Structure-derived potentials and protein simulations

Robert L Jernigan*‡ and Ivet Bahar*†

There has recently been an explosion in the number of structure-derived potential functions that are based on the increasing number of high-resolution protein crystal structures. These functions differ principally in their reference states; the usual two classes correspond either to initial solvent exposure or to residue exposure of residues. Reference states are critically important for applications of these potentials functions. Inspection of the potential functions and their derivation can tell us not only about protein interaction strengths themselves, but can also provide suggestions for the design of better folding simulations. An appropriate goal in this field is achieving self-consistency between the details in the derivation of potentials and the applied simulations.

Addresses
*Laboratory of Mathematical Biology, MSC 5677, Room B-116, Bldg 12A, National Institutes of Health, Bethesda, MD 20892-5677, USA
†Bogazici University and TUBITAK Polymer Research Center, Istanbul, Turkey
‡e-mail: jernigan@lmmb.nci.nih.gov

Current Opinion in Structural Biology 1996, 6:195–209

Abbreviations

H	hydrophobic
MC	Monte Carlo
MD	molecular dynamics
P	polar
PDB	Protein Databank
rms	root mean square

Introduction

In recent years, there has been a growing interest in deriving potentials of mean force from experimentally observed frequencies of non-bonded pairs of amino acids in sets of structures. The analysis of inter-residue contact preferences in X-ray elucidated structures has contributed substantially to understanding the dominant forces stabilizing native structures in globular proteins. Such protein potentials have an important role in simplified models of protein structure [1]. Low-resolution methods appear to be the most practical approach for unraveling the complex issues in protein folding and recognition [2]. Here we will consider their utility for potential energy derivation and computer simulation.

The idea of extracting such knowledge-based potentials was first conceived by Tanaka and Scheraga [3]. A rigorous determination of the effective inter-residue contact potentials, including both solvent and size effects, was achieved by Miyazawa and Jernigan [4]. The latter contact potentials have been extensively tested and used to analyze and simulate protein structures. Eisenberg and co-workers [5] evaluated the environment of each amino acid in each structure on the basis of several properties. Sippl [6] studied inter-residue interactions, which are expressed in terms of distance-dependent potentials. Crippen and Maiorov [7] have fitted large numbers of parameters to structures. One important area of application of potential functions has been the evaluation of sequences threaded onto known structures [8]. These studies have been followed by many others along the same lines [9–17,18•,19,20••]. In this review, however, the focus will instead be on the use of potential functions in conformational simulations. It is worth remarking that the physical situations requiring assessment by energy functions for threading, binding, and folding simulations are much the same.

Motivation

There is a long history of the use of interaction energies as adjustable parameters to reflect varied circumstances. In part, this was an attempt to deal with highly varied situations: for instance, there is evidence that the relative populations of rotational isomers are affected by their environment. In the case of polymer theory, the interactions between two atomic groups at close range have typically been evaluated from physical measurements of the entire chain, such as its overall dimensions. Much of this adjustable energy parameter approach has been summarized by Flory [21] and Mattice and Suter [22]. Admittedly, circumstances prevailing in proteins are complex, and it can readily be presumed that different effective potentials might be operative at different stages of folding because the local environment changes. For instance, at extremely early stages, the solvent state might correspond somewhat to dilute solution conditions, whereas at later stages a high density of residue contacts would hold and could affect individual conformational preferences in significant ways.

The quest for simplified representations of protein structure and statistical potentials suitable for representing interactions at a low level of resolution has been motivated by two main factors, one theoretical and the other experimental. From the theoretical point of view, a full atomic description of protein structure, even though more rigorous, has limited applicability, in view of the time and length scales of many observed phenomena. It is impossible to explore completely phenomena of interest, such as the folding and interactions in proteins, within reasonable computation time with present computational technology. In addition to computational limitations, another reason for directing efforts to less detailed models and potentials is the fact that the semi-empirical potentials commonly used in atomic descriptions discriminate poorly between correctly folded and misfolded structures [23–25]. These potentials cannot efficiently select the native fold;

recent studies have demonstrated the further requirement for a sufficiently pronounced energy minimum [26,27•]. The energy minimum also needs to be surrounded by relatively low potential-energy barriers for the protein to fold into the native state [28•]. These requirements are not usually satisfied by conventional atomic potentials. A lower resolution description can reduce the number of folding states and smooth out the barriers between local minima, and is therefore of fundamental utility in simulations. One recent study has investigated the correlation between the energy level and the root mean square (rms) deviations from the native form [29•].

From the experimental point of view, more than 1000 protein structures have now been resolved to less than 2.5 Å by X-ray crystallography, offering a wealth of information on long-range interaction preferences and the propensities of individual amino acids. The experimental data may be used for extracting knowledge-based potentials of mean force which, after proper averaging, may yield effective free energies associated with various inter-residue contacts in globular proteins. Since the original work of Miyazawa and Jernigan [4], the number of non-homologous protein structures deposited in the Protein Databank (PDB) [30,31] increased by one order of magnitude [32]. A total of 1661 protein subunits were considered by Miyazawa and Jernigan in a recent re-evaluation of effective inter-residue contact energies [20••]. This number may be compared to the 42 structures explored in their original work [4].

The larger sample size does not substantially change the original set [4] of inter-residue potentials, but rather confirms its validity [20••]. The utility of these effective potentials has been demonstrated in a number of studies (reviewed in [33]), including the selection of good conformations from large sets of conformations [34], the evaluation of different sequences threaded onto known structures [20••], the assessment of sequence similarities [35], the prediction of the effects of amino acid substitutions on stability [36], and the prediction of binding peptides to a specific MHC complex [37].

Other researchers have derived contact potentials for atoms from structures in a similar way (C Zhang, G Vasmatzis, JL Cornette, C DeLisi, personal communication). A somewhat different approach for calculating energies, especially atomic surface interaction energies, has been to base them directly on the extent of mutual surface between two molecules [38••,39]. Interestingly, the estimate of hydrophobic potentials by Kurochkina and Lee [39], which was based on the surface area buried by an interacting pair of atoms, was shown to be strongly correlated with the contact potentials of Miyazawa and Jernigan. A recently published comparison [40••] showed that these functions differ principally in their reference states; the usual two choices correspond either to initial solvent exposure of residues or to residue exposure.

The advantages and usefulness of knowledge-based potentials are clear [41]. These inherently include perturbations due to solvent mediation. Many details are averaged out but some essential features, such as hydrophobicity, are retained. A much larger range of conformations may be explored particularly when on-lattice simulations are employed. In fact, potentials based on knowledge of residue–residue interactions have been widely used in both on- and off-lattice simulations of proteins [28•], but these energies do involve approximations and they do have limitations [42••].

The general approach and some of the assumptions involved in reduced models and in knowledge-based energy parameter estimations will now be discussed. The uses and limitations of these models in simulating proteins will be discussed in the next section, together with some recent illustrative examples. Some refinements to models and extracted potentials aimed at spanning the gap between coarse-grained structure and atomic structure will be described in the section after that.

Model, method and approximations

In the most general case, the potential of mean force W between two interaction sites A and B located in a distance range $r \pm \Delta r$ from one another is given by the Boltzmann relationship

$$W_{AB}(r) = -RT \ln [P_{AB}(r \pm \Delta r)/P_{XX}(r \pm \Delta r)] \qquad (1)$$

where $P_{AB}(r \pm \Delta r)$ is the probability of observing the specific pair [A,B] at separation $r \pm \Delta r$, $P_{XX}(r \pm \Delta r)$ is the corresponding reference probability, independent of residue type, R is the gas constant and T is the absolute temperature. In some treatments, the number of intervening residues between A and B along the chain sequence have been considered [6].

Equation 1 already entails several approximations: firstly, the choice of the reference state; secondly, the application of Boltzmann statistics; thirdly, the choice of the interaction sites to characterize residue–residue interactions; and fourthly dividing the conformational space into finite intervals of width $2\Delta r$. Two additional fundamental issues implicit in Equation 1 are that the potential energy is expressed as a sum of pairwise interactions and that each pair of specific residue types [A,B] is assumed to behave independently, regardless of the chain connectivity, constraints imposed by specific sequential neighbors, and context or environmental conditions. Finally, the experimental data set is assumed to be large enough to represent the full spectrum of inter-residue energetics manifested in protein structures.

Reference state

Some aspects of the resemblance of these protein energies to thermodynamic quantities deserve comment. Just as,

in thermodynamics, specification of the standard state is critical to knowing the meaning of values, here, the definition of the reference state is critical. Furthermore, by analogy to thermodynamics, both intensive and extensive properties have roles to play. In some cases, the sum of all the interactions within the protein is important; in other cases, such as when proteins of different sizes are compared, the energy per residue serves best. The analogous thermodynamic quantities would be the total free energy and the chemical potential, or the free energy per unit. Cases in which this intensive contribution per residue would be particularly useful are in threading, where the effects of insertions and gaps are unknown, or in the process of folding, where only part of the protein might be folded. Comparing conformations of one whole protein is simpler because of the protein's fixed size and composition. Binding is simpler because the reference state is more clearly definable as being the two molecules completely separated, as long as there are no conformational changes.

