 1989. Protein motifs and data-base searching. *Trends Biochem. Sci.*, 14(7): 300-4.

Wilbur, W. J. and Lipman, D. J. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc Natl Acad Sci U S A*, 80(3): 726-30.