Vol 14 No 2 April 2004 pp 117–260

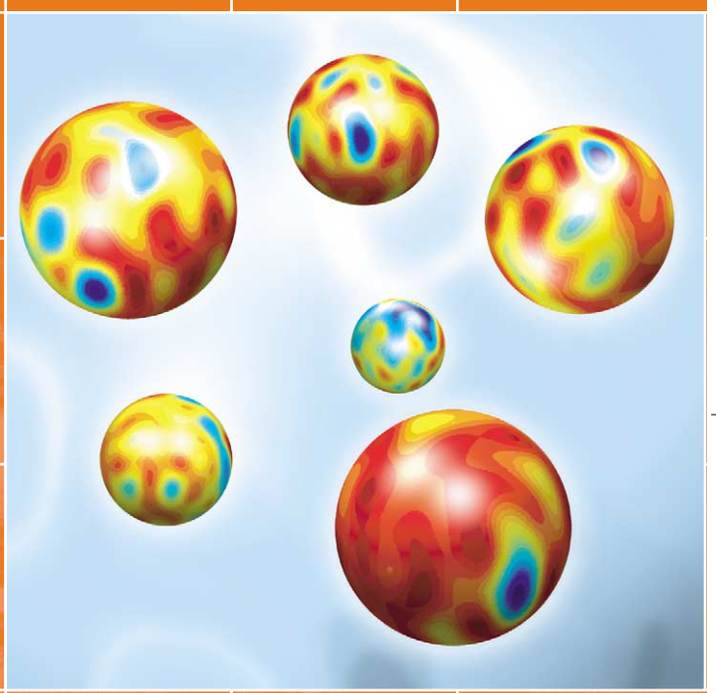
Current Opinion in Structural Biology | Macromolecular assemblages • Theory and simulation

ELSEVIER

Vol 14 No 2 **April 2004**

# Current Opinion in Structural Biology

**Wayne A Hendrickson & Tom L Blundell, Editors**



## April 2004

### Macromolecular assemblages

Edited by R Anthony Crowther and BV Venkataram Prasad

### Theory and simulation

Edited by Joël Janin and Thomas Simonson

**June 2004** Nucleic acids • Sequences and topology

**August 2004** Membranes • Engineering and design

**October 2004** Biophysical methods • Carbohydrates and glycoconjugates

**December 2004** Proteins • Catalysis and regulation

**February 2005** Protein–nucleic acid interactions • Folding and binding

# Development of novel statistical potentials for protein fold recognition

N-V Buchete[1], JE Straub[2] and D Thirumalai[3]

The need to perform large-scale studies of protein fold recognition, structure prediction and protein–protein interactions has led to novel developments of residue-level minimal models of proteins. A minimum requirement for useful protein force-fields is that they be successful in the recognition of native conformations. The balance between the level of detail in describing the specific interactions within proteins and the accuracy obtained using minimal protein models is the focus of many current protein studies. Recent results suggest that the introduction of explicit orientation dependence in a coarse-grained, residue-level model improves the ability of inter-residue potentials to recognize the native state. New statistical and optimization computational algorithms can be used to obtain accurate residue-dependent potentials for use in protein fold recognition and, more importantly, structure prediction.

**Addresses**
[1]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA
[2]Department of Chemistry, Boston University, Boston, Massachusetts 02215, USA
e-mail: straub@bu.edu
[3]Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA
e-mail: thirum@glue.umd.edu

**Current Opinion in Structural Biology** 2004, **14**:225–232

This review comes from a themed issue on
Theory and simulation
Edited by Joel Janin and Thomas Simonson

0959-440X/$ – see front matter


**DOI 10.1016/j.sbi.2004.03.002**

**Abbreviations**

| | |
|---|---|
| **BB** | backbone |
| **PDB** | Protein Data Bank |
| **Pep** | virtual interaction site located in the geometric center of the peptide bond |
| **SC** | sidechain |
| **SHA** | spherical harmonic analysis |
| **SHS** | spherical harmonics synthesis |

## Introduction

With the exponential growth of known protein structures in the Protein Data Bank (PDB, [1]), proteome- and genome-wide studies [2,3•,4•] become feasible. The emphasis on genome-wide analysis has made it urgent to devise new algorithms and methods that rely on coarse-grained yet accurate descriptions of polypeptide chains. Simple models of peptides and proteins have a significantly reduced number of degrees of freedom compared to all-atom treatments of the corresponding systems. Minimal models of proteins are widely used to obtain insights into folding mechanisms [5] of proteins, as well as in structure prediction [6–14,15••]. Improvements in the design of minimal models of proteins have made it possible to assess their accuracy by direct comparison with experiments [16••,17•,18••]. In the simplest approach, the polypeptide chains are modeled using only the α-carbon representation. Such models are useful in providing a global picture of folding and may also be used to obtain low-resolution structures. Although the simplest models allow detailed even exhaustive studies of proteins, it is increasingly clear that a certain degree of complexity is needed for more realistic applications [15••,19–22].