In their quasi-chemical approximation, Miyazawa and Jernigan [4,20••] introduced the random mixing approximation for describing the reference state, in which the number of contacts between a particular pair of species is directly proportional to their relative concentrations. Three types of effective contact potentials were given in conjunction with three different reference states. The first, the preference of an A-type residue for a B-type residue over their self-interactions, is expressed as

$$A{\cdot}A + B{\cdot}B \leftrightarrow 2\,A{\cdot}B \qquad (2)$$

and is accounted for by the intramolecular effective contact energy

$$e_{AB}'(r_c) = W_{AB}(r_c) - (W_{AA}[r_c] + W_{BB}[r_c])/2 \qquad (3)$$

where r_c is the cutoff separation over which averages are taken, in the original case, 6.5 Å between side-chain centers and termed broad here, and W_{AB} is the potential energy of interaction between A and B. For this reference state, the opposite charge interactions are the most favored pairs and the hydrophobic pairs manifest quite weak interactions, corresponding to strong specificity.

The more interesting scheme involves desolvation of residues A and B prior to their association, as

$$A{\cdot}0 + B{\cdot}0 \leftrightarrow A{\cdot}B + 0{\cdot}0. \qquad (4)$$

Here, '0' indicates solvent molecules. The corresponding solvent-mediated effective contact energy reads

$$e_{AB}(r_c) = W_{AB}(r_c) + W_{00}(r_c) - W_{A0}(r_c) - W_{B0}(r_c). \qquad (5)$$

The solvent-residue potentials, $W_{A0}(r_c)$ and $W_{B0}(r_c)$, are determined by assuming the coordination of residues A and B to be completed by solvent molecules whenever the respective numbers of non-bonded side chains in their 'broad' neighborhood (within a spherical shell of radius $r_c \approx 6.5$ Å) is smaller than those observed in their fully coordinated, buried state.

Sample values of e_{AB}(broad) are given in Table 1 [20••].

Table 1

Sample values of e_{AB}(broad) for residue types phenylalanine (F), leucine (L), alanine (A), glutamic acid (E), and lysine (K).

e_{AB}(broad)*	F†	L†	A†	E†	K†
F	–7.26	–7.28	–4.81	–3.56	–3.36
L		–7.37	–4.91	–3.59	–3.37
A			–2.72	–1.51	–1.31
E				–0.91	–1.80
K					–0.97

*e_{AB} is the solvent-mediated effective contact energy between residues A and B. †All values are given in RT units, where R is the gas constant and T is the absolute temperature.

Here, because exposure to water was the reference state, large favorable hydrophobic interactions are seen.

In a third scheme, the interactions replaced in the reaction are those with average residues. This corresponds to the transition

$$A{\cdot}r + B{\cdot}r \leftrightarrow A{\cdot}B + r{\cdot}r. \qquad (6)$$

Here, 'r' indicates an average residue. The corresponding residue-mediated effective contact energy is

$$e_{AB}''(r_c) = W_{AB}(r_c) + W_{rr}(r_c) - W_{Ar}(r_c) - W_{Br}(r_c). \qquad (7)$$

The values for the same residue types are given in Table 2.

Table 2

Sample values of e_{AB}''(broad) for residue types phenylalanine (F), leucine (L), alanine (A), glutamic acid (E), and lysine (K).

e_{AB}''(broad)*	F†	L†	A†	E†	K†
F	–0.29	–0.26	0.03	0.44	0.37
L		–0.30	–0.08	0.46	0.41
A			–0.13	0.30	0.23
E				0.12	–1.04
K					–0.48

*e_{AB}'' is the residue-mediated effective contact energy between residues A and B. †All values are given in RT units, where R is the gas constant and T is the absolute temperature.

With this reference state, the energy scale is intermediate between the previous two cases, showing both favorable hydrophobic and electrostatic interactions.

It is critically important to consider the unfolded state; although this denatured state is usually poorly characterized, it is plausible to postulate, however, that it is intermediate between complete exposure of residues to solvent and complete burial, that is, interactions with average residues. The exposed case might be appropriately represented by the process in Equation 4 and the buried case by Equation 6. Consequently, it seems quite general to postulate that the process overall could have new contact energies formed in the native state as some average of the two energies in Equations 5 and 7 above. Hence we would define a folding potential as a mixture of two fractional contributions

$$\epsilon_{AB}(r_c) = x\, e_{AB}(r_c) + (1-x)\, e_{AB}''(r_c). \qquad (8)$$

This actually corresponds to a definition of the denatured state as having an initial fraction x of residues A and B exposed to water and the remaining fraction $(1-x)$ randomly buried.

Park and Levitt [43••] have recently demonstrated the superiority of such a combination for picking out native conformations (equivalent to $x=0.5$) over either type of energy reference state individually. Their success raises several questions. Should x be different for different proteins? Should x be different for various residue types: for example, should it depend on their hydrophobicities? Our knowledge of the denatured state is woefully limited and insufficient to tell us how to define the denatured state completely. From Tables 1 and 2 above, however, it can be seen that the relative preferences given to a specific residue pair in the folded form depend critically on this reference state.

The solvent-mediated contact potentials are particularly useful for representing the behavior of amino acids exposed to solvent, and therefore could be more appropriate at the initial stages of folding, whereas the intramolecular contact potentials e″ would be more appropriate for portraying interactions between residue pairs buried in the core. This raises the possibility that folding simulations and the potentials also could change progressively, with less and less water in the reference state as folding proceeds, that is, $x \rightarrow 0$.

Boltzmann statistics

In a strict sense, Boltzmann statistics apply if the radial distribution of residues represents an equilibrium ensemble, and the native state is at thermodynamic equilibrium. Knowledge-based potentials then ought to yield the lowest free energy for the native conformation. Threading experiments indicate that the conformational energies of amino acid sequences evaluated for a number of alternative folds indeed assume their lowest values when the sequences are in their native conformation. This type of threading, referred to as the 'sequence recognizes structure' protocol [44••], has been applied extensively [9,10,16,20••,45–47]. Also, a 'structure recognizes sequence' protocol has been used, which is equivalent to an inverse protein folding analysis [14,26,48,49]. Proteins with low levels of sequence identity, for which classical sequence alignment methods are not applicable, can be subjected to this type of screening test to locate structural homologies. In the former test, the sequences would correctly recognize their native fold on the basis of potentials. In the latter, the sequence that best fits a given three-dimensional structure would be detected, dependent on the potentials. Successes for both applications lend support to the validity of Boltzmann equilibrium statistics.

Interaction sites

The choice of interaction sites depends on the degree of complexity one adopts in the model. The most common approach has been to represent each residue by a single interaction site. This representation is particularly useful in on-lattice simulations. The single site per residue could be identified with the side-chain centroids or the Cα atoms or the Cβ atoms, etc. More detailed descriptions of protein structure distinguish between backbone and side-chain groups. The main chain is typically represented by virtual bonds connecting Cα atoms, following the model introduced by Brant, Miller and Flory [50], which was further developed and used in simulations [1,51]. The side chain is usually represented by one or more points. The recent comparative study by Kocher, Rooman and Wodak [44••] demonstrates that the knowledge-based potentials computed from the separations between average side-chain centroids perform significantly better in threading tests than those computed from inter-Cα or inter-Cβ distances. In a recent study (I Bahar, RL Jernigan, unpublished data), the side-group interaction centers have been determined on the basis of a selection of multiple-side group atoms expected to be subject to the most distinctive interactions. Different atoms were selected for each type of amino acid, with a tendency for them to be near the side-chain terminus. The potentials of mean force were evaluated on the basis of multiple interactions taking place between the selected atoms. There are two advantages in adopting such multiple site correlations. Firstly, strongly attractive or repulsive specific interactions involving particular atoms are explicitly taken at their actual locations and not smoothed out. An extreme approach in this direction was that of Godzik and Skolnick [52], in which residue contacts were identified on the basis of the closest approach of any pair of atoms. Secondly, the use of multiple interactions substantially enlarges the sample size and enhances the smoothness of the data.

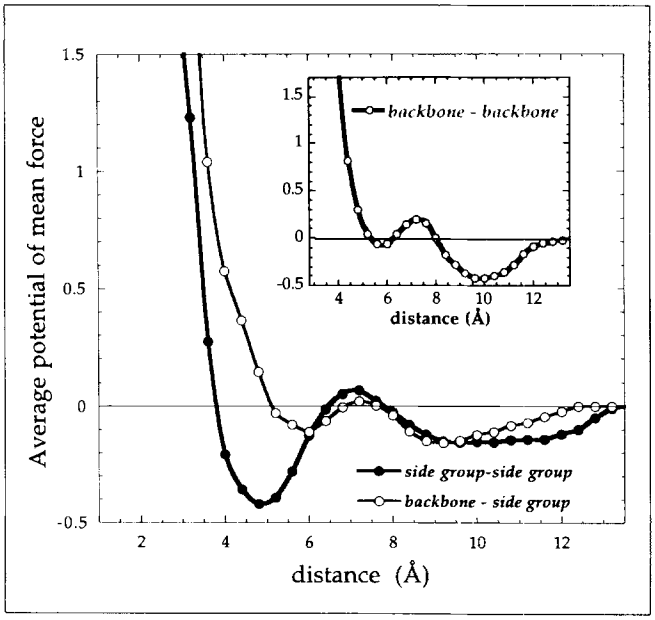
The division of conformational space

Discretization makes the problem computationally more tractable. It has been argued in several recent studies

that the protein folding problem is NP-hard in spite of minimal frustration arguments [53]; in other words, the complexity of the problem increases exponentially, and not via a polynomial (NP), with increasing molecular size. Thus, it is only by considering discrete states that a substantial portion of the conformational space can possibly be sampled. Such discretization is a necessity for lattice simulations. For off-lattice simulations, the discretization of the geometric variables places an upper boundary on the accuracy achievable in computations. Another issue that is not usually considered is how to modify empirical contact energies to adjust for different lattice types.

