Local sequence–structure correlations in proteins

Christopher Bystroff*, Kim T Simons†, Karen F Han‡ and David Baker§

Considerable progress has been made in understanding the relationship between local amino acid sequence and local protein structure. Recent highlights include numerous studies of the structures adopted by short peptides, new approaches to correlating sequence patterns with structure patterns, and folding simulations using simple potentials.

Addresses
*†§ Department of Biochemistry, Box 357350, University of Washington, Seattle, WA 98195, USA
* e-mail: bystroff@ben.bchem.washington.edu
† e-mail: ksimons@ben.bchem.washington.edu
§ e-mail: baker@ben.bchem.washington.edu
‡ Northwestern University Medical School, Box 182, Chicago, IL 60611, USA; e-mail: han@msg.ucsf.edu

Current Opinion in Biotechnology 1996, 7:417–421

Abbreviations

3D	three-dimensional
TFE	2,2,2-trifluoroethanol

Introduction

It is well established that the three-dimensional (3D) structures of proteins are determined by their amino acid sequences, yet the prediction of structure from sequence remains an unsolved problem. The importance of interactions between residues distant in the linear sequence is one of the features of proteins that makes the problem difficult. These interactions play a critical role in stabilizing proteins: unique well-defined structure in water is rare in peptides of less than ~30 amino acids [1•,2•,3••,4].

Despite the importance of nonlocal interactions in determining protein structures, the relationship between local sequence and local structure remains an important and active area of research. Understanding such interactions is important for predicting protein secondary structure, often a first step in 3D structure modeling and prediction. The relationship is also important for understanding the process of folding. It is clear that a folding polypeptide chain cannot exhaustively search conformational space; instead, local sequence preferences are likely to limit the number of configurations available to each portion of a polypeptide chain and so are likely to decrease greatly the effective size of the space that must be searched.

In this review, we focus on recent advances in predicting structural properties from local amino acid sequence and for probing the relationship between local sequence and structure. Some attention is also paid to the types of interactions responsible for the observed sequence–structure relationships. Because excellent reviews of secondary-structure prediction and protein sequence–structure relationships have only recently appeared [5••,6••], the classical secondary-structure prediction problem is not covered in detail, and the discussion is, for the most part, limited to papers that have appeared during the past year.

Recurrent structural patterns

In recent years, considerable work has been directed at better defining local structural motifs and analyzing their sequence preferences. In general, structural motifs have been identified by inspection of the ever-increasing database of protein crystal structures. Thornton and collaborators [7•] have carried out much important work in characterizing local structural motifs; a program (PROMOTIF) that identifies a large variety of such motifs in a protein structure file is now available. Once defined, the frequencies of occurrence of the amino acids in each position in the motif can be calculated from the protein structure database. These frequencies can then be used to predict the occurrence of the motifs in new sequences. For example, the sequence preferences of the various types of β turns have recently been re-evaluated using a larger structural database [8].

Much work during the past year has focused on the structural characterization of peptide models of previously identified motifs. Some of the strongest local sequence–structure correlations are observed at the amino and carboxyl termini of α helices. The Schellman motif [9] is frequently observed at the carboxyl termini of α helices, and contains a conserved glycine residue immediately following the last residue in the helix. Peptide studies have shown that this motif is not significantly populated in aqueous solution [10•]. In contrast, studies of peptides with an amino-terminal helix capping motif, the 'hydrophobic staple' [11•] or 'extended capping box' [12], which contains two conserved hydrogen bonds involving a serine and a glutamate residue, have identified significant native-like structure [11•,13]. Thus, local interactions are sufficient to stabilize the latter helix cap motif but not the former. Nonetheless, both helix caps can be predicted from sequence with a fairly high degree of confidence.

Studies of peptides corresponding to β-hairpin regions of proteins have shown ordered structure in some cases [14,15•,16•] but not in others [1•,17•]. Peptides with sequences designed based on observed turn propensities adopt β-hairpin structures [18], but in at least one case the strands are held together by interactions between hydrophobic side chains rather than by backbone hydrogen bonds [19••]. Several studies have utilized 2,2,2-trifluoroethanol (TFE) as a structure-enhancing solvent, but this may artificially induce helix formation [2•,20•], and the significance of such results is unclear, given the importance of solvent in local structure formation [21]. In all of the above peptide studies, it should be noted [22••] that given the loss in conformational entropy, the observation of even low levels of occupancy of a particular conformation requires that the conformation be low in energy relative to the other possible conformations. Thus, local interactions may contribute substantially to protein stability even if structure is not observed in isolated peptides.

