Ab initio protein structure prediction

Corey Hardin*, Taras V Pogorelov† and Zaida Luthey-Schulten*†

Steady progress has been made in the field of *ab initio* protein folding. A variety of methods now allow the prediction of low-resolution structures of small proteins or protein fragments up to approximately 100 amino acid residues in length. Such low-resolution structures may be sufficient for the functional annotation of protein sequences on a genome-wide scale. Although no consistently reliable algorithm is currently available, the essential challenges to developing a general theory or approach to protein structure prediction are better understood. The energy landscapes resulting from the structure prediction algorithms are only partially funneled to the native state of the protein. This review focuses on two areas of recent advances in *ab initio* structure prediction – improvements in the energy functions and strategies to search the caldera region of the energy landscapes.

Addresses
*Center for Biophysics and Computational Biology, University of Illinois, 600 South Mathews Avenue, Urbana, Illinois 61801, USA
†School of Chemical Sciences, University of Illinois, 600 South Mathews Avenue, Urbana, Illinois 61801, USA

Current Opinion in Structural Biology 2002, 12:176–181

0959-440X/02/$ – see front matter

Abbreviations
CASP Critical Assessment of Structure Prediction
PDB Protein Data Bank
rmsd root mean square deviation
UNRES united residue

Introduction

The prediction of a protein's structure and folding mechanism from knowledge only of its amino acid sequence has been described as the determination of the second half of the genetic code [1]. Because of its importance for both practical and theoretical purposes, this challenge has been steadily pursued for over a decade, a pursuit that has yielded a wide variety of novel computational techniques and much progress. The approaches used to predict protein structure range from comparative modeling using a homologous protein that already exists in the structural databases to *ab initio* folding, which results in a novel fold. In between these two extremes is the technique of threading, a method of fold recognition whereby one attempts to construct a model of the protein using as a template the structure of a protein in the PDB that has little or no obvious sequence relation to the target protein. These definitions are somewhat vague and, in the case of very low sequence identity, the distinction between threading and *ab initio* folding becomes blurred as virtually all successful *ab initio* methods utilize information from the sequence and structural databases in some form. Indeed, the development of techniques that make use of regions of local similarity between globally dissimilar proteins is one of the areas that has seen dramatic recent progress [2••,3,4•].

The three approaches mentioned above also define the categories of the Critical Assessment of Structure Prediction (CASP) experiments, which take place every two years. These community-wide blind tests of prediction methods are useful to gauge the progress of the field. To date, the most successful method for structure prediction is homology-based comparative modeling. Advances in homology modeling and threading have been recently reviewed [5]; we will concentrate here on the recent progress in *ab initio* protein folding methods [6].

Often, the term *ab initio* is interpreted to mean to start with potentials that are based entirely on physiochemical interactions, such as the empirical potentials used in CHARMM and AMBER. Although there has been progress in the use of full-atom simulations with explicit and implicit solvent models to predict the folding of small peptides and to discriminate between the native state and static decoy sets (van Gunsteren and co-workers, this issue, pp 190–196; [7•]), a more practical use has been as an adjunct to reduced-model *ab initio* protocols [8••]. As the most successful prediction methods all use structural information to some degree, in this review we will strictly use the term *ab initio* to mean to start without knowledge of globally similar folds and to produce a structure that has a novel fold. With this more general definition, we include both statistics- and physics-based energy functions, which derive the parameters appearing in their potentials from the structural databases. The results of CASP4, completed this past summer, provide a snapshot of the current state of the field. Compared to previous years, longer fragments of proteins were predicted within 6 Å of the crystal structure. Figure 1, a Hubbard plot [9,10] for one of the eight proteins in the novel fold category, indicates that the methods of several groups were able to predict long contiguous segments of the protein [2••,3,4•,11,12•,13], with the best results being obtained by the Rosetta statistics-based approach of the Baker group [2••]. The striking feature of Figure 1 is that many different methods were able to obtain comparable results, but none were able to predict the structure of the entire protein. One goal of this review is to examine what common features of the methods were important in achieving this improvement and to interpret these features in terms of the effects they have on the energy landscape of the prediction algorithm. Energy landscape theory describes how the energy of the system changes with the geometry of the protein; its usefulness as a framework to analyze protein folding has been reviewed in many places [14,15].