In adopting a low-resolution model, a compromise must be reached between sufficient flexibility to afford a reasonable account of the structural characteristics of the protein and sufficiently limited degrees of freedom to reduce the conformational space. In some lattice models, a high coordination number, up to 90 possible ways of choosing a vector connecting two consecutive α-carbon atoms in the absence of restrictions, has been adopted. Crystal structures from the PDB [30,31] may be projected onto such a lattice with an average rms deviation of 0.6–0.7 Å [54]. Yet such systems can be explored only by Monte Carlo (MC) methods. In the other direction [55•], a diamond lattice, the three-dimensional lattice with the smallest coordination number, was used for mounting protein structures. The possibility of achieving an exhaustive generation of all conformations of small proteins makes this lattice particularly attractive [55•].

Figure 1



Non-specific average potentials of mean force between side group–side group (S–S), backbone–side group (B–S) and backbone–backbone (B–B) pairs of interaction centers, obtained from 150 X-ray-elucidated protein structures (I Bahar, RL Jernigan, unpublished data). Backbone interaction centers are Cα atoms. Side chains are represented by single interaction sites defined by a selected group of atoms. All S–S, B–S and B–B pairs separated by three, four and five residues along the sequence are considered, yielding a potential that is residue non-specific, characteristic of compact globular structures. Multiple minima are observable in the curves, characteristic of consecutive coordination shells of dense systems.

Representation of energies through pairwise interactions

Although the overall energy is expressed as a sum of pairwise interactions, the two-body residue potentials of mean force extracted from structures do include the average effects of other residues upon the target residue pair. The potentials represent effective interaction energies, in a bath of other closely packed amino acids. The multiple minima in the potential curves are a direct manifestation of the close packing of other residues (see Fig. 1). Accordingly, one should not necessarily expect the potentials of mean force, W_{AB}, between a specific pair of residues [A,B] to have functional forms identical to individual interaction energies, E_{AB}, that are used as input in simulations. The former W_{AB} encompass both entropic effects and perturbations induced by many-body interactions, which may be strong in dense media such as the interiors of proteins. The latter (E_{AB}) is a purely energetic effect operating between the two particles; the environment is either ignored altogether or considered as a continuum. In this respect, a criticism of all extracted potentials [42••] that is based on an inability to extract from a set of conformers the input potentials may be too extreme. There are really two points of view: either the functions are just effective and representative of the proteins, or, more strictly, one must be able to extract the individual interactions precisely.

The fact that the extracted potentials represent effective free energies in compact globular environments is likely to be the major reason for the success of these potentials in threading experiments or other exercises that detect native-like characteristics. However, this property also raises the important issue of their limited applicability to denatured structures or other more varied conformations. An approximate way of overcoming this difficulty would be to derive potentials that are designed for specific media. Distinguishing between the behavior of buried and solvent-exposed residues, as in Equations 5 and 7, by defining effective contact potentials with different reference states, is one approximate way to treat different environments. This approach can be further refined by focusing on different close-distance intervals (I Bahar, RL Jernigan, unpublished data). It is interesting that the use of distance dependence may also aid in smoothing results; such an improvement can be inferred from the simulation results of Park and Levitt [43••], in which the incorporation of distance dependence in a way similar to Wallqvist and Ullner [56] appears to aid in the discrimination of the native conformation. The process of

smoothing over distance was also employed by Crippen and Maiorov [7].

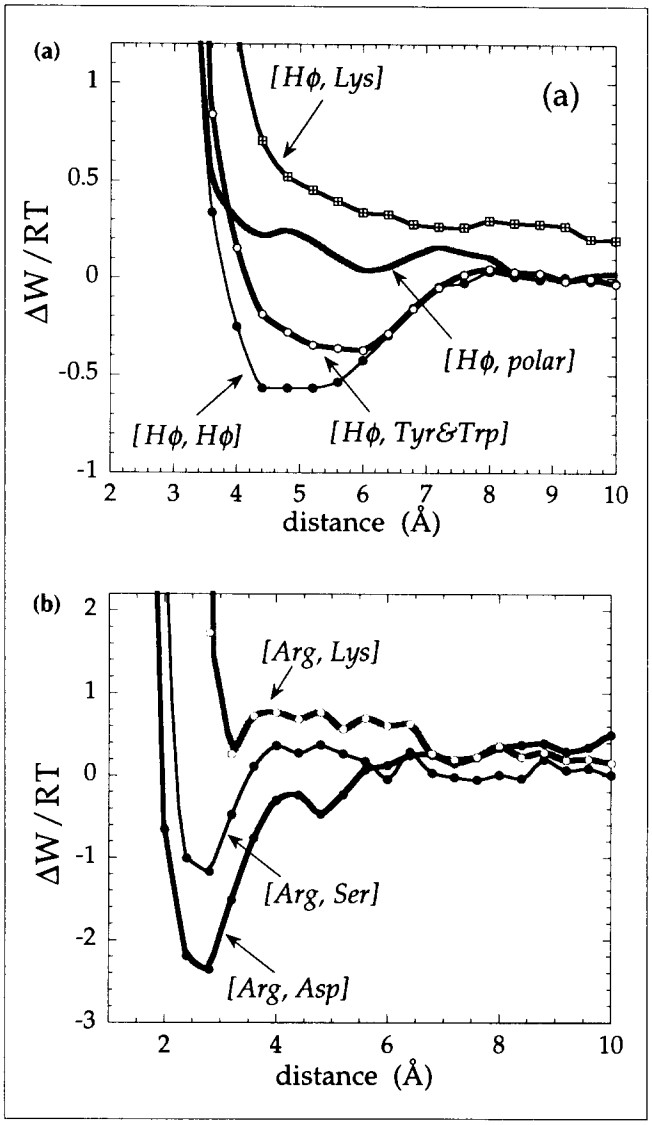
Neglect of chain connectivity and correlations between interactions

This is a basic assumption that underlies all these potentials and is also one of the reasons for observing an attractive potential between a pair of positively charged residues, even though from electrostatic considerations these ought to repel each other [20••]. The apparent attraction is a consequence of the clustering of charged groups by the dominant hydrophobic effect. Also, because of the coarse graining, there is the possibility of the interposition of favorable counter-ions or waters.

The strongest attractions are those that take place between pairs of hydrophobic side chains for a solvent-exposed reference state, when the effective potentials are over the broad distance range up to r = 6.5 Å; interactions between hydrophilic groups, on the other hand, are relatively weak [20••]. A recent closer examination of inter-residue effective contact energies (I Bahar, RL Jernigan, unpublished data) reveals that the most favorable attractive potentials between hydrophobic groups occur in the range $4 \leq r \leq 6$ Å, whereas pairs of polar and charged groups experience stronger attractions in the close interval $2 \leq r \leq 4$ Å. This feature is illustrated in Figure 2. At short distances, therefore, a different class of inter-residue potentials operates: between pairs of hydrophobic residues the interactions are highly specific and predominantly attractive, whereas between pairs of hydrophobic residues the interactions are relatively weak. This dual character of effective contact energies has important implications for the refinement of low-resolution protein structures.

The closer approach of residues leads to greater specificity. This is observed in both the distance dependence of interactions and in the surface–surface energies [38••]. In both cases, the polar pairs are stronger and much more specific. Atomic surface area formulations (for close approach) also favor polar pairs, but they exhibit an additional interesting feature: a segregation between aromatic and aliphatic carbon atoms. The origin of the difference is presumed to be that, in the broad range, hydrophobic pairs dominate because there are larger numbers of atom pairs counted, and most hydrophobic atom pairs are at least somewhat favourable, whereas in the close range, atom pairs are fewer so the stronger polar pairs are more important.

Figure 2



Specific potentials of mean force $\Delta W_{AB}(r)$, relative to the non-specific S–S potential curve displayed in Figure 1 between particular pairs [A,B]. **(a)** [Hφ,Hφ], [Hφ, polar], [Hφ, Tyr and Trp] and [Hφ, Lys] pairs. Hφ represents the group of hydrophobic residues isoleucine, leucine, valine, phenylalanine and methionine. Polar represents the polar residues asparagine, glutamine, serine, threonine and histidine. **(b)** [Arg, Ser], [Arg, Asp] and [Arg, Lys]. Results are obtained at 0.4 Å intervals, starting from 2.0 Å. The most favorable inter-residue separation decreases from 5.0 ± 1.0 Å in the case of hydrophobic residues to 2.5 ± 0.5 Å for charged or polar residues.

Simulations

Overview

In a broad sense, low-resolution simulations of proteins can be classified into two groups according to the type of model chains. The first group considers heteropolymers, or perturbed homopolymers, that consist of only a few types of monomers, or sites subject to independent interactions, on two- or three-dimensional lattices. These elementary models have provided some fundamental insights [57] into the principles governing the stability and folding mechanism of proteins (some recent applications of this approach are given in [27•,58,59•,60–64]). The main advantage of such simple models is in their being exact, in the sense that often a complete search of the conformational space for a given sequence, or sequence space for a given conformation, is possible. In the case of longer sequences that cannot be explored by complete enumeration, it is also possible to identify the most

probable conformations using folding strategies based on a hydrophobic-zipper model [65] or a constraint-based hydrophobic core construction procedure [66], as was recently performed for specific H/P (hydrophobic/polar) sequences of 48-mers on a cubic lattice [64]. Degenerate lowest-energy conformations are found, with the use of a two-letter (H/P) code for residue specification. This class of simulation is not reviewed here. For more information, the reader is referred to two excellent recent reviews [28•,57].