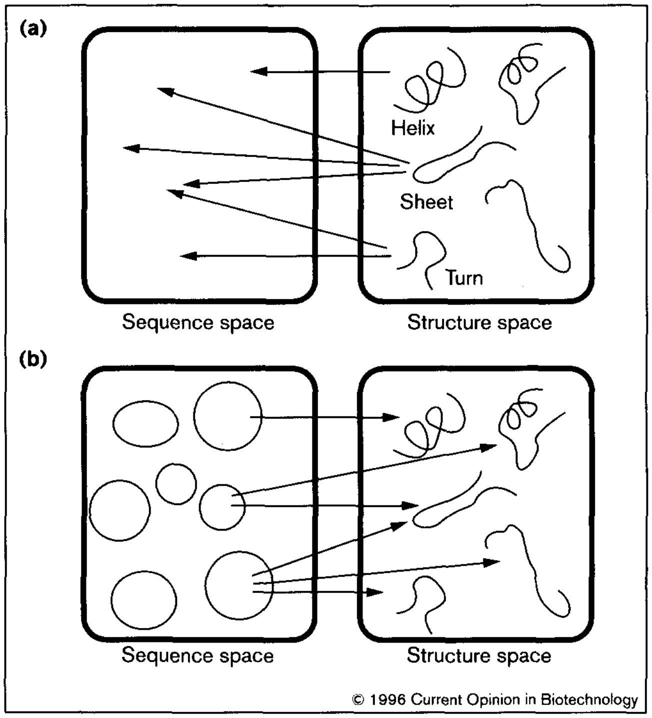
When calculating the sequence preferences of structural motifs, it is commonly assumed that the residue preferences at each position in a motif are independent. This approximation may be rather poor, but the consideration of covariances between residue preferences at pairs of positions generally requires more data than is available from the structure database [23]. Within the past year, several important advances in this area have taken place. An elegant mutation study of a pair interaction between spatially adjacent β-sheet residues in protein G showed significant preference for complementary charge pairs and particular pairs of hydrophobic residues over that expected from the analysis of single substitutions [24••]. These covariances mirror the statistical trends observed in the protein database. Pair correlations in β strands have been used to predict β-strand pairings with remarkable success [25••,26]. Pair correlations also form the basis of a new algorithm for predicting coiled coils in proteins, which appears to do significantly better than previous approaches which utilized only single residue preferences [27•,28].

Because of the importance of residue hydrophobicity in protein folding, a natural way to reduce the complexity of sequence–structure mapping is to convert amino acid sequences into a two-letter code: H (hydrophobic) or P (polar). Studies of peptides with periodic hydrophobicity patterns show that amphipathicity can outweigh the intrinsic preferences of the different amino acids for the different secondary-structure types. HP patterns are thus sufficient conditions for the formation of helix and sheet in short peptides, although they are not necessary conditions [29•]. Analysis of the structural database has shown a strong correlation between pentapeptide HP patterns and α helices, but less correlation for β sheets [30•].

Recurrent sequence patterns

The underlying approach in the studies mentioned thus far is to study the sequence correlates of predefined structural properties using the database of sequences whose structures are known, and then to use the results to predict the structural characteristics of new sequences (Fig. 1a). The converse approach is to search for sequence patterns first, and to then study their structural correlates (Fig. 1b). Because the important structural properties need not be specified in advance, new structural motifs can potentially be identified. A potential advantage of this approach is that one-dimensional amino acid sequences may be more amenable to pattern-recognition approaches than 3D protein structures.

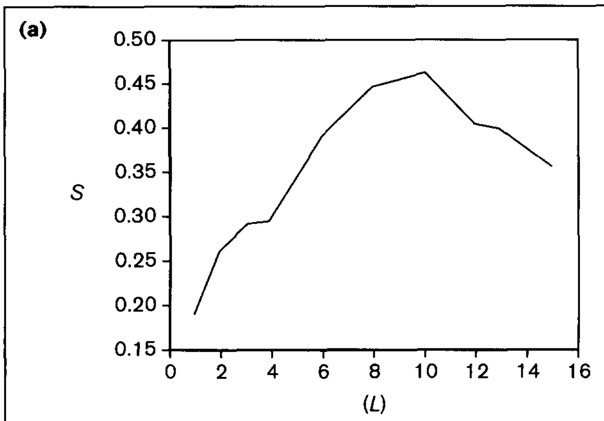
Figure 1



Two approaches to studying local sequence–structure relationships. **(a)** Determination of the sequence correlates of predefined local structures. **(b)** Determination of structural correlates of sequence patterns. The circles indicate groups of protein segments with similar sequences.