All protein structure prediction techniques comprise a representation of the protein, a force field commensurate with

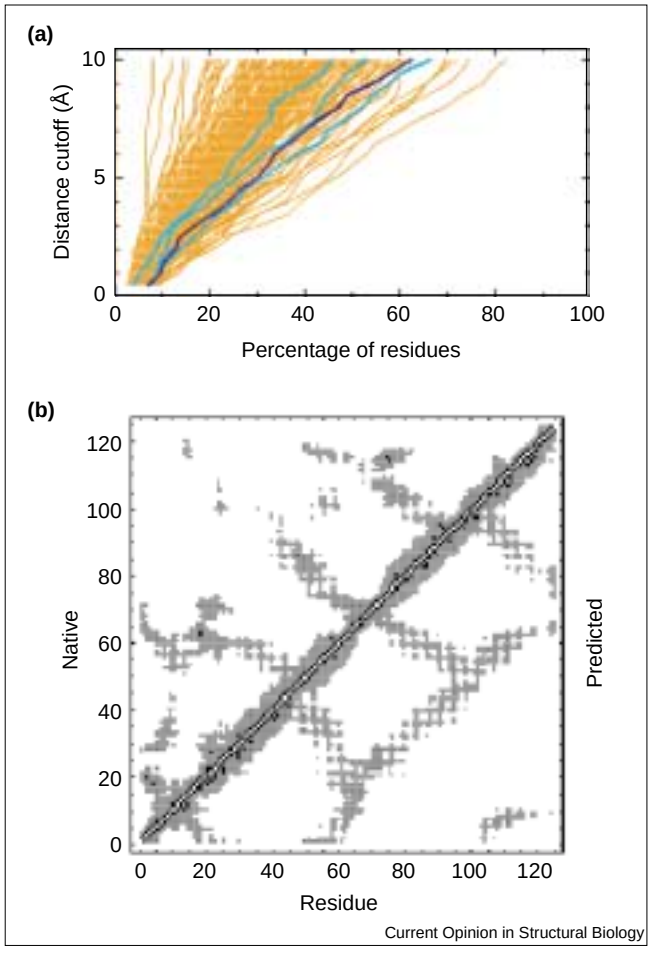
this representation, a technique for searching the resultant energy landscape [16] and a method for evaluating the prediction scheme [17••]. Most of the *ab initio* protein structure prediction methods discussed here use reduced representations of the protein, at least in the initial stages. Interactions are typically assigned between sites located at the Cα atom, the Cβ atom, the peptide bond or at the center of mass of the sidechains. For each representation, a corresponding set of interaction potentials are developed and used to guide the sampling of conformation space. In the construction of the interaction potential, statistical information from the sequence and structural databases is used to optimize the weights assigned to the various interactions. The topology of the optimized landscape is critical to the success of the search procedure to find the best predicted structure [17••]. Even with the recent improvements in the *ab initio* energy functions, the correlation between energy and measures of similarity to the native state, such as rmsd, weakens as the native state is approached [4•,8••,11,18••,19]; rather than being funnel-like, the landscape resembles a caldera. Searching such a landscape results in a large ensemble of structures with similar energies and widely varying similarities to the native state. To overcome the flattening of the landscape, several groups have incorporated rather extensive filtering and clustering techniques for the final stages of the prediction process. This review focuses on two areas: improvements in the energy functions and strategies for searching the caldera region of the landscape.

Improvements in the energy functions

The essential requirement for protein folding, no less for model proteins than for real ones, is the ability to efficiently search a rugged energy landscape for the (presumably) minimum energy native state. Bryngelson and Wolynes [20] formulated this discrimination requirement as the 'principle of minimal frustration'. Parameterization of the prediction energy functions must maximize the ratio of the gap in energy between the folded state (E_n) and the ensemble of unfolded (or non-native) states ($<E_u>$) to the energetic variation of the unfolded states (ΔE) generated during the search process: $(E_n - <E_u>)/\Delta E$. Many schemes for structure prediction directly optimize this dimensionless measure of foldability [21,22•] or a related quantity [23]. In these cases, the optimization procedure requires the generation of large decoy sets of non-native structures containing a controlled amount of secondary structure. The functional forms of the energy functions depend on the features selected to distinguish folded native conformations from the decoy conformations. Although the forms may vary, the features usually occur in the scoring functions with linear weights, so that optimization is a straightforward task. Dill and co-workers [24] have recently suggested a global optimization method to adjust the energy parameters for any given search strategy that gives promising results for simple models with few adjustable parameters.