The second type of simulation takes into consideration more of the structural and energetic diversity of proteins at low resolution and involves on-lattice simulations, for which knowledge-based potentials are used. The overall tertiary fold, rather than local structural details, is explored in these simplified models of protein structure. Presumably, such low-resolution approaches presently offer the most promising method for investigating the global behavior of proteins with sufficient fidelity. In general, one has to choose a relatively simple model in order to restrict the conformational space, or a sufficiently accurate model in order to give a realistic description of proteins. The relationship between the complexity of a model and its accuracy was shown to be approximately of the form (accuracy) ~ (complexity)$^{-1/2}$ [67••]. There, the complexity was measured in terms of the number of conformational states per residue and accuracy refers to the rms deviation from the X-ray structure, both purely geometric quantities. On- and off-lattice models are pointed out to be of comparable complexity [67••] for a given level of accuracy. We now concentrate on such simulations.

The above two types are not so sharply defined. For example, recent simulations by Hinds and Levitt [55•] are remarkable in that they bridge the two former classes and allow for exhaustive enumeration of the complete set of conformations for real small proteins. On the other hand, attempts to combine the second class with atomic simulations [68–71] are particularly promising as a tool for predictive purposes and protein design. In some other studies, the distribution of atoms, groups of atoms, or amino-acid fragments has been considered [72–74].

Simulations with simplified models

The recent on-lattice simulations by Hinds and Levitt [55•] deserve special attention, as mentioned above. The advantage of the model they describe is that although it is computationally simple, at the same time it offers the possibility of emulating real proteins. Structures were generated on a tetrahedral lattice. Each lattice point was associated with a specific residue. Between zero and three residues were placed between adjacent lattice points, the details being determined for each conformation by a clever, simple dynamic programming algorithm. All the conformations for 11 proteins occupying up to ~40 vertices were enumerated exhaustively, each conformation being evaluated on the basis of empirical contact energies to find a locally optimal pattern of tertiary interactions. The contact energies were determined following the procedure of Miyazawa and Jernigan [4] or Bryant and Lawrence [16]. A volume constraint within an ellipsoid significantly reduced the number of generated structures. The alignment of residues between vertices and the elimination of extended conformations effectively decreased the energy.

The overall path of the chain is the principal structural property that can be captured in this lattice representation; secondary structures or side-chain orientations are difficult to reproduce. Yet, the low energy lattice models are observed to contain some native features; this suggests that it is fair to make the assumption that the native fold encoded in the amino-acid sequence is robust enough to survive even in such a low-resolution approach.

The most native-like conformations generated in the work of Hinds and Levitt [55•] typically have no more than 20% or 30% of native contacts; this number increases to about 50% in simulations carried out by Covell [75,76] on eight small proteins. Covell pointed out that about 25% of native contacts form even when hydrophobic polyleucine sequences are used, and a further increase of about 25% in the number of native contacts occurs when the specific protein sequence is considered. Miyazawa–Jernigan contact potentials were adopted in both of the studies carried out by Covell [75,76]. Constraints were placed on the size, surface area and total number of contacts. A dynamic MC scheme was adopted. The algorithm is, however, highly directed, even though each new conformation is selected, using the Boltzmann law, after evaluation of all possible moves at all positions for a given step. Covell's simulations demonstrate that a simple lattice model with effective inter-residue contact energies may ultimately find predictive application.

A higher resolution, or lower rms deviation from native structures, was achieved with more detailed lattice representations, which also included several additional energy terms [54,68,77•,78•]. The simulations of the folding of the B domain of staphylococcal protein A, ROP, crambin [78•] and GCN4 leucine zipper [68] are recent examples. The basic differences in these simulations, compared with those described above, lay firstly in finer lattice descriptions conforming more closely with the geometric characteristics of real proteins, including side-chain orientations, and secondly in more detailed potentials incorporating both local and non-local contributions, even including a many-body term.

In these studies, the Cα trace of the protein backbone was mounted on a cubic lattice, using up to 90 different types of vectors connecting the successive Cα atoms [54,68,77•,78•]. Cα traces of native proteins were typically reproducible with an accuracy of 0.7–1.0 Å in this representation. Side chains were represented by single off-lattice

sites, whose orientation with respect to the backbone was determined from a knowledge-based rotamer library. The potentials included a Ramachandran-like potential accounting for torsions and chiralities at Cα atoms regardless of residue type; an effective hydrogen-bond energy; a side-group rotameric state energy; sequence-specific orientation potentials between side chains; a residue-specific burial energy; contact potentials for non-bonded amino acid pairs; and residue-specific cooperative pairwise interaction energies implemented when two pairs of residues were simultaneously in contact. Additionally, to prevent aggregation in unstructured clusters, the presence of too many contacts was penalized. With this model and these parameters, Kolinski and Skolnick [78•] successfully simulated the folding of the above-mentioned two simple helical proteins and of crambin, a small α/β protein, starting from random, expanded conformations.

Hierarchical approaches: combining atomic and less detailed information

The approach of Kolinski and Skolnick [77•], further refined with a molecular dynamics (MD) simulation protocol, has been applied to the GCN4 leucine zipper [68]. Here, the final structure exhibits a backbone rms deviation of only 0.81 Å from the crystal structure, which is remarkable. Thus, a hierarchical algorithm combining the results from MC lattice dynamics with restrained MD-simulations to produce full atomic models, and a subsequent MD simulated annealing algorithm with explicit water, was demonstrated to yield high-resolution coiled coils [68]. The method was further used in the prediction of the quaternary structures of coiled coils [54]. The pattern of hydrophobic and hydrophilic residues alone is pointed out to be insufficient to define a protein's three-dimensional structure. The importance of specific side chain packing preferences and the entropy reduction in higher order multimers due to side chain burial was emphasized. We also note that a knowledge-based approach, utilizing pairwise residue correlations in heptad-repeat positions of coiled coils, was recently developed by Berger *et al.* [19]. Their approach successfully distinguished coiled coils from simple α helices.

A hierarchical approach has also been adopted in the recent off-lattice simulations of Gunn *et al.* [69] and Monge *et al.* [71]. The original structures for simulations consisted of sequences of cylinders and spheres. Cylinders represented the helices and β strands, and spheres represented the loops. The model thus implicitly assumes that secondary structural elements are formed prior to the tertiary organization of the protein. Conformations were generated by changing the dihedral angles of residues in the loop regions from a discrete set of rotamers. These were evaluated using the Casari–Sippl hydrophobic potentials [11]. The method was applied to two small proteins: the four α helix bundle myohemerythrin and cytochrome *b*562 [70], demonstrating that the basic topology of the four-helix bundle was recovered in the lowest energy conformers. However, in the application of the same methodology to the C-terminal fragment of the L7/L12 ribosomal protein, a β-sheet-containing segment [71], an *ad hoc* energy function was added to favor the antiparallel configuration of the β strands. The lowest energy structure in this simulation showed an rms deviation of 5.0 Å from the native structure.

Another hierarchical procedure has been proposed recently by Srinivasan and Rose [79] to predict the folds of proteins. Their algorithm was generated on the basis of constraining the conformations of local units of structures that appear to be independently nucleated during the random sampling of conformational space. Although good results were obtained for the secondary and supersecondary structures of 50-residue fragments in seven X-ray-elucidated proteins, this algorithm cannot yield all tertiary folds.

Parameters

There have been inconsistencies in the potentials used in simulations. For example, in many cases, redundancies and couplings between different contributions are neglected; each contribution is weighted with a different, somewhat arbitrary, scaling factor. We note the duplicate inclusion of some specific interactions. Some researchers [54] pointed out that it was necessary to reduce the strength of short-range interactions in order to prevent trapping within local minima indicated, and others indicated that scaling factors were required to correct for the incomplete separation of the contributions in various potential terms [78•].

Monge *et al.* [71] fixed the secondary structure and employed a simple knowledge-based potential for simulations. The idea of fixing secondary structures originated in the work of Ptitsyn and Rashin [80]. This assumption has two drawbacks. Firstly, completely predictive calculations are not possible, because *a priori* knowledge of secondary structure is required. Secondly, the simulated folding pathway may not be physically realistic; that is, it is unlikely that all secondary structure is formed prior to all tertiary packing. It is widely held that the overall hydrophobic collapse occurs first, driving the formation of secondary structure. Locking in all portions of the chain, except for the loops, is clearly an oversimplification.

Recent refinements in potentials

Recently, Šali, Shakhnovitch and Karplus [59•], together with others, have suggested that the lack of a suitable potential function, rather than the design of a folding algorithm, could be the bottleneck in structure prediction. In addition, the combination of potentials extracted from different structural and energetic considerations, being either short range or long range, has proven useful [13,15,44••,46,68,77•,81].

Among the major issues related to obtaining more accurate potentials, cognate to low-resolution models, are the

following: effective inter-residue contact potentials applicable at finer distance intervals in different environments; short-range conformational correlations governing the coupled torsional behavior of consecutive bonds on the main chain; backbone–side group interactions; and packing and coordination of side chains. Recent developments concerning these issues will now be presented.

A new generation of effective contact potentials

The attractive inter-residue contact energies recently re-evaluated by Miyazawa and Jernigan [20••] were derived from a significantly larger set of protein crystal structures, but results are nearly identical to those previously determined [4]. An additional repulsive inter-residue energy was introduced in this study, which consisted of a non-specific hard core potential and a soft residue-specific packing potential that was operative when the number of contacts surrounding an amino acid exceeded a threshold value. This many-body term serves to offset the physically unrealistic overly dense clustering of residues, which is observed as a difficulty in simulations that use only attractive contact potentials.