If proteins contain a finite number of different local structural motifs, multiple sequence alignments should also exhibit a finite number of patterns of sequence variation. Starting with this assumption, recurring sequence patterns that transcend protein family boundaries were identified in the HSSP database of multiple sequence alignments for proteins of known structure using cluster analysis [31•]. The recurrent sequence patterns are in part recognizable patterns of hydrophobic and hydrophilic residues, and in part less obvious combinations [31•]. Because protein structural information was not used in the identification of the patterns, any correlations between pattern and local structure reflect structural information in local sequence. The correlation between sequence and structure increases as the pattern length increases from three to eight residues, and then slowly decreases for longer pattern lengths (Fig. 2). The decrease may reflect the average number of residues required to span a protein; the patterns are based on ungapped alignments and thus do not cover variable length turns and loops. The limited size of the protein database also becomes an increasingly important problem for longer segment lengths.

Figure 2



(b) $$S_L = \frac{1}{200L}\sum_{k=1}^{200}\sum_{j=1}^{L}\sum_{i=H,T,S} P_{ijk}\ln\left(\frac{P_{ijk}}{P_i}\right)$$

Sequence–structure correlations for different segment lengths. Segments of proteins of known structure were partitioned into 200 groups based on sequence similarity [31•]. **(a)** The relative entropy, *S*, is plotted as a function of the segment length in amino acids (*L*) used in the partitioning. By this measure, a segment length of 10 contains the greatest amount of local sequence-dependent structural information. **(b)** The similarity in secondary structure within a group of segments is reflected in the relative entropy. P_{ijk} is the fraction of segments in group *k* that have secondary-structure type *i* at position *j*, and P_i is the fraction of secondary structure type *i* in the database overall. Each position in each segment has a secondary structure assignment: H (α helix), S (β sheet) or T (other).

Patterns for which one and two local structures predominate account for 45% and 28% of the protein database, respectively [32••]. The first set of patterns probably includes virtually all of the short sequence patterns in proteins that consistently occur in a particular local structure. Many of the patterns discussed in the preceding section, as well as several new sequence–structure relationships, have been reidentified by this automated approach.

A disadvantage of the simple clustering procedure used in these studies is the lack of an underlying statistical model. An important recent development in this area is the use of a Dirichlet mixture model to describe the major types of amino acid distributions found in columns of multiple sequence alignments for proteins belonging to the same family [33••]. Because of their different contexts in protein 3D structures, some positions accept primarily hydrophobic residues, others accept small residues, etc. Each component of the mixture model describes one such distribution, but rather than being fixed at the outset, the parameters describing each distribution are estimated from a training set of multiple sequence alignments using a maximum likelihood approach. The Dirichlet mixtures essentially cluster amino acid distributions into prototypical classes of distributions. Because it provides a recipe for generalizing from a small amount of data, the mixture model is extremely useful in predicting the amino acid variation likely to be observed at a particular position in a protein given only a small number of starting aligned sequences.

Origin of sequence–structure correlations

Why do some local sequences have a high tendency to occur in particular types of local structure? A variety of factors to account for the observed secondary-structure propensities of the amino acids have been proposed. These include side-chain entropy, buried surface area and steric factors. It has been proposed recently that electrostatic interactions between backbone atoms are largely responsible for the observed preferences, and a model in which the different amino acids differentially screen these electrostatic interactions performs quite well in accounting for the observed preferences [34•].

Interesting developments within the past year include approaches to predicting the configuration of peptides and short proteins starting from simple physical principles. In addition to the obvious usefulness of a program for predicting tertiary structures, such approaches have the potential to illuminate the basis for observed sequence–structure correlations if they can reproducibly generate native structures. One such approach utilized a simple treatment of hydrophobic interactions, hydrogen bonding, and steric overlap together with a hierarchical assembly procedure: at the start of a simulation, only local interactions are considered, and any persisting structure is fixed in the later stages of the simulation when longer-range interactions are considered [35••]. Surprising features of the results are the striking accuracy of the secondary-structure predictions, and the fixing of isolated β strands early in simulations despite the relatively weak local interactions. A related potential function which emphasizes hydrogen bonds between buried hydrogen-bond donors and acceptors was used in conjunction with a novel extensive searching procedure to fold small proteins and peptides with a reasonable degree of success [36••].