Reduced models of proteins have been used in protein structure prediction [23], in studies of the dynamics of protein folding [24] and, more recently, to investigate protein–protein interactions [25••]. The general strategy in many applications is to employ minimal models (typically using centers of mass of sidechains attached to Cα atoms) to obtain the topology of structures. Subsequently, a higher resolution prediction can be obtained using all-atom representations. This dual strategy has found success in the challenging *ab initio* prediction of protein structures. Minimal models were also studied in conjunction with detailed atomistic approaches [26••,27••]. The importance of minimal models is expected to increase when treating protein–protein interactions in which substantial conformational changes occur upon interface formation. This necessitates sampling a large space of conformations for both proteins — a task that can be more easily achieved using coarse-grained models. Indeed, results of the most recent structure prediction (CASP, [23]) and protein–protein interaction prediction (CAPRI, [25••]) 'community-wide experiments' emphasize the need for better methods that can probe such conformational changes. This need becomes even more important as fast and accurate automatic structure prediction servers become widely available to the structural biology community [28••,29,30].

Accurate residue-based potentials are needed to obtain reliable minimal models of proteins. The growing number of structures available in the PDB [1] has enabled the analysis of factors that control packing in folds with different architectures. Several studies have shown that pairwise contact or isotropic distance-dependent potentials are inadequate for predicting or describing sidechain

packing in folded structures [14,31–33]. These studies argue the need for obtaining anisotropic potentials or multi-particle interactions for use in coarse-grained models. Recently, it has been shown that it is also crucial to incorporate interactions between sidechains and backbone in the construction of statistical potentials [16••]. In the past few years, there has been considerable progress in obtaining reliable interaction schemes for use in structure prediction, fold recognition and folding kinetics studies. In this article, we review some of the key ideas that have contributed to the design of statistical potentials. In most applications, the efficacy of these schemes has been tested for fold recognition only. The success of anisotropic potentials can be assessed in future applications. The review concludes with a description of some of the outstanding issues that need to be addressed before statistical potentials can be most effectively employed in structure prediction and protein–protein interaction studies.

## Coarse-grained representations of polypeptide chains

### Cα models

Depending on the nature of the application, several levels of coarse-grained representations of polypeptide chains have been used [16••,17•]. In the simplest representation, the polypeptide chain is represented using only the Cα atoms [34]. Models that represent the polypeptide chain as a polymer of connected Cα atoms have been potentially useful in obtaining insights into the folding kinetics of proteins.

### Cα–SC models

To account for the various sizes and specific packing features of the 20 different types of amino acids, more detailed models are used. Dense packing of the native states can be captured using models [6,8,35] in which the backbone is described using non-interacting backbone sites located at the positions of the Cα atoms, with a second type of interaction center, $S^i$, which corresponds to each sidechain (SC). The $S^i$ interaction centers are typically located at the geometric center of the heavy atoms in each sidechain, with the exception of glycine, where it coincides with the position of the Cα atom. The use of geometric centers to represent sidechain interaction centers is a better choice than using the Cα or Cβ atoms [15••]. In the Cα–SC model, the $C\alpha^i$ sites are used to describe the backbone connectivity of the polypeptide chain structure, but only the SC interaction centers are considered to interact with each other. This type of model has been successfully used to obtain contact-based sidechain–sidechain (SC–SC) interaction potentials [10,11,36] distance-dependent potentials [9,37] and, more recently, distance- and orientation-dependent potentials [15••,16••]. These models do not include explicitly the interactions between sidechain–backbone and backbone–backbone atoms.

### Cα–SC–Pep models

The dense packing in the interior of most proteins arises through not only favorable interactions between the buried hydrophobic sidechains but also a preponderance of backbone–sidechain and backbone–backbone interactions. Analysis of a large number of single-domain protein structures reveals that, whenever two sidechains are in contact, their backbones also interact with each other with high probability [38]. Indeed, it has been recently estimated [16••] that the number of backbone–backbone (BB–BB) contacts can range from 12% to as much as 35% of the total number of SC–SC, SC–BB and BB–BB contacts depending on the protein class (e.g. α, β, mixed α/β [39]). The importance of including the backbone interactions is also supported by the results of previous statistical derivations of backbone potentials that used virtual bond and torsion angles [40], and secondary structure information [41]. Therefore, more complex models have been used [16••,20] that include an additional interaction center located on the backbone [35] at the geometric center of each peptide bond ($Pep^i$). In the resulting Cα–SC–Pep models, it is assumed that the local conformation of residue i is described by the corresponding $C\alpha^i$, $S^i$ and $Pep^i$ interaction centers.

With these coarse-grained representations of polypeptide chains, interaction potentials, including both distance and orientational dependence, can be extracted from a non-homologous subset of the PDB database of known structures using the standard Boltzmann scheme.