In an examination of the distance dependence of effective contact potentials (I Bahar, RL Jernigan, unpublished data), we determined inter-residue potentials using the radial distribution function $g_{AB}(r)$ for pairs of side chains A and B separated by three or more residues. Pair radial distributions differ from probabilities or directly counted frequencies in that the observed frequencies are normalized with respect to the radial distance. Precisely, the observed number $N_{AB}(r \pm \Delta r)$ of neighbors of type B, located in a given spherical shell of thickness Δr centered about A, is divided by the volume of that shell, $4\pi r^2 \Delta r$. This avoids the overweighting of neighbors in more distant, larger sized volume elements. The radial distribution function $g_{AB}(r)$, as r increases, approaches unity in pure systems, or the product of the equilibrium mole fractions of species A and B in multi-component mixtures. The corresponding potential of mean force is $W_{AB}(r) = -RT \ln g_{AB}(r)$ [82], which vanishes at large separations after normalization with respect to composition. Our potentials become negligibly small beyond $r \approx 13$ Å, in contrast to those of Sippl [6], where strong physically unrealistic preferences persist over larger separations.

The average potential $W_{XX}(r)$, existing between all pairs of X side-chain groups, is the reference state in evaluating the potential of mean force $\Delta W_{AB}(r) \equiv W_{AB}(r) - W_{XX}(r)$ specific to the pair of residues [A,B]. $W_{XX}(r)$ was shown in Figure 1 for three types of interactions: S–S, S–B and B–B, where B and S refer to backbone and side-chain sites respectively. $W_{XX}(r)$ may be viewed as a homogeneous sequence non-specific contribution, driving the overall compactness of the globular structure. The passage to intramolecular effective contact energies defined by Equation 3 is made by

$$e'_{AB}(r_c) = -RT \ln \int_0^{r_c} \bar{g}_{AB}(r)\, dr \left[\int_0^{r_c} \bar{g}_{AA}(r)\, dr \int_0^{r_c} \bar{g}_{BB}(r)\, dr \right]^{-1/2} \quad (9)$$

where $\bar{g}_{AB}(r)$ is the normalized radial distribution function. The evaluation of $e_{AB}(r_c)$, on the other hand, requires consideration of the average coordination numbers of residues in order to estimate the effective solvent–residue potentials with a mean-field approximation.

We evaluated the effective contact energies $e_{AB}(r_c)$ and $e_{AB}'(r_c)$ for two distance ranges, $2.0 \le r \le r_c = 6.4$ Å and $2.0 \le r \le r_c = 4.0$ Å, referred to as the 'broad' and 'close' distance ranges. The first interval matches that considered by Miyazawa and Jernigan [4,20••] in their derivation of effective contact potentials. Our $e_{AB}(r_c)$ and $e_{AB}'(r_c)$ values in this range show a close correspondence to those of Miyazawa and Jernigan: the respective total correlation coefficients are 0.98 and 0.94, confirming the consistency of the two studies. The basic feature in the broad distance regime is the predominance of hydrophobic interactions. However, this behavior does not persist into the close distance range, $r \le 4.0$ Å. The most striking differences between the two sets of effective potentials operating in the two distance intervals, observed after superposition of the values obtained for glycine–glycine pairs in the two ranges (I Bahar, RL Jernigan, unpublished data), are summarized below.

The solvent-mediated effective contact potentials $e_{AB}(r_c)$ between hydrophobic groups in the broad distance regime are of the order of –5.5 RT; this decreases to approximately –1.5 RT in the close range. For charged residues, the situation is reversed: the oppositely charged side chains are subject to the strongest interactions in the close regime, with interactions of the order of –6.0 RT. The strongest attraction (–7.02 RT) occurs between arginine and glutamic acid. The pairs of side chains with like charges are also favored, being about –4.0 RT. Interactions between polar and charged side chains also emerge as an important group of highly attractive potentials at close separations. The corresponding effective contact potential is about 2 RT more favorable than that occurring in the broad distance range. Sample values for the close range are given in Table 3.

As to the intramolecular effective contact potentials $e_{AB}'(r_c)$, the energy range of effective potentials broadens from 2.3 RT to 5.5 RT as one shifts from the broad to the close regime. This indicates a significant increase in specificity within the close distance range. The pair Lys–Glu is subject to the strongest attraction, –2.3 RT. Interestingly, pairs involving tyrosine such as Tyr–Gly, Tyr–His, Tyr–Pro, Tyr–Val and Tyr–Leu experience significantly more favorable contact energies in the close range, compared to their potentials in the broad range. His–Asp and Asp–Gly are two other pairs distinguished

Table 3

Sample values of e_{AB}(close) for residue types phenylalanine (F), leucine (L), alanine (A), glutamic acid (E), and lysine (K).

e_{AB}(close)*	F†	L†	A†	E†	K†
F	–0.64	–0.66	–2.21	–0.71	–1.02
L		–0.82	–2.12	–2.05	–1.58
A			–3.34	–3.09	–2.42
E				–3.72	–5.74
K					–3.14

*e_{AB} is the solvent-mediated effective contact energy between residues A and B. †All values are given in RT units, where R is the gas constant and T is the absolute temperature.

by an enhanced attraction at short separations, whereas Met–Phe, Ser–Phe, Pro–Ile and Tyr–Phe represent some examples of opposite character, that is, pairs whose intramolecular effective contact potential becomes less favorable at shorter distances.

The same residue types have the values for the residue-mediated reference state at close approach given in Table 4.

Table 4

Sample values of e_{AB}''(close) for residue types phenylalanine (F), leucine (L), alanine (A), glutamic acid (E), and lysine (K).

e_{AB}''(close)*	F†	L†	A†	E†	K†
F	–1.07	–0.55	–0.85	1.15	0.48
L		–0.67	–0.23	0.34	0.45
A			–0.20	0.55	0.86
E				0.44	–1.96
K					0.28

*e_{AB}'' is the residue-mediated effective contact energy between residues A and B. †All values are given in RT units, where R is the gas constant and T is the absolute temperature.

We note that the classification of amino acids into two crude groups, H (hydrophobic) and P (polar), is inadequate even from an examination of the contact potentials listed above for only five residues. Here, phenylalanine and leucine would naturally belong to the H group and the charged residues glutamic acid and lysine would necessarily be assigned to the P group. If such a classification were adequate, we could expect the values in the upper left-hand corner of Tables 1–4, which represent the H–H interactions, to be approximately equal to each other, and the values in the lower right-hand corner, which are representative of the P–P interactions, to be comparable as well. In addition, the four values in the upper right-hand corner, representing the H–P interaction terms, would be expected to be approximately equal to one another. We observe that these requirements are not fulfilled. In particular, the residue-mediated contact energies, e_{AB}'' for P–P interactions have the values 0.44 RT, –1.96 RT and 0.28 RT for Glu–Glu, Glu–Lys and Lys–Lys pairs respectively, which could not be approximated by a single value. Also, the solvent-mediated contact energies e_{AB} for members of the H–P group exhibit significant diversity. For example, e_{AB} for phenylalanine–glutamic acid, –0.71 RT, is closer to the average value for H–H interactions, than to the other H–P contact energies.

Short-range conformational potentials

Effective contact potentials cannot adequately represent the conformational preferences of the main chain on a local scale. These are controlled by interactions operating between near neighbors along the chain backbone. Also, the chain chirality ought somehow to be introduced in low-resolution simulations [68,76,77•,79].

Several studies have been carried out that deal with the torsional preferences of backbone bonds [13,15,81,83,84] (I Bahar, RL Jernigan, unpublished data). In particular, the Cα–Cα virtual bond formalism has proven useful in treating short-range conformational statistics. This model [1] was proposed long ago and was revisited in numerous studies. A recent study has demonstrated again the utility of adopting such a virtual bond representation with a small number of states per residue for efficiently generating conformations that satisfactorily fit crystal structures [67••]. Our recent unpublished work shows that an advantage is that the virtual bond angles and virtual torsional angles exhibit highly correlated bimodal distributions, which can characterize secondary structures well. In addition, strong coupling between consecutive bond torsions is observed [81]. Such pairwise correlations provide the basis for constructing short-range conformational energies in Markov chains [21,22]. The short-range interactions are different in origin from the contact potentials operating between non-bonded neighbors. One can, therefore, in principle, combine the two contributions without including redundant interactions. The potentials we obtained (I Bahar, RL Jernigan, unpublished data) for the pairwise interdependent torsions of two virtual bonds adjacent to a given residue type A are correlated strongly with the secondary structure propensities of the particular residue type. It was demonstrated that α-helix and even β-sheet preferences, which depend on context [85], are correlated to a significant extent with the doublet energies for pairwise coupled torsions. The determinants of secondary structure have recently been studied [86•].

Backbone–side-group and backbone–backbone interactions

Potentials of mean force between side-chain (S) and backbone (B) interaction sites (S–B), and between pairs of Cα atoms (B–B) have been evaluated in much the same way as the potentials between pairs of side chains. S–B interactions are residue specific (I Bahar, RL Jernigan, unpublished data). The average behavior of all residues is displayed in Figure 1. These are obtained by considering

S–B pairs belonging to residues i and i+k where $k \geq 4$, and thus are classified as long-range interactions. Likewise, B–B potentials are determined on the basis of backbone Cα atoms i and i+k, where $k \geq 5$.