A somewhat different potential function, which emphasizes main-chain electrostatic effects, has been used in conjunction with torsional space Monte Carlo to fold fragments of proteins thought to be folding initiation sites. In all but one of the examples, the lowest energy configuration was very similar to the structure found experimentally in the context of the entire protein structure [22••]. Dissection of the potential function suggested that main-chain hydrogen bonding, main-chain electrostatics and the burial of hydrophobic groups all contribute to the stabilization of the native-like structures. A blind test using the same potential function, but with the genetic algorithm rather than Monte Carlo, resulted in a roughly correct prediction for one of the three peptides studied [37•].

Conclusions

Continued progress with such simulation efforts should provide insight into the energetic origins of sequence–structure relationships. Conversely, continued progress in understanding local sequence–structure correlations should contribute to the prediction of protein tertiary structure: the size of the conformational space that must be searched can potentially be greatly reduced by confining short segments to likely local structures. The power of statistical approaches will grow as the size of the protein structure database increases, making possible the elucidation of more subtle sequence–structure relationships.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Viguera AR, Serrano L: **Conformational analysis of peptides corresponding to β-hairpins and a β-sheet that represent the entire sequence of the α-spectrin SH3 domain.** *J Mol Biol* 1996, **255**:507–521.

The structures adopted by peptides spanning the entire length of the spectrin SH3 domain are analyzed by NMR and circular dichroism, and are found not to contain significant amounts of native-like structure.

2. • Yang JJ, Buck M, Pitkeathly M, Kotik M, Haynie DT, Dobson CM, Radford SE: **Conformational properties of four peptides spanning the sequence of hen lysozyme.** *J Mol Biol* 1995, **252**:483–491.

Peptides from the β-sheet domain of lysozyme are unstructured in water, whereas a peptide from the carboxy-terminal portion of the helical domain adopts a partially helical structure. It is proposed that the carboxy-terminal portion of the helical domain plays an important role in early folding events.

3. •• De Prat Gay G, Ruiz-Sanz J, Neira JL, Corrales FJ, Otzen DE, Ladurner AG, Fersht AR: **Conformational pathway of the polypeptide chain of chymotrypsin inhibitor-2 growing from its N terminus *in vitro*. Parallels with the protein folding pathway.** *J Mol Biol* 1995, **254**:968–979.

This paper (together with [4]) describes the structures of progressively longer peptides that start at the amino terminus of chymotrypsin inhibitor-2. Consistent with the highly cooperative acquisition of structure observed in the folding of the intact protein, little persistent structure is observed in all but the nearly full-length peptides.

4. De Prat Gay G, Ruiz-Sanz J, Neira JL, Itzhaki LS, Fersht AR: **Folding of a nascent polypeptide chain *in vitro*: cooperative formation of structure in a protein module.** *Proc Natl Acad Sci USA* 1995, **92**:3683–3686.

5. •• Cordes MHJ, Davidson AR, Sauer RT: **Sequence space, folding and protein design.** *Curr Opin Struct Biol* 1996, **6**:3–10.

An excellent recent review of topics related to those discussed here.

6. •• Barton GJ: **Protein secondary structure prediction.** *Curr Opin Struct Biol* 1995, **5**:372–376.

An excellent recent review of secondary-structure prediction.

7. • Hutchinson EG, Thornton JM: **PROMOTIF – a program to identify and analyze structural motifs in proteins.** *Protein Sci* 1996, **5**:212–220.

A very useful tool for finding defined local structural motifs in proteins and compiling statistics on their sequence preference.

8. Hutchinson EG, Thornton JM: **A revised set of potentials for beta-turn formation in proteins.** *Protein Sci* 1994, **3**:2207–2216.

9. Schellman C: **The αL conformation at the ends of helices.** In *Protein Folding: Proceedings of the 28th Conference of the German Biochemical Society: 1979 Sep 10–12; Regensburg.* Amsterdam: Elsevier/North-Holland Biomedical Press; 1980:53–56.

10. • Viguera AR, Serrano L: **Experimental analysis of the Schellman motif.** *J Mol Biol* 1995, **251**:150–160.

Further characterization of the Schellman α helix carboxyl cap. The frequencies of occurrence of the conserved glycine residue, the hydrophobic interaction and the conserved polar residue in capping elements are determined. A designed peptide folds to the motif weakly in water but strongly in TFE.