Energy functions are conveniently categorized according to the degree to which they make use of data from experimentally determined structures. At one end of the spectrum are models that explicitly score trial structures according to their similarity to a database of known structures. It is important to distinguish these approaches from homology modeling, as databases for *ab initio* prediction are not required to contain any proteins with global structural similarity to the target. At the other end of the spectrum are more physics-based approaches, which use the sequence and structural databases merely to derive the parameters occurring in their energy functions.

Figure 1



Ab initio structure prediction results. (a) Hubbard plot of *ab initio* predictions for CASP4 entry t0106, a 128 amino acid secreted frizzled protein from mouse [53]. Each group could submit up to five ranked models, four of which are shown in blue to represent data from our group, with dark blue corresponding to our highest ranked model. (b) The contact distance map of the native structure and our best predicted model in (a) indicates that most of the secondary units are predicted correctly, but that their packing is not always correct, which is shown by the presence of additional off-diagonal contacts in the lower right-hand half of the map.

Statistics-based potentials

Early on, Go [25] observed that efficient folding requires consistency between a protein's tertiary structure and the local conformational preferences of its sequence.

Simons *et al.* [26] expanded on this observation by suggesting that the collection of conformations in which a given sequence fragment can be found in the database of known structures approximates the ensemble of its local structural preferences. This forms the basis of the successful Rosetta program [26]. A predicted structure is built up from matching segments in a fragment library via a Monte Carlo procedure whose scoring function is the Bayesian probability of structure/sequence matches. Recent improvements to this program included the use of secondary structure prediction to bias the selection of fragments and terms to favor the assembly of strands into sheets and the burial of hydrophobic residues [23,27]. In addition to modifications to the energy function, the improved performance of this method is attributed to a variety of post-Monte Carlo filters of the *ab initio* structures, which are described below.

The work by Wolynes, Luthey-Schulten and co-workers [28,29] addresses the protein folding problem as one of information processing. The short-range interactions between residues are treated as associative memory-like potentials that learn sequence/structure associations from a database of known structures. Residue pairs in the target and memory proteins are associated by a sequence/structure threading algorithm [30]. The theory of such energy functions is quite advanced [31] and suggests that it should be possible to construct an energy function that is funneled to the correct native state even from a database that contains only short, local regions of structural similarity. As there are no globally similar folds in the memory proteins, the interactions between residues distant in sequence are now determined by a series of piecewise contact potentials whose forms are chosen to roughly mimic the observed behavior of pair correlations between distant pairs in known structures [4•,22•]. As with many other methods, it was found that the performance of the simulation is improved when the parameters are separately optimized for proteins belonging to different structural classes. Correct β-sheet formation is promoted by the inclusion of an explicit hydrogen bond potential [4•].

Skolnick, Kolinski and co-workers [3,32•,33] have developed a hierarchical approach to *ab initio* folding on a high-coordination lattice that uses a combination of multiple sequence comparisons, threading, clustering and refinement. The profiles obtained from the multiple sequence alignments are used to construct pair distance restraints and secondary structure biasing in the scoring function of their threading algorithm [33]. In the prediction of novel folds, the threading algorithm provides fragmentary templates for starting lattice models. As threading is limited by the inaccuracy of the scoring energy functions, averaging over a set of homologous sequences can improve the consistency and discrimination scores of these methods [34,35]. Although groups have differed on the details of its implementation [2••,17••], there is widespread agreement that the use of information from multiple sequence alignments invariably improves performance. The force fields used in the lattice simulations also consist of statistical potentials for pairwise and multibody sidechain interactions. The conformational space is sampled by replica exchange Monte Carlo, a technique that may be helpful in overcoming the slow dynamics associated with rough energy landscapes. The strength of this hierarchical prediction approach is that it can be used with little modification for either fold recognition or *ab initio* folding.