## Sidechain packing and orientation-dependent statistical potentials

It has been appreciated for some time that the orientation of sidechains in the native state influences the dense packing found in the interior of proteins [20,42–45]. The formation of hydrogen bonds between atoms of sidechains that are in contact requires preferred orientations. An analysis of native structures has revealed residue-specific coordination and preferred orientations of sidechains [46]. A more recent study based on orientational order parameters [15••] shows that there may be certain symmetries associated with sidechain packing. Although the preferred orientational symmetry is difficult to ascertain, it is clear that it depends on the topology of the native state. Certain classes of proteins (immunoglobulins with β-sheet topology and α-helical hemoglobins) exhibit a mixture of icosahedral and face-centered cubic arrangements. However, the orientational symmetry in myoglobins, as quantified by orientational order parameters, is different. These results show that the orientational packing of sidechains in the native state is subtle. Nevertheless, these studies point to the need for orientational potentials to describe packing in proteins.

To extract orientational and distance-dependent statistical potentials ($U_{DO}$), we assume that the known native

protein structures represent an equilibrium ensemble of 'states'. This assumption allows one to compute the statistical potentials from the distribution functions:

$$U_{\mathrm{DO}}^{\mathrm{ij}}(\mathrm{r},\phi,\theta) = -\mathrm{k_B T}\ln\left[\frac{\mathrm{P^{ij}}(\mathrm{r},\phi,\theta)}{\mathrm{P_{ref}}(\mathrm{r},\phi,\theta)}\right]. \quad (1)$$

In the above equation, $P^{ij}(r,\phi,\theta)$ is the probability of having sidechain j in a spherical volume element corresponding to distance separation r with respect to sidechain i, and with relative orientations $\theta$ and $\phi$. The computation of $P^{ij}(r,\phi,\theta)$ requires an appropriate choice of reference frame to define the local orientations of $\theta$ and $\phi$. The quality of the statistical potentials depends not only on the precise computation of $P^{ij}(r,\phi,\theta)$ but also on the reference state. Bahar and Jernigan [46], who were the first to capture relative orientation probabilities using a simple-body fixed-coordinate system, showed clearly that the preferred orientation of sidechains depends on the nature of the residue. Their results also implied that statistical potentials would be sensitive to orientations. Recently, we have introduced [16••] a new way to calculate $P^{ij}(r,\phi,\theta)$ using a local reference frame for each amino acid sidechain, as well as the virtual interaction center that represents the peptide backbone (Pep).

Once the local reference systems for special groups of atoms (e.g. the heavy atoms in sidechains, or the C, O and N of the peptide link) are defined, the statistics collected from a database of non-homologous proteins can be used to estimate the pair distributions for each specific type of site–site interaction. The $P^{ij}(r,\phi,\theta)$ distribution functions may be computed from the set of non-homologous proteins used by Scheraga *et al.* [20–22] for similar purposes. A larger training set of protein structures could be used if higher accuracy is necessary. The pair distributions can be further normalized by considering the corresponding volume elements and the total number of observations for building orientational probabilities $P^{ij}(r,\phi,\theta)$ for each type of interaction.

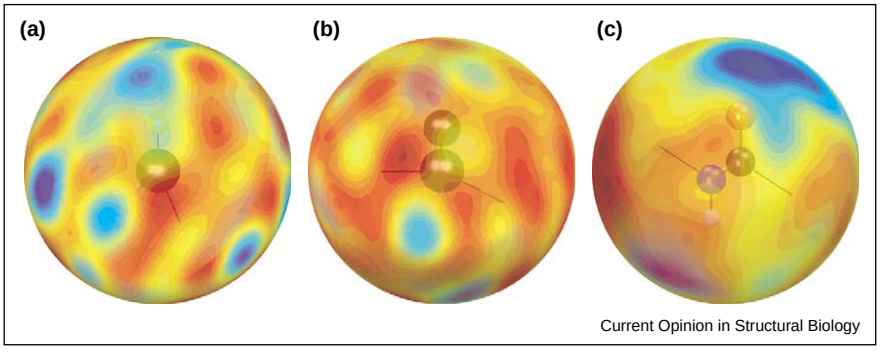
An important technical issue that appears when using probability density functions with the Boltzmann device is 'the problem of small datasets'. As noted by Sippl [9], dividing the SC–SC pair frequencies by both sidechain type and distance intervals can lead to situations in which the available data are too small in number for conventional statistical procedures. This problem was solved by Sippl, who proposed a 'sparse data correction' formula that builds the correct probability densities as linear combinations of the measured data and the reference [9,47,48].

We used the Boltzmann method to extract potentials for three distance ranges by considering short-range (2.0 → 5.6 Å), medium-range (5.6 → 9.2 Å) and long-range (9.2 → 12.8 Å) SC–SC interactions [16••]. For example, Figure 1 shows 3D representations of the extracted residue-based orientation-dependent potentials for Gly–Gly (Figure 1a, short-range interactions), Ala–Gly (Figure 1b, short-range interactions) and Pep–Pep (Figure 1c, medium-range interactions).

In this notation (e.g. Ala–Gly), the orientational potential is calculated in the local reference frame of the first amino acid (i.e. alanine) for its interactions with the second amino acid type (i.e. glycine).