11. • Muñoz V, Blanco FJ, Serrano L: **The hydrophobic-staple motif and a role for loop-residues in α-helix stability and protein folding.** *Nat Struct Biol* 1995, **2**:380–385.

A hydrophobic interaction between a residue located before an amino terminal helix cap and a residue within the helix is frequently observed. NMR and circular dichroism studies demonstrate that peptides with this 'hydrophobic staple' motif adopt helix cap structures in solution.

12. Harper ET, Rose GD: **Helix stop signals in proteins and peptides: the capping box.** *Biochemistry* 1993, **32**:7605–7609.

13. Muñoz V, Serrano L: **Analysis of i,i+5 and i,i+8 hydrophobic interactions in a helical model peptide bearing the hydrophobic staple motif.** *Biochemistry* 1995, **34**:15301–15306.

14. Blanco FJ, Rivas G, Serrano L: **A short linear peptide that folds into a native stable beta-hairpin in aqueous solution.** *Nat Struct Biol* 1994, **1**:584–590.

15. • Blanco FJ, Serrano L: **Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements.** *Eur J Biochem* 1995, **230**:634–649.

The solution structure of peptides corresponding to the secondary-structure elements of protein G are studied. The second β hairpin has considerably more structure than either the first β hairpin or the α helix, suggesting that it may be a folding initiation site.

16. • Ilyina E, Mayo KH: **Multiple native-like conformations trapped via self-association-induced hydrophobic collapse of the 33-residue β-sheet domain from platelet factor 4.** *Biochem J* 1995, **306**:407–419.

An NMR study shows that a length of 33 residues, along with some stabilizing interchain salt bridges and hydrophobic contacts, is sufficient for the formation of a stable tetramer of three-stranded β sheets.

17. • Itzhaki LS, Neira JL, Ruiz Sanz J, De Prat Gay G, Fersht AR: **Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2.** *J Mol Biol* 1995, **254**:289–304.

NMR studies of peptides 5–28 residues in length corresponding to β-strand, β-turn and α-helix regions of CI2 show very low populations of native structure in water, and only slightly higher concentrations in TFE.

18. De Alba E, Jiménez MA, Rico M, Nieto JL: **Conformational investigation of designed short linear peptides able to fold into β-hairpin structures in aqueous solution.** *Folding Des* 1996, 1:133–144.

19. •• Sieber V, Moe GR: **Interactions contributing to the formation of a β-hairpin-like structure in a small peptide.** *Biochemistry* 1996, **35**:101–188.

NMR studies of a 12-residue designed peptide show formation of a stable β hairpin without the backbone hydrogen-bonding network. Instead, interstrand pairing is seen between hydrophobic side chains.

20. • Hamada D, Kuroda Y, Tanaka T, Goto Y: **High helical propensity of the peptide fragments derived from beta-lactoglobulin, a predominantly beta-sheet protein.** *J Mol Biol* 1995, **254**:737–746.

Peptides corresponding to β strands in β-lactoglobin were found to form α helices in TFE, calling into question the use of TFE to stabilize structure in peptides.

21. Waterhous DV, Johnson WC Jr: **Importance of environment in determining secondary structure in proteins.** *Biochemistry* 1994, **33**:2121–2128.

22. •• Avbelj F, Moult J: **Determination of the conformation of folding initiation sites in proteins by computer simulation.** *Proteins* 1995, **23**:129–141.

Folding initiation sites are sought using simple physical principles. A torsion space Monte Carlo sampling procedure together with an all-atom free-energy function incorporating local main-chain electrostatics, main-chain hydrogen bonds, and burial of nonpolar area is found to produce conformations that in some cases are close to those observed experimentally.

23. Gibrat JF, Garnier J, Robson B: **Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs.** *J Mol Biol* 1987, **198**:425–443.

24. •• Smith CK, Regan L: **Guidelines for protein design: the energetics of beta sheet side chain interactions.** *Science* 1995, **270**:980–982.

Substantial pair interactions are observed between two positions in protein G that are hydrogen-bonded in an antiparallel β hairpin. The results highlight the importance of such pair interactions in protein folding and design.

25. Hubbard TJ, Park J: **Fold recognition and *ab initio* structure**
•• **prediction using hidden Markov models and β-strand pair potentials.** *Proteins* 1995, **23**:398–402.
Impressive use of pair correlations between residues in adjacent β strands to predict protein structure.