The side chains that display the strongest affinity for the backbone were observed to be glycine, serine and threonine (I Bahar, RL Jernigan, unpublished data). These are subject to an attractive B–S interaction of about −1.3 RT, relative to the average potential shown in Figure 1. The residues that exhibit the next most favorable interactions were lysine > arginine ≈ asparagine > glutamine ≈ aspartic acid > glutamic acid. Asparagine and aspartic acid had two energy minima, which may be attributed to the specific interactions of the two terminal atoms of their side chains. The behavior of glutamic acid was quite distinct from that of aspartic acid, despite the similarity of these residues.

Packing of side chains

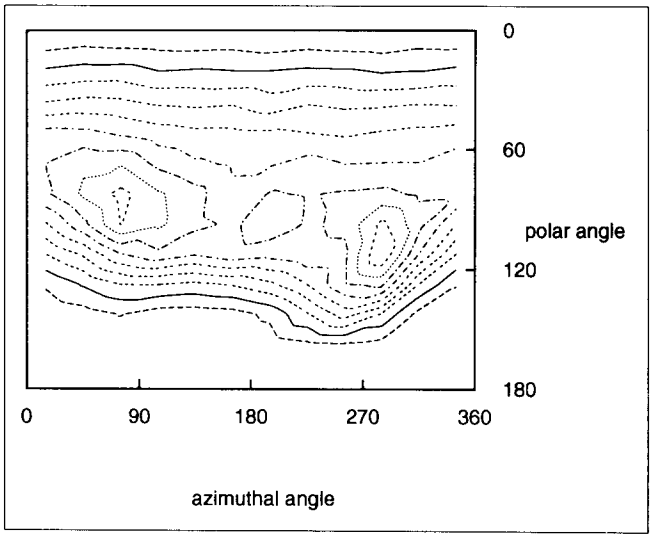
Interior packing is particularly relevant to simulations, because the reduction of angular freedom in the dense state that is imposed by packing is highly effective in reducing conformations. A highly detailed study of the packing of side chains in proteins, considering both spatial and orientational distributions, was performed by Singh and Thornton [87,88]. The existence of side chain packing specificity has been questioned [58], and elsewhere its existence has been pointed out to be the major determinant of proteins' unique structure [89,90•].

Our recent observations on PDB structures indicate that there is indeed non-random packing of side chains, and that the coordination geometry of amino acids is residue specific. Specificity exists both on a local scale and on a non-local scale. The overall non-randomness of the coordination geometry may be verified on the map shown in Figure 3. Here, all amino acids in 150 non-homologous PDB structures were examined and the angular positions of their non-bonded neighbors within $r \leq 6.8$ Å, separated by at least three intervening residues along the chain, are included. A preference for three well-defined loci is observed. This non-randomness is evident despite the fact that the cumulative behavior of all residues is shown. When coordination geometry about specific amino acids is considered, there is a substantial increase in the specificity and variability of the preferred loci. Most of the probable coordination loci for the twenty different amino acids, with their respective probabilities, have been determined, and these can be used in on- or off-lattice simulations of proteins (I Bahar, RL Jernigan, unpublished data).

Conclusions

Recently, the accumulation of substantial numbers of protein–DNA structures and their binding data indicates the possibility [91•] of obtaining base–amino acid potentials for those cases.

Figure 3



General coordination geometry of all amino acids in globular proteins. The map represents the observed probability distribution for the two spherical angles specifying the location of side chains in the neighborhood ($r \leq 6.8$ Å) of a given side chain. The polar ϑ_i and azimuthal φ_i angles are defined as follows. ϑ_i is the angle between b_i and r_{ij}, where b_i is the virtual bond vector connecting $C\alpha_i$ to side chain S_i, and r_{ij} is the vector pointing from S_i to S_j. φ_i is the dihedral angle defined by atoms $C\alpha_{i-1}$, $C\alpha_i$, S_i and S_j. This angle is 0° for the *trans* position with respect to atoms $C\alpha_{i-1}$, $C\alpha_i$ and S_i. Three regions of highest probability are distinguished: $(\vartheta_i,\varphi_i)=(110°,280°)$, (80°,75°) and (95°,195°), in the order of decreasing heights of peaks.

Threadings have been used extensively for recognizing protein structures or confirming the validity of knowledge-based potentials. However, the question of the validity of threading experiments as a sufficiently stringent test of potentials has been raised [44••]. Thomas and Dill [42••] asked whether only two residue types, hydrophobic and polar, are sufficient. However, examination of the contact energies for the five illustrative residues presented above demonstrates the impossibility of combining the behavior of all residues into two classes. Demonstrations of success with fewer classes of residues may indicate that threading is not the most demanding test of potentials. The fact that sequences correctly recognize their native fold does not necessarily indicate that the extracted potentials are sufficient to discriminate effectively against all possible non-native folds in simulations. Success in threading using limited residue types does not prove that no additional types are required.

Combining potentials associated with different degrees of freedom, incorporating the specific preferences on both local and global scales, seems important so long as this is done in a self-consistent way. Figure 1 gives a description of non-specific interactions in globular proteins. These are clearly important in driving the overall condensation of the protein and the formation of native-like secondary structures. Yet they are not residue specific and hence cannot distinguish between folds exhibiting the same

compactness and secondary structure. The unique conformation of globular proteins is, in fact, determined by specific interactions, some of which are illustrated in Figure 2. Furthermore, side-chain packing seems to be an important additional determinant of particular tertiary folds. Lumb and Kim [90•] call attention to the important role of specific interactions between buried polar groups in imparting structural uniqueness. They suggested that non-specific hydrophobic interactions do contribute to protein stability, but that structural uniqueness is imparted by the requirement of satisfying hydrogen bonds by buried polar groups in the hydrophobic environment. This is an important observation which motivates a rigorous analysis of the packing of side chains, and particularly of buried polar groups.

As pointed out above, potentials operating between pairs of side chains, $W_{AB}(S–S)$, may be expressed in terms of two contributions: a predominantly attractive homogeneous part, $W_{XX}(S–S)$ for all pairs of amino acids (shown in Figure 1), and a residue-specific contribution, $\Delta W_{AB}(S–S)$ (illustrated in Figure 2), imparting much of the unique sequence–structure correspondence. Our recent analysis of these two contributions (I Bahar, RL Jernigan, unpublished data) demonstrated that the homogeneous contribution to the overall S–S energy is stronger than that of specific interactions by about a factor of five. A contribution of about –2.1 RT per residue is induced by the homogeneous interactions between side chains. A similar scaling was also apparent from examining B–S interactions, confirming that specific interactions are substantially weaker in magnitude than non-specific potentials. This weaker contribution is essential, however, for selecting the correct fold from various compact forms. We have also observed (unpublished data) that the overall interaction energy of native structures increases with the size of the protein by $n^{1.28}$, where n is the number of residues in the protein. This implies that larger proteins possess enhanced stability. This is consistent with observations of more frequent disulphide bridges in smaller proteins to compensate for their weaker non-bonded energies.

The analysis of effective contact energies for two different distance regimes is important for gaining an understanding of interactions at different inter-residue separations. At close distances, that is, $r \leq 4$ Å, specific interactions between pairs of hydrophilic residues are predominantly important, whereas at greater separations, hydrophobic interactions supersede in this role. The latter are much stronger and effectively dominate the apparent behavior over the broad range $r \leq 6.5$ Å. These observations have important implications as far as the simulations with low-resolution models are concerned. Broad distance potentials, such as those of Miyazawa and Jernigan [4,20••], have proven useful in numerous studies for recognizing native-like folds. However, a finer level of description may possibly be achieved by the subsequent use of the new close distance effective potentials, provided that a relatively compact native-like intermediate structure has previously been attained with the broad distance potentials. Thus, refinement using this second class of contact potentials may help bridge the gap between low resolution and atomic resolution.

An important overall question is how much detail is needed in potentials for simulations: the examples here show that two residue types are not sufficient. The validation of simple contact potentials for additional proteins [20••] shows that the differences among individual values are real and no longer dependent on the amount of available structural data. The close values show that the close approach of polar side chains is extremely important. The level of detail in a given simulation ought to be commensurate, as far as is possible, with the details in the applied potential functions. Also, the reference state should reflect, as nearly as possible, the details of initial interactions.

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. Levitt M: **A simplified representation of protein conformations for rapid simulation of protein folding.** *J Mol Biol* 1976, **104**:59–107.

2. Jernigan RL: **Protein folds.** *Curr Opin Struct Biol* 1992, **2**:248–256.

3. Tanaka S, Scheraga HA: **Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.** *Macromolecules* 1976, **9**:945–950.

4. Miyazawa S, Jernigan RL: **Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation.** *Macromolecules* 1985, **18**:534–552.

5. Lüthy R, Bowie JU, Eisenberg D: **An assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356**:83–85.

6. Sippl MJ: **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.** *J Mol Biol* 1990, **213**:859–883.

7. Crippen GM, Maiorov VN: **Contact potential for global identification of correct protein folding.** In *The Protein Folding Problem and Tertiary Structure Prediction.* Edited by Merz KM Jr, Le Grand SM. Boston: Birkauser; 1994:231–277.

8. Bowie JU, Lüthy R, Eisenberg D: **A method to identify protein sequences that fold into a known 3-dimensional structure.** *Science* 1991, **253**:164–170.

9. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ: **Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force.** *J Mol Biol* 1990, **216**:167–180.

10. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition.** *Nature* 1992, **358**:86–89.

11. Casari G, Sippl MJ: **Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of**

globular proteins is able to identify native folds. *J Mol Biol* 1992, **224**:725–732.

12. Sun S, Luo N, Ornstein R, Rein R: **Protein structure prediction based on statistical potential.** *Biophys J* 1992, **62**:104–106.