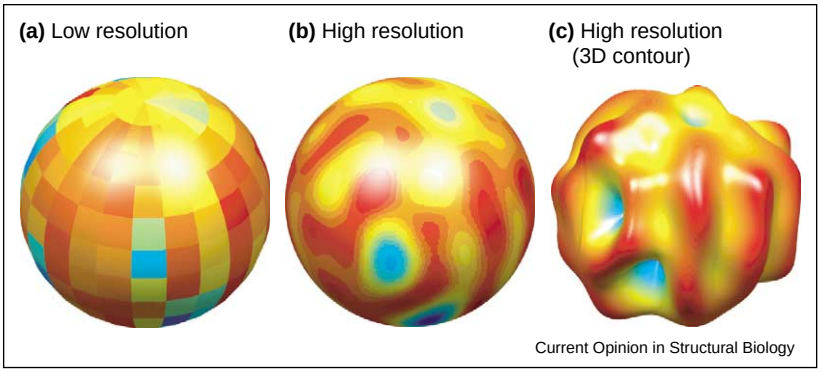
To efficiently use the orientational dependence of the inter-residue coarse-grained potentials, it is necessary to provide an easily computable functional representation. Assuming that the angular-dependent potential function is sufficiently smooth, one can decompose the potentials in terms of spherical harmonic analysis (SHA) [16••]. Alternatively, using the expansion coefficients, the potential function $U(\theta,\phi)$ can be reconstructed using

**Figure 1**




3D representations of the residue-based orientation-dependent potentials for **(a)** Gly–Gly short-range interactions, **(b)** Ala–Gly short-range interactions and **(c)** Pep–Pep medium-range interactions. The potential values, corresponding to a color scale ranging from blue (most attractive values) to red (repulsive), are projected on the surface of a spherical grid centered on the interaction center that corresponds to each type of sidechain ($S_i$).

**Figure 2**




3D representations of the SHS-reconstructed short-range residue-based orientation-dependent potentials for Ala–Gly interactions **(a)** on a $12 \times 24$ angular grid of the same size as the grid used for collecting the statistical data and **(b,c)** on a grid with a resolution that is ten times higher. In (c), the magnitude of the potential is proportional to both the distance from the interaction center of the sidechain and the color scale (i.e. red, repulsive, regions are located distant from the sidechain interaction center [$S_i$], whereas the blue, attractive, regions are closer to $S_i$).

the spherical harmonics synthesis (SHS) [16••]. This method provides a realistic representation of the orientation-dependent statistical potentials as smoothed, continuous functions.

In Figure 2, we show 3D representations of the SHS-reconstructed short-range residue-based orientation-dependent potentials for Ala–Gly interactions.

The new continuous orientation-dependent potentials lead to results that are consistent with and, in many cases, improved when compared to the raw potentials constructed directly from orientational interaction probabilities [16••]. Results from decoy tests show that the smoothing of the orientational potentials using the SHA/SHS approach does not necessarily lead to a loss of accuracy. In this context, we have shown that the Cα–SC–Pep representation improves protein fold recognition [16••].

## The choice of reference states

Regardless of the level of description used for polypeptide chains, the derivation of knowledge-based potentials from PDB structures hinges on the 'quasi-chemical' approximation (i.e. the interacting sidechains are in equilibrium with the solvent) [8,36]. Therefore, to minimize the errors introduced by entropy losses due to chain connectivity, without specifically correcting for it (as in [49]), one needs to choose a reference state as close to the random mixing approximation as possible [31]. In obtaining the anisotropic potentials, we have constructed the reference state by averaging over all interaction types. Recently, potentials of mean force constructed using a distance-scaled, finite ideal-gas reference state were shown to result in improved performance in native fold recognition at both atomistic [50•] and residue [51] levels. Alternatively, one can consider as a reference an amino acid such as threonine, with intermediate hydrophobic properties, closely correlated to experimental estimates [31].