A number of other methods make use of predicted secondary structure (predictions that themselves are derived from propensities observed in the databases) to reduce the prediction task to that of the assembly of preformed elements [11,12•,36]. The potentials used in the assembly phase have been parameterized on known structures. Friesner and his co-workers [11] point out that there are good reasons for separating the prediction problem into separate parts for secondary and tertiary structure. To this we would add the comment that, as schemes for secondary structure prediction are now quite reliable, it is quite natural to focus on the assembly problem as the most urgent. Eyrich *et al.* [11] combine knowledge of predicted secondary structure with a contact potential among sidechain centroids and an excluded volume term fit to observed pair distances in a set of known structures. They have noted that it was important to modify the potential function according to the size of the protein. Levitt and colleagues [12•] also used predicted secondary structure, but with the addition of two cooperative terms to their energy function, one for hydrogen bonds and one to confine hydrophilic residues to the protein surface. Yue and Dill [37] have also developed a technique that seeks to assemble secondary structure elements as rigid bodies and have recently augmented the secondary structure prediction with small homologous fragments taken from a database that is similar to the ISITES library, as implemented in Rosetta.

Physics-based potentials

Scheraga and co-workers [38–40] have developed a physics-based reduced model in which the interaction terms for a united residue (UNRES) description are derived by averaging over the neglected degrees of freedom in the all-atom ECEPP/3 force field. The weights of the different terms appearing in UNRES are determined by maximizing the Z-score, a quantity similar to the foldability definition above. Over the past two years, the authors continued to improve their force fields with the aid of a cumulant expansion to introduce more multibody terms important for describing β-sheet formation [13,41•]. They obtain low-resolution structures by using force fields optimized for the various structural classes and, once the structures are clustered, each model is then converted to an all-atom structure and refined by the ECEPP/3 force field.

The method of Crivelli *et al.* [42•] is notable for its extensive use of a modified all-atom AMBER5 force field, rather than statistical potentials. Constraints derived

from secondary structure prediction are used to smooth the energy landscape and reduce the phase space searched by their global optimization method. To account for solvation effects in the optimization, the authors add a hydrophobic contact potential to the AMBER molecular mechanics potentials. Although the procedure is computationally intensive, it performs well on small α-helical proteins [10].

Searching the caldera

None of the methods detailed above yield an energy surface that is funneled all the way to very native-like structures. The most successful groups have implemented some type of additional filtering or clustering procedure to pick out native-like conformations from the energetically degenerate ensemble of *ab initio* predictions. Searches in the flat caldera region of the energy landscape are necessarily undertaken with energy functions that are sensitive to features either missing or poorly represented in the *ab initio* functions.

Most of the scoring functions in this final stage of the *ab initio* prediction require the positions of all the backbone and sidechain atoms. In models using reduced atomic details, the missing atoms are typically generated using either a homology modeling program such as Modeller [43] or MaxSprout [44], or other hybrid knowledge- and physics-based sidechain builders [45•]. Given the importance of sidechain atoms in determining interactions for full-atom potentials and surface accessibility, which is often used in statistics-based potentials, the refinement will influence the degree of discrimination that can be achieved by the scoring functions. The most intriguing of these scoring functions, because it holds out the promise of yielding very high-resolution predictions (2.0–3.0 Å), is the use of full-atom force fields with and without implicit solvent models [7•,8••,46]. Most groups use some sort of clustering [2••,11,18••] to reduce the ensemble of predicted structures to a few structural families. Before clustering, the results of the *ab initio* simulations can be filtered for β-sheet formation, incorrect physical interactions and contact order, or they can be ranked by a threading potential. Recently, Petrey and Honig [47] have suggested using a simplified energy function that combines a Coulomb term with a hydrophobic contact term to differentiate the conformations. A measure of how many different structures are to be expected in this caldera region for any prediction scheme can be estimated from a free energy analysis of the prediction energy function using conventional multiple histogram sampling methods [17••]. Such an analysis of the energy landscape will also reveal the best structures to be expected from the given energy function and, along with the foldability criterion, provide a quantitative method to guide improvements to the energy function.