26. Hubbard TJ: **Use of β-strand interaction pseudo-potentials in protein structure prediction and modelling.** In *Proceedings of the Biotechnology Computing Track, Protein Structure Prediction Minitrack of the 27th HICSS.* Los Alamitos, CA: IEEE Computer Society Press; 1994:336–354.

27. Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS:
• **Predicting coiled coils by use of pairwise residue correlations.** *Proc Natl Acad Sci USA* 1995, **92**:8259–8263.
An algorithm for predicting coiled-coil domains in protein sequences using pairwise residue correlations is described. A program, PAIRCOIL, implementing the method does not produce any obvious false postives or negatives in searches of the protein database for coiled-coil domains.

28. Berger B: **Algorithms for protein structural motif recognition.** *J Comput Biol* 1995, **2**:125–138.

29. Xiong H, Buckwalter BL, Shieh HM, Hecht MH: **Periodicity of**
• **polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides.** *Proc Natl Acad Sci USA* 1995, **92**:6349–6353.
A peptide design experiment shows that polar/nonpolar side-chain periodicity controls the choice of secondary structure (α or β) over intrinsic propensity.

30. West MW, Hecht MH: **Binary patterning of polar and nonpolar**
• **amino acids in the sequences and structures of native proteins.** *Protein Sci* 1995, **4**:2032–2039.
Pentapeptide patterns of polar and nonpolar amino acids are analyzed in known protein structures. Periodic patterns are found to be better predictors of α helix than β sheet.

31. Han KF, Baker D: **Recurring local sequence motifs in proteins.**
• *J Mol Biol* 1995, **251**:176–187.
Cluster analysis is used to identify recurrent patterns of sequence variation at single positions and in short segments of contiguous positions in multiple sequence alignments for a nonredundant set of protein families.

32. Han KF, Baker D: **Global properties of the mapping from local**
•• **amino acid sequence to local structure in proteins.** *Proc Natl Acad Sci USA* 1996, **93**:5814–5818.
The structural correlates of sequence patterns identified using the methods described in the previous paper [31•] are investigated. A first class of patterns consistently occur in a single type of local structure in proteins; other patterns occur in one of two or three types of local structures. The frequencies of occurrence of the three classes of patterns in the protein database are considerably higher than in a simulated dataset. The sequence and associated structural features of a subset of the first class of patterns are described.

33. Sjolander K, Karplus K, Brown MP, Hughey R, Krogh A, Main IS,
•• Haussler D: *Dirichlet Mixtures: a Method for Improving Detection of Weak but Significant Protein Sequence Homology.* UCSC Technical Report UCSC-CRL-96-09; 1996.
This report summarizes the development of a Dirichlet mixture model for the amino acid substitution patterns in proteins. Not only is the model elegant, but the results of searches for distant homologs show a substantial increase in performance.

34. Avbelj F, Moult J: **Role of electrostatic screening in determining**
• **protein main chain conformational preferences.** *Biochemistry* 1995, **34**:755–764.
An interesting argument is made for the importance of electrostatic interactions between main-chain polar atoms in determining the secondary-structure propensities of different side chains (different side chains screen these interactions to different extents). A model based on electrostatics accounts quite well for the observed conformational preferences of the different amino acids.

35. Srinivasan R, Rose GD: **LINUS: a hierarchic procedure to**
•• **predict the fold of a protein.** *Proteins* 1995, **22**:81–99.
The supersecondary structure of protein fragments is predicted using a relatively simple energy function and a hierarchical procedure. At each stage, segments of the chain that consistently adopt a particular local configuration are fixed in that configuration before increasing the size of the interaction window. Secondary structure is successfully predicted.

36. Yue K, Dill K: **Folding proteins with a simple energy function**
•• **and extensive conformational searching.** *Protein Sci* 1996, **5**:254–261.
Structures of very small proteins are predicted using a very simple energy function and a novel algorithm for conformational searching.

37. Pederson J, Moult J: ***Ab initio* structure prediction for small**
• **polypeptides and protein fragments using genetic algorithms.** *Proteins* 1995, **23**:454–460.
Blind predictions are made for the structure of three short peptides using a potential similar to that in the previous paper [36••], but using a search based on a genetic algorithm rather than Monte Carlo. For one of the three peptides, the lowest conformation had a root mean square deviation of 4.4 Å from the true structure (over 22 amino acid residues).