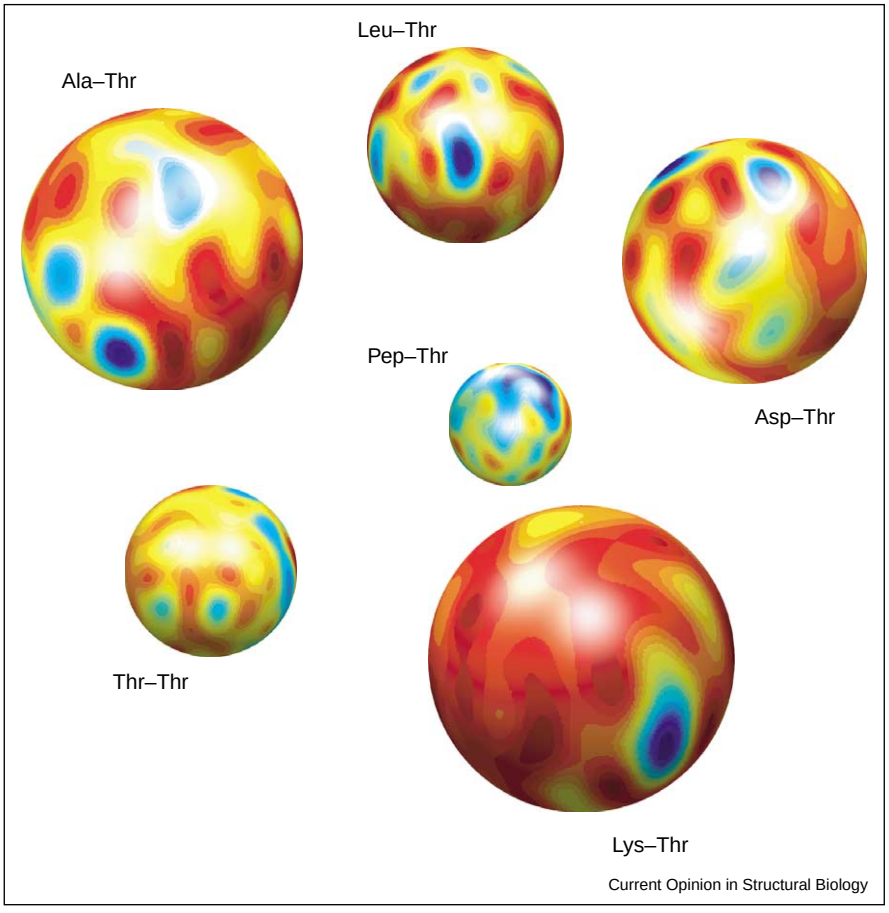
In Figure 3, we show a qualitative spatial representation of the orientation-dependent potentials that control the non-bonded 'quasi-chemical' interactions responsible for the specific sidechain packing in proteins.

For illustration purposes, interactions with respect to threonine are considered for all the interaction centers shown here. The degree of observed orientational interaction anisotropy is significant in most cases, even for interactions involving small amino acids such as alanine and glycine. The specific locations of statistically preferred interaction loci are observable. In particular, the preferred orientations for hydrogen bonding are clearly visible in the Pep interaction maps. Although the importance of physically motivated reference states is recognized, a rigorous theoretically based means of choosing one is not clear.

## Decoy sets: standard tests for native fold recognition

To assess the efficacy of various models and interaction schemes in protein fold recognition, standard tests are needed. Based on the observation that there are various ways in which one could improve even the simplest scoring functions, Samudrala and Levitt [52] have understood the necessity to organize a standard database of computer-generated, alternative protein conformations that are native-like, but are not the true native states. The 'decoy structures' are generated by various methods [53–55], with the specific aim of 'fooling scoring functions' [52,56–58,59•]. Our results from tests using the decoy sets of Samudrala and Levitt [52] confirmed that considering explicitly the Pep interaction centers on

**Figure 3**




Qualitative spatial representations of the orientation-dependent potentials that control the non-bonded, 'quasi-chemical' interactions responsible for specific sidechain packing in proteins. For illustration purposes, the interactions with respect to threonine are considered for all the interaction centers shown here. The relatively different sizes of the anisotropic residue-dependent potentials depicted in this figure are a qualitative reflection of the 3D spatial sidechain packing.

the backbone offers significant improvement over simpler models for the ability to recognize the native conformations of proteins [15••]. We have further used these decoy tests to assess the statistical improvement in the ability to recognize native protein folds resulting from the explicit inclusion in the residue-based statistical potentials of the orientation dependence. As newer scoring functions are being developed, new sets of decoys are also being generated and studied [56–58,59•]. There is a need to generate robust decoy sets for standard evaluation studies of the potentials.

## All-atom statistical potentials for proteins

All-atom models with various degrees of sophistication have been used in structure prediction studies for close homology modeling [60], native fold recognition [50•] and folding of small proteins [61]. Recently, Skolnick *et al.* [62,63] compared residue-based knowledge-based potentials with their heavy-atom-based statistical pair potential. They concluded that, although more time-consuming, the atom-based potential performs better in identifying near-native structures from docking-generated decoys. On the other hand, they suggest that the residue-based potential is well suited to genome-scale protein interaction prediction and analysis (e.g. in threading-based algorithms [3•]). Moreover, reduced models tend to perform better in more challenging *ab initio* structure predictions or structure refinement in distant homology modeling and threading calculations [3•,16••,17•,64,65]. Despite the success of anisotropic coarse-grained potentials in protein fold recognition, it is important to develop other physically motivated constructions of all-atom forcefields. For example, the profile method introduced by Wilmanns and Eisenberg [66] can be used in principle to compute all-atom potentials. Chang *et al.* [67] have already used this idea to obtain a simple class of mean field one-body potentials using a learning procedure. More recently, Kussell *et al.* [61] have constructed

all-atom pairwise potentials including a one-body term that accounts for sidechain exposure to solvent. Using this force-field, they successfully studied the folding kinetics of a three-helix bundle.