13. Sun S: **Reduced representation model of protein structure prediction: statistical potential and genetic algorithms.** *Protein Sci* 1993, **2**:762–785.

14. Willmans M, Eisenberg D: **Three-dimensional preferences from residue-pair preferences.** *Proc Natl Acad Sci USA* 1993, **90**:1379–1383.

15. Nishikawa K, Matsuo Y: **Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies.** *Protein Eng* 1993, **6**:811–820.

16. Bryant SH, Lawrence CE: **An empirical energy function for threading protein sequence through the folding motif.** *Proteins* 1993, **16**:92–112.

17. Brauer A, Beyer A: **An improved pair potential to recognize native protein folds.** *Proteins* 1994, **18**:254–261.

18. • Bowie JU, Eisenberg D: **An evolutionary approach to the folding of small α-helical proteins from sequence information using an empirical, guiding fitness function.** *Proc Natl Acad Sci USA* 1994, **91**:4436–4440.

This work selects fragments on the basis of their sequence compatibilities and then links them to form hundreds of starting structures that are then put through an evolutionary algorithm. This consists of random changes in conformation as well as cutting and pasting (recombination) for choices of segments. Success in finding native small helical proteins at 2.5–4.0 Å rms deviations is reported.

19. 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.

20. •• Miyazawa S, Jernigan RL: **Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J Mol Biol* 1996, **256**:623–644.

The procedures in [4], to derive residue–residue potentials from counting interactions in a set of protein structures, are applied to 1661 protein subunits. A useful weighting scheme based on sequence identities is presented. This yields values of interaction parameters nearly identical to old values, except in the cases of tryptophan, methionine and leucine. Also, an unfavorable residue-packing term is derived that depends on the number of neighbors around a central residue when the number exceeds a threshold value. The application of the potential to threading yields the correct native form for 73 of 89 representative proteins. A new reference state is developed which is appropriate for use of these types of residue–residue contact potentials in threading.

21. Flory PJ: *Statistical Mechanics of Chain Molecules.* Oxford: Hanser Publishers; 1988.

22. Mattice WL, Suter UW: *Conformational Theory of Large Molecules. The Rotational Isomeric State Model in Macromolecular Systems.* New York: John Wiley & Sons, Inc; 1994.

23. Novotny J, Bruccoleri RE, Karplus M: **An analysis of incorrectly folded protein models. Implications for structure predictions.** *J Mol Biol* 1984, **177**:788–818.

24. Novotny J, Rashin AA, Bruccoleri RE: **Criteria that discriminate between native proteins and incorrectly folded models.** *Proteins* 1988, **4**:19–30.

25. Wang Y, Zhang H, Li W, Scott RA: **Discriminating compact non-native structures from the native structure of globular proteins.** *Proc Natl Acad Sci USA* 1995, **92**:709–713.

26. Shakhnovich EI: **Proteins with selected sequences fold into unique native conformation.** *Phys Rev Lett* 1994, **72**:3907–3990.

27. • Šali A, Shakhnovich E, Karplus M: **Kinetics of protein folding: a lattice model study of the requirements for folding to the native state.** *J Mol Biol* 1994, **235**:1614–1636.

The presence of a large energy gap between the native and all higher-energy states is asserted to be the common attribute for correctly folding 27-mer sequences on a cubic lattice. A three step random search (3SRS) mechanism of protein folding is proposed that consists of rapid collapse from a random coil to a molten globule, followed by a slow, rate-determining step to find a transition state, and a final rapid folding into the native state.

28. • Karplus M, Šali A: **Theoretical studies of protein folding and unfolding.** *Curr Opin Struct Biol* 1995, **5**:58–73.

This review discusses the role of both simplified and all-atom models for the theoretical investigation of the mechanism of protein folding. Emphasis is placed on lattice simulations of short heteropolymers, which is asserted to be a useful approach for resolving the Levinthal paradox.

29. • Huang ES, Subbiah S, Tsai J, Levitt M: **Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations.** *J Mol Biol* 1996, in press.

This paper describes the testing of the authors' potential [92] against sets of conformations generated by molecular dynamics with water. Five small proteins without disulfide bridges, metals or other hetero-atom groups were tested. One thousand backbone forms were generated at 298K and 498K. Hydrophobic fitness is evaluated for each, based on burial and the number of non-polar contacts compared to random. Two considerations were used: near-native structures should have energies intermediate between the native forms and badly folded forms, and there should be a monotonic increase in energy with increase in the rms away from the native form. For the low temperature samples, more of the non-native 1000 conformers were below the native than at 498K or for threading. Overall greater rms deviations corresponded to less favorable cores but not in every case. The authors claim to be able to clearly discriminate as bad conformations that deviate by >~ 4 Å.

30. Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M: **The protein databank: a computer-based archival file for macromolecular structures.** *J Mol Biol* 1977, **112**:535–542.

31. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J: **Protein Databank.** In *Crystallographic Databases – Information Content Software Systems, Scientific Applications.* Edited by Allen FH, Bergerhoff G, Sievers R. Bonn: Data Commission of the International Union of Crystallography; 1987:107.

32. Hobohm U, Sander C: **Enlarged representative set of protein structures.** *Protein Sci* 1994, **3**:522–524.

33. Jernigan RL, Young L, Covell DG, Miyazawa S: **Applications of empirical amino acid potential functions.** In *Modelling of Biomolecular Structures and Mechanisms.* Edited by Pullman A *et al.*. Netherlands: Kluwer Academic Publishers; 1995:151–166.

34. Covell DG, Jernigan RL: **Conformations of folded proteins in restricted spaces.** *Biochemistry* 1990, **29**:3287–3294.

35. Miyazawa S, Jernigan RL: **A new substitution matrix for protein sequence searches based on contact frequencies in protein structures.** *Protein Eng* 1993, **6**:267–278.

36. Miyazawa S, Jernigan RL: **Protein stability for single substitution mutants and the extent of local compactness in the denatured state.** *Protein Eng* 1994, **7**:1209–1220.

37. Altuvia Y, Schueler O, Margalit H: **Ranking potential binding peptides to MHC molecules by a computational threading approach.** *J Mol Biol* 1995, **249**:244–250.

38. •• Wallqvist A, Jernigan RL, Covell DG: **A preference-based free energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design.** *Protein Sci* 1995, **4**:1881–1903.

Surface interactions in a set of 38 crystal structures of complexes are analyzed. In a surface area formulation of interaction energies, preferences between atom pairs across the intermolecular boundary are derived. Results give a quantitative evaluation of the relative importance of hydrogen bonding, charge pairs, and hydrophobic pairs. The latter are weaker than the first two, but there is an interesting appearance of segregation between aromatic and aliphatic carbons.

39. Kurochkina N, Lee B: **Hydrophobic potential by pairwise surface area sum.** *Protein Eng* 1995, **8**:437–442.

40. •• Godzik A, Kolinski A, Skolnick J: **Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets.** *Protein Sci* 1995, **4**:2107–2117.

This is a serious and useful comparison of several different residue–residue potential functions. A new version of those originally derived in reference [49] is presented. The authors point out that the functions can be divided into two major classes, depending on whether the reference state is the unfolded or the compact state. They make an interesting comparison between their individual interaction parameters derived from high-resolution and low-resolution X-ray structures and find a surprisingly high correlation of 0.91. However, a comparison between energy parameters derived from NMR structures and from high-resolution X-ray-derived structures gives a correlation of only 0.46.

A related finding for NMR structures is given in [20••] where the total protein stability of X-ray-derived structures is found to be consistently lower than those of NMR-derived structures with X-ray-based functions. Does this mean that solution and crystal structures are different in the strengths of their internal interactions? And does it imply that NMR structures require different interaction potentials?

41. Sippl MJ: **Knowledge-based potentials for proteins.** *Curr Opin Struct Biol* 1995, **5**:229–235.

42. •• Thomas PD, Dill K: **Statistical potentials extracted from protein structures: what is wrong with them?** *J Mol Biol* 1996, in press.
The authors apply the Miyazawa and Jernigan scheme [4] to a set of conformations generated on lattices for short chains with simple contact potentials, with the result that for the longer chains they could almost retrieve the starting contacts that were used as input. They also apply the Sippl procedure to extract distance-dependent functions. These results appear to reflect the overall characteristics of the folded forms: the positions of the minima are at the closest distance for H–H pairs, at intermediate values of distance for H–P pairs and furthest apart for P–P pairs, reflecting the overall size and globular nature of the conformations.

43. •• Park B, Levitt M: **Energy functions that discriminate X-ray and near-native folds from well-constructed decoys.** *J Mol Biol* 1996, in press.
The authors obtain 35 000–200 000 'decoys' for testing potential functions and compare many different individual energy functions. Surprisingly, the use of pairs of functions improved discrimination abilities. They find that improvements are made both with some combination of functions such as $e_{AB} + e'_{AB}$ works best and using smoothing functions with distance dependence, such as that of Wallqvist and Ullner.

44. •• Kocher J-PA, Rooman MJ, Wodak S: **Factors influencing the ability of knowledge-based potentials to identify native sequence–structure matches.** *J Mol Biol* 1994, **235**:1598–1613.
This is an excellent review on the use of knowledge-based potentials for structure–sequence recognition. It emphasizes the importance of combining short-range and long-range interactions on the basis of different structural descriptions and questions the use of threading experiments as a test for evaluating the quality of potentials.

45. Maiorov VN, Crippen GM: **Contact potential that recognizes the correct folding of globular proteins.** *J Mol Biol* 1992, **227**:876–888.