## Optimization of potential energy functions

The performance in fold recognition of orientation-dependent statistical potentials can be enhanced by a variety of methods, such as increasing and improving the quality of the training set of non-homologous structures or considering architecture-specific potentials. Recently, general methods have been developed for constructing new classes of effective contact-based [68] or distance-dependent [69] potentials by employing novel optimization techniques. For potentials expressed as linear combinations of pairwise additive functions, the optimal interaction parameters are calculated by employing linear programming. New algorithms, based on linear programming techniques such as the interior-point method, have been shown to be useful in both native state recognition and threading computations [64,70]. The results for potentials optimized by linear programming suggest that multibody interactions play a significant role in native state recognition. Large-scale potential optimization can be performed on a few hundred parameters and tens of millions of constraints [71], which makes this new class of algorithm generalizable to more complex potential functions that can explicitly include the relative orientation dependence of protein sidechains.

## Conclusions

Detailed force-fields are often needed for capturing the specific, essential features of native protein folds, protein kinetics and protein–protein interactions. It is reasonable to expect that an accurate residue-level description can be constructed to successfully meet the expectations of modern proteome- and genome-scale studies [3•,4•], and to simplify the rather complex current *ab initio* [72] protein design methods. Recent developments suggest that the specific orientation-dependent interactions responsible for sidechain packing and for backbone contacts should be included explicitly. This inclusion seems to partially alleviate the errors introduced by the assumption of pairwise interactions.

We have shown [15••] that the performance of energy-based scoring functions can be improved by using statistical information extracted from the relative residue–residue orientations. Our recent results [16••] suggest that the statistical data extracted from protein structural databases can be successfully used to build orientation-dependent potentials that have sufficient continuity properties to make possible their SHA. The resulting smooth, continuous interaction potentials are represented using separate spherical harmonic expansions of the orientation-dependent potential for short-, medium- and long-range interactions.

The choice of the reference state for statistical interactions is also important. The ideal-gas reference state can be generalized to sidechain interactions [50•,51], but further tests are needed to determine if this is a better choice than using a reference amino acid type (i.e. threonine [31]) or considering all the observed sidechain–sidechain interactions [15••]. New large-scale linear programming based optimization techniques [71] or machine learning procedures could be used to circumvent this type of question and to derive potential parameters directly optimized for native fold recognition.

The new continuous distance- and orientation-dependent statistical potentials could be useful in developing more efficient computational methods for protein structure prediction, as well as for Monte Carlo or molecular dynamics simulations of coarse-grained models of peptides and proteins. For structure prediction, the new residue-level statistical orientational potentials could be connected to the local backbone structure, using the information from a detailed rotamer library [73•] or a simplified SC–BB energy function [74].

## Acknowledgements

This work was supported by the National Institutes of Health (R01 NS41356-01, JES and DT). The data visualization was carried out using Matlab (The Mathworks Inc, Natick, MA). NVB is thankful to Gerhard Hummer for helpful discussions and support during the preparation of this review. We thank the editors, J Janin, R Page and T Simonson, for their useful suggestions and help with this review.

## 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. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**:235-242.

2. Service RF: **Proteomics - Public projects gear up to chart the protein landscape**. *Science* 2003, **302**:1316-1318.

3. • Lu L, Arakaki AK, Lu H, Skolnick J: **Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome**. *Genome Res* 2003, **13**:1146-1154.
A novel, multimeric threading algorithm (MULTIPROSPECTOR) for the prediction of protein–protein interactions is presented and applied on a large scale to search for possible interactions between more than 6000 encoded proteins.

4. • Janin J, Seraphin B: **Genome-wide studies of protein-protein interaction**. *Curr Opin Struct Biol* 2003, **13**:383-388.
A review of two large-scale experimental studies of protein–protein interactions. It is suggested that the majority of proteins exist in the cell as parts of multicomponent assemblies.

5. Thirumalai D, Klimov DK: **Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models**. *Curr Opin Struct Biol* 1999, **9**:197-207.

6. 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.

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

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

9. Sippl MJ: **Calculation of conformational ensembles from potentials of mean force**. *J Mol Biol* 1990, **213**:859-883.

10. 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.

11. Miyazawa S, Jernigan RL: **Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues**. *Proteins* 1999, **34**:49-68.

12. Meller J, Elber R: **Linear programming optimization and a double statistical filter for protein threading protocols**. *Proteins* 2001, **45**:241-261.

13. Kolinski A, Godzik A, Skolnick J: **A general method for the prediction of the three dimensional structure and folding pathways of globular proteins: application to designed helical proteins**. *J Chem Phys* 1993, **98**:7420-7433.

14. Gatchell DW, Dennis S, Vajda S: **Discrimination of near-native protein structures from misfolded models by empirical free energy functions**. *Proteins* 2000, **41**:518-534.

15. •• Buchete N-V, Straub JE, Thirumalai D: **Anisotropic coarse-grained statistical potentials improve the ability to identify native-like protein structures**. *J Chem Phys* 2003, **118**:7658-7671.
This is the first paper that uses sidechain-specific definitions of a local reference frame to derive and test distance- and orientation-dependent residue-based statistical potentials for proteins.