46. Wang Y, Lai L, Han Y, Xu X, Tang Y: **A new protein folding recognition potential function.** *Proteins* 1995, **21**:127–129.

47. Flockner H, Braxenthaler M, Lackner P, Jaritz M, Ortner M, Sippl M: **Progress in fold recognition.** *Proteins* 1995, **23**:376–386.

48. Hinds DA, Levitt M: **From structure to sequence and back again.** *J Mol Biol* 1996, in press.

49. Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, **91**:5803–5807.

50. Brant DA, Miller WG, Flory PJ: **Conformational energy estimates for statistically coiling polypeptide chains.** *J Mol Biol* 1967, **23**:47–65.

51. Levitt M, Warshel A: **Computer simulation of protein folding.** *Nature* 1975, **253**:694–698.

52. Godzik A, Skolnick J: **Sequence–structure matching in globular proteins. Applications to supersecondary and tertiary structure determination.** *Proc Natl Acad Sci USA* 1992, **89**: 98–102.

53. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: **Funnels, pathways, and energy landscape of protein folding: a synthesis.** *Proteins* 1995, **21**:167–195.

54. Vieth M, Kolinski A, Brooks CL, Skolnick J: **Prediction of quaternary structure of coiled coils. Application to mutants of the GCN4 leucine zipper.** *J Mol Biol* 1995, **251**:448–467.

55. • Hinds DA, Levitt M: **Exploring conformational space with a simple lattice model for protein structure.** *J Mol Biol* 1994, **243**:668–682.
The authors develop a highly simplified representation of proteins on a bounded diamond lattice. The representation allows all conformations to be generated for small proteins. Selectivity for native folds is reduced when a shuffled sequence is tested. The importance of forming a core in establishing a proteins' overall chain fold is pointed out, as well as the need for a low-resolution model tailored to the structural features commonly found in proteins.

56. Wallqvist A, Ullner M: **A simplified amino acid potential for use in structure predictions of proteins.** *Proteins* 1994, **18**:267–280.

57. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding. A perspective from simple exact models.** *Protein Sci* 1995, **4**:561–602.

58. Bromberg S, Dill KA: **Side-chain entropy and packing in proteins.** *Protein Sci* 1994, **3**:997–1009.

59. • Šali A, Shakhnovich E, Karplus M: **How does a protein fold?** *Nature* 1994, **369**:248–251.
The importance of a pronounced global minimum is demonstrated with lattice calculations.

60. Bahar I, Jernigan RL: **Cooperative structural transitions induced by non-homogeneous intramolecular interactions in compact globular proteins.** *Biophys J* 1994, **66**:467–477.

61. Bahar I, Jernigan RL: **Stabilization of intermediate density states in globular proteins by homogeneous intramolecular attractive interactions.** *Biophys J* 1994, **66**:454–466.

62. Chan HS, Dill KA: **Transition states and folding dynamics of proteins and heteropolymers.** *J Chem Phys* 1994, **100**:9238–9257.

63. Socci ND, Onuchic JN: **Folding kinetics of protein-like heteropolymers.** *J Chem Phys* 1994, **100**:1519–1528.

64. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA: **A test of lattice protein folding algorithms.** *Proc Natl Acad Sci USA* 1995, **92**:146–150.

65. Fiebig KM, Dill KA: **Protein core assembly processes.** *J Chem Phys* 1993, **98**:3475–3487.

66. Yue K, Dill KA: **Forces of tertiary structural organization of globular proteins.** *Proc Natl Acad Sci USA* 1995, **92**:146–150.

67. •• Park BH, Levitt M: **The complexity and accuracy of discrete state models of protein structure.** *J Mol Biol* 1995, **249**:493–507.
This paper constitutes a thorough analysis of the degree of accuracy one can achieve with on- or off-lattice models of varying complexity. Low complexity off-lattice models, having, for example, six accessible states per residue, are argued to be more appropriate for protein structure prediction than higher complexity on-lattice models. An optimized set of four-state models is presented, which fits native structures to an average of 2.4 Å and may preserve 85% of the native contacts occurring between residues located at a distance 8 Å from each other.

68. Vieth M, Kolinski A, Brooks CL, Skolnick J: **Prediction of the folding pathways and structure of GCN4 leucine zipper.** *J Mol Biol* 1994, **237**:361–367.

69. Gunn JR, Monge A, Friesner RA: **Hierarchical algorithm for computer modeling of protein tertiary structure: folding of myoglobin to 6.2 Å resolution.** *J Phys Chem* 1994, **98**:702–711.

70. Monge A, Friesner RA, Honig B: **An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure.** *Proc Natl Acad Sci USA* 1994, **91**:5027–5029.

71. Monge A, Lathrop EJ, Gunn JR, Shenkin PS, Friesner RA: **Computer modeling of protein folding: conformational and energetic analysis of reduced and detailed protein models.** *J Mol Biol* 1995, **247**:995–1012.

72. Vriend G, Sander C: **Quality control of protein models: directional atomic contact analysis.** *J Appl Crystallogr* 1993, **26**:47–60.

73. Delarue M, Koehl P: **Atomic environment energies in proteins defined from statistics of accessible and contact surface areas.** *J Mol Biol* 1995, **249**:675–690.

74. Wang Y, Zhang H, Scott RA: **A new computational model for protein folding based on atomic solvation.** *Protein Sci* 1995, **4**:1402–1411.

75. Covell DG: **Folding protein α-carbon chains into compact forms by Monte Carlo methods.** *Proteins* 1992, **14**:409–420.

76. Covell DG: **Lattice model simulation of polypeptide chain folding.** *J Mol Biol* 1994, **235**:1032–1043.

77. • Kolinski A, Skolnick J: **Monte Carlo simulation of protein folding. I. Lattice model and interaction scheme.** *Proteins* 1994, **18**:338–352.

This paper describes a hierarchical lattice method that applies Monte Carlo lattice dynamics with a complex potential function. First stage procedures give 4–5 Å rms fits of all heavy atoms. Subsequent refinement with a more highly articulated lattice improves the fits.

78. • Kolinski A, Skolnick J: **Monte Carlo simulation of protein folding. II. Application to protein A, ROP and crambin.** *Proteins* 1994, **18**:353–366.

This paper constitutes the application of the method in [77•] to three small proteins. The authors find results with rms values of 2.25–3.65 Å.

79. Srinivasan R, Rose GD: **LINUS: A hierarchic procedure to predict the fold of a protein.** *Proteins* 1995, **22**:81–99.

80. Ptitsyn OB, Rashin AA: **A model of myoglobin self-organization.** *Biophys Chem* 1975, **3**:1–20.

81. DeWitte RS, Shakhnovich EI: **Pseudodihedrals: simplified protein backbone representation with knowledge-based energy.** *Protein Sci* 1994, **2**:1570–1581.

82. Ben-Naim A: *Statistical Thermodynamics for Chemists and Biochemists.* New York: Plenum Press; 1992.

83. Rooman MJ, Kocher J-PA, Wodak SJ: **Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions.** *J Mol Biol* 1991, **221**:961–979.

84. Laiter S, Hoffman DL, Singh RK, Vaisman II, Tropsha A: **Pseudotorsional OCCO backbone angle as a single descriptor of protein secondary structure.** *Protein Sci* 1995, **4**:1633–1643.

85. Minor DL Jr, Kim PS: **Context is a major determinant of β-sheet propensity.** *Nature* 1994, **371**:264–267.

86. • Hunt NG, Gregoret LM, Cohen FE: **The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search.** *J Mol Biol* 1994, **241**:214–225.

This is an investigation of the importance of compactness and hydrogen bonds in stabilizing secondary structures. The authors use an energy function that is a simple combination of two terms and apply a simulated annealing minimization scheme. Compactness alone yields less than 20% of the expected amount of secondary structure. Hydrogen bonding alone gives between 66% and 90% of the expected secondary structure, depending on the definition of secondary structures. Favoring both hydrogen bonds and compactness gives nearly the experimental range of secondary structures.

87. Singh J, Thornton JM: **SIRIUS. An automated method of analysis of preferred packing arrangements between protein groups.** *J Mol Biol* 1990, **211**:595–615.

88. Singh J, Thornton JM: *Atlas of protein side-chain interactions,* vols 1,2. New York: Oxford University Press; 1992.

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

90. • Lumb KJ, Kim PS: **A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil.** *Biochemistry* 1995, **34**:8642–8648.

On the basis of experimental observations (circular dichroism, hydrogen exchange NMR, fluorescence) on designed multimeric coiled coils, the important role of specific interactions between buried polar groups in imparting structural uniqueness, at the possible expense of stability is pointed out. A single interaction between a pair of asparagines belonging to the respective monomers of a designed hetero-dimeric coiled coil was responsible for the formation of a unique structure. Variants obtained by substituting leucine for Asn14 are shown to lack structural uniqueness in spite of their higher stability and ability to form heterotetramers.

91. • Lustig B, Jernigan RL: **Consistencies of individual DNA base–amino acid interactions in structures and sequences.** *Nucleic Acids Res* 1995, **23**:4707–4711.

This is an attempt to extract interaction strengths between amino acids and DNA bases from non-structural data. Amino acid–base interaction strengths from combinatorial library binding studies on zinc fingers are compared with separate binding studies of several DNA-binding proteins. These are correlated and indicate the same strongest binding pairs. Operator DNA sequence frequencies are also analyzed together with known repressor structures to find relative strengths.

92. Huang ES, Subbiah S, Levitt M: **Recognizing native folds by the arrangement of hydrophobic and polar residues.** *J Mol Biol* 1995, **252**:709–720.