16. •• Buchete N-V, Straub JE, Thirumalai D: **Orientational potentials extracted from protein structures improve native fold recognition**. *Protein Sci* 2004, in press.
In this paper, the authors show that there are substantial contacts between the backbone and sidechains in the native states of proteins. Using this observation, they introduce the Cα–SC–Pep model, whereby the Pep virtual interaction center represents the peptide bond. The orientational potentials using this model greatly improve fold recognition.

17. • Kolinski A, Skolnick J: **Reduced models of proteins and their applications**. *Polymer* 2004, **45**:511-524.
Various reduced models developed for proteins are presented and classified.

18. •• Head-Gordon T, Brown S: **Minimalist models for protein folding and design**. *Curr Opin Struct Biol* 2003, **13**:160-167.
A review of the development of minimalist protein models in the context of their application to current research issues in protein folding and design.

19. Lee J, Liwo A, Scheraga HA: **Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein A and to apo calbindin D9K**. *Proc Natl Acad Sci USA* 1999, **96**:2025-2030.

20. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA: **A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data**. *J Comput Chem* 1997, **18**:849-873.

21. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA: **A united-residue force field for off-lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization**. *J Comput Chem* 1997, **18**:874-887.

22. Liwo A, Kazmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Wawak RJ, Rackovsky S, Pincus MR, Scheraga HA: **United-residue force field for off-lattice protein-structure simulations: III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials**. *J Comput Chem* 1998, **19**:259-276.

23. Moult J, Fidelis K, Zemla A, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP)-round V**. *Proteins* 2003, **53**:334-339.

24. Srinivas G, Bagchi B: **Study of the dynamics of protein folding through minimalistic models**. *Theor Chem Acc* 2003, **109**:8-21.

25. •• Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vasker I, Wodak SJ: **CAPRI: A Critical Assessment of PRedicted Interactions**. *Proteins* 2003, **52**:2-9.
An analysis of the results of CAPRI, a community-wide experiment to assess the ability of current protein docking methods to predict protein–protein interactions. Recent results ''stress the need for new scoring functions and for methods handling the conformation changes that were observed in some of the target systems''.

26. •• Murphy J, Gatchell DW, Prasad JC, Vajda S: **Combination of scoring functions improves discrimination in protein-protein docking**. *Proteins* 2003, **53**:840-854.
This study suggests that the discrimination strategies that perform best in protein–protein docking combine an RPScore (residue pair potential score) filter with an ACP (atomic contact potential)-based scoring function.

27. •• Hummer G, Kevrekidis IG: **Coarse molecular dynamics of a peptide fragment: free energy, kinetics, and long-time dynamics computations**. *J Chem Phys* 2003, **118**:10762-10773.
A new molecular dynamics algorithm is presented. It is shown that the evolution of conformationally coarse variables for a small peptide can be inferred from an ensemble of short, appropriately initialized all-atom simulations.

28. •• Fischer D, Rychlewski L, Dunbrack RL, Ortiz AR, Elofsson A: **CAFASP3: The third critical assessment of fully automated structure prediction methods**. *Proteins* 2003, **53**:503-516.
This report shows the significant progress achieved in automatic structure prediction and its important implications for the prospects of automated structure modeling in the context of structural genomics.

29. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Grana O, Pazos F, Valencia A, Sali A, Rost B: **EVA: evaluation of protein structure prediction servers**. *Nucleic Acids Res* 2003, **31**:3311-3315.

30. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A: **Tools for comparative protein structure modeling and analysis**. *Nucleic Acids Res* 2003, **31**:3375-3380.

31. Betancourt MR, Thirumalai D: **Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes**. *Protein Sci* 1999, **8**:361-369.

32. Vendruscolo M, Domany E: **Pairwise contact potentials are unsuitable for protein folding**. *J Chem Phys* 1998, **109**:11101-11108.

33. Meller J, Wagner M, Elber R: **Maximum feasibility guideline in the design and analysis of protein folding potentials**. *J Comput Chem* 2002, **23**:111-118.

34. Honeycutt JD, Thirumalai D: **Metastability of the folded states of globular proteins**. *Proc Natl Acad Sci USA* 1990, **87**:3526-3529.

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

36. Miyazawa S, Jernigan RL: **Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space**. *Proteins* 2003, **50**:35-43.

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

38. Dima RI, Thirumalai D: **Asymmetry in the shapes of folded and denaturated states of proteins**. *J Phys Chem B* 2004, in press.

39. Pearl FMG, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA: **Assigning genomic sequences to CATH**. *Nucleic Acids Res* 2000, **28**:277-282.

40. Bahar I, Kaplan M, Jernigan RL: **Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches**. *Proteins* 1997, **29**:292-308.

41. Miyazawa S, Jernigan RL: **Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition**. *Proteins* 1999, **36**:347-356.

42. Tsai J, Taylor R, Chothia C, Gerstein M: **The packing density in proteins: standard radii and volumes**. *J Mol Biol* 1999, **290**:253-266.

43. Bagci Z, Jernigan RL, Bahar I: **Residue packing in proteins: uniform distribution on a coarse-grained scale**. *J Chem Phys* 2002, **116**:2269-2276.

44. Bagci Z, Jernigan RL, Bahar I: **Residue coordination in proteins conforms to the closest packing of spheres**. *Polymer* 2002, **43**:451-459.

45. Banavar JR, Maritan A, Seno F: **Anisotropic effective interactions in a coarse-grained tube picture of proteins**. *Proteins* 2002, **49**:246-254.

46. Bahar I, Jernigan RL: **Coordination geometry of nonbonded residues in globular proteins**. *Folding Des* 1996, **1**:357-370.

47. 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.

48. Thomas PD, Dill KA: **Statistical potentials extracted from protein structures: how accurate are they?** *J Mol Biol* 1996, **257**:457-469.

49. Skolnick J, Jaroszewski L, Kolinski A, Godzik A: **Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?** *Protein Sci* 1997, **6**:1-13.

50. • Zhou HY, Zhou YQ: **Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction**. *Protein Sci* 2002, **11**:2714-2726. [Correction appears in *Protein Sci* 2003, **12**:2121-2121].

The authors suggest that the use of the ideal-gas reference state greatly improves fold recognition.

51. Zhang C, Liu S, Zhou H, Zhou YQ: **An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled ideal-gas reference state**. *Protein Sci* 2004, **13**:400-411.

52. Samudrala R, Levitt M: **Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction**. *Protein Sci* 2000, **9**:1399-1401.

53. Park B, Levitt M: **Energy functions that discriminate X-ray and near native folds from well-constructed decoys**. *J Mol Biol* 1996, **258**:367-392.

54. Park B, Huang ES, Levitt M: **Factors affecting the ability of energy functions to discriminate correct from incorrect folds**. *J Mol Biol* 1997, **266**:831-846.

55. Simons KT, Kooperberg C, Huang ES, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions**. *J Mol Biol* 1997, **268**:209-225.

56. Keasar C, Levitt M: **A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics**. *J Mol Biol* 2003, **329**:159-174.

57. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D: **An improved protein decoy set for testing energy functions for protein structure prediction**. *Proteins* 2003, **53**:76-87.

58. Chen W, Mirny L, Shakhnovich EI: **Fold recognition with minimal gaps**. *Proteins* 2003, **51**:531-543.

59. • Zhu J, Zhu QQ, Shi YY, Liu HY: **How well can we predict native contacts in proteins based on decoy structures and their energies?** *Proteins* 2003, **52**:598-608.

Decoys based on Gromos96 were generated and used to predict native contacts.

60. Sali A, Potterton L, Yuan F, Vanvlijmen H, Karplus M: **Evaluation of comparative protein modeling by Modeller**. *Proteins* 1995, **23**:318-326.

61. Kussell E, Shimada J, Shakhnovich EI: **A structure-based method for derivation of all-atom potentials for protein folding**. *Proc Natl Acad Sci USA* 2002, **99**:5343-5348.

62. Lu H, Lu L, Skolnick J: **Development of unified statistical potentials describing protein-protein interactions**. *Biophys J* 2003, **84**:1895-1901.

63. Lu H, Skolnick J: **A distance-dependent atomic knowledge-based potential for improved protein structure selection**. *Proteins* 2001, **44**:223-232.

64. Meller J, Elber R: **Protein recognition by sequence-to-structure fitness: bridging efficiency and capacity of threading models**. *Adv Chem Phys* 2002, **120**:77-130.

65. Betancourt MR: **A reduced protein model with accurate native-structure identification ability**. *Proteins* 2003, **53**:889-907.

66. Wilmanns M, Eisenberg D: **Inverse protein-folding by the residue pair preference profile method - Estimating the correctness of alignments of structurally compatible sequences**. *Protein Eng* 1995, **8**:627-639.

67. Chang I, Cieplak M, Dima RI, Maritan A, Banavar JR: **Protein threading by learning**. *Proc Natl Acad Sci USA* 2001, **98**:14350-14355.

68. Tobi D, Shafran G, Linial N, Elber R: **On the design and analysis of protein folding potentials**. *Proteins* 2000, **40**:71-85.

69. Tobi D, Elber R: **Distance-dependent, pair potential for protein folding: results from linear optimization**. *Proteins* 2000, **41**:40-46.

70. Meller J, Elber R: **Linear optimization and a double statistical filter for protein threading protocols**. *Proteins* 2001, **45**:241-261.

71. Wagner M, Meller J, Elber R: **Large-scale linear programming techniques for the design of protein folding potentials**. *Mathematical Programming* 2004, in press.

72. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy**. *Science* 2003, **302**:1364-1368.

73. • Canutescu AA, Shelenkov AA, Dunbrack RL Jr: **A graph-theory algorithm for rapid protein side-chain prediction**. *Protein Sci* 2003, **12**:2001-2014.

The newest (third) version of the SCWRL program for sidechain prediction using a backbone-dependent rotamer library is presented.

74. Kazmierkiewicz R, Liwo A, Scheraga HA: **Addition of side chains to a known backbone with defined side-chain centroids**. *Biophys Chem* 2003, **100**:261-280.