Conclusions

The resolution of current *ab initio* structure prediction techniques, although much improved during the past two years, is clearly not yet good enough for detailed studies such as docking and drug design. But even at the present stage, the low-resolution structures that they generate may be sufficient for genome annotation. Several authors have begun to explore this application. Baker and co-workers [48,49•] have demonstrated several cases in which a structure-based comparison of their models to known structures in the PDB found related proteins, yielding information on function by analogy. Skolnick and co-workers [50] have taken a more direct approach, designing active site signatures, termed fuzzy-functional forms, which can be used to assign function directly to a predicted structure. By including knowledge from structural and comparative genomics in the analysis of local sequence/structure patterns [27,51,52], the statistics-based energy functions in the *ab initio* approaches will continue to improve.

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. Gierasch LA, King J: Preface. In *Protein Folding: Deciphering the Second Half of the Genetic Code*. Edited by Gierasch LA, King J. Washington: AAAS; 1990:vii-viii.

2 •• Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D: Rosetta in CASP4: progress in ***ab initio*** protein structure prediction. *Proteins* 2001, 5(suppl):119-126.
This paper contains a self-assessment of the performance of the Rosetta *ab initio* prediction algorithm in CASP4. The authors attribute the improved performance to a faster program, which enabled more complete conformational searches, a series of 'filters' for selecting from the large number of structures constructed by Rosetta and the use of evolutionary information in the form of multiple sequence alignments.

3. Skolnick J, Kolinski A, Kihara D, Betancourt M, Rotkiewicz P, Boniecki M: ***Ab initio*** protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins* 2001, 5(suppl):149-156.

4. • Hardin C, Eastwood MP, Prentiss M, Luthey-Schulten Z, Wolynes PG: Folding funnels: the key to robust protein structure prediction. *J Comput Chem* 2002, 23:138-146.
The authors present a detailed discussion of the energy functions used in their associative memory contact potentials for *ab initio* protein structure prediction.

5. Moult J, Fidelis K, Zemla A, Hubbard T: Critical assessment of methods of protein structure prediction CASP — round IV. *Proteins* 2001, 5(suppl):2-7.

6. Lesk AM, Hubbard TJP: Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins* 2001, 5(suppl):98-118.

7. • Lazaridis T, Karplus M: Effective energy functions for protein structure prediction. *Curr Opin Struct Biol* 2000, 10:139-145.
A review of statistical and physical effective energy functions, including implicit solvation models.

8. •• Lee MR, Tsai J, Baker D, Kollman PA: Molecular dynamics in the endgame of protein structure prediction. *J Mol Biol* 2001, 313:417-430.
An excellent review of the progress and limitations of the use of all-atom molecular dynamics calculations to refine the results of *ab initio* Rosetta predictions.

9. Hubbard TJP: RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions. *Proteins* 1999, 3(suppl):15-21.

10. Fourth community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP4) on the World Wide Web URL: http://predictioncenter.llnl.gov/casp4/

11. Eyrich VA, Standley DM, Flets AK, Friesner RA: Protein tertiary structure prediction using a branch and bound algorithm. *Proteins* 1999, 35:41-57.

12. • Xia Y, Huang ES, Levitt M, Samudrala R: ***Ab initio*** construction of protein tertiary structure using a hierarchical approach. *J Mol Biol* 2000, 300:171-185.
The authors describe the combination of a lattice model with a statistical energy function and a knowledge-based atomic level potential that is used for protein tertiary structure prediction.

13. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova EA *et al.*: Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc Natl Acad Sci USA* 2001, 98:2329-2333.

14. Onuchic JN, Luthey-Schulten Z, Wolynes PG: Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 1997, 48:545-600.

15. Brooks CL III, Onuchic JN, Wales DJ: Taking a walk on a landscape. *Science* 2001, 293:612-613.

16. Osguthorpe DJ: ***Ab initio*** protein folding. *Curr Opin Struct Biol* 2000, 10:146-152.

17. •• Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG: Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J Res Dev* 2001, 45:475-497.
The authors use multiple histogram sampling methods to measure the quality of their energy function for *ab initio* protein structure prediction.

18. •• Betancourt MR, Skolnick J: Finding the needle in a haystack: deducing native folds from ambiguous ***ab initio*** protein structure predictions. *J Comput Chem* 2001, 22:339-353.
The problem of finding the most native-like structure from an energetically degenerate ensemble of low-resolution predicted structures is addressed. The authors' method involves clustering the ensemble of predicted structures, then developing an average structure for each cluster by minimizing a harmonic potential constructed from average distances of the cluster structures.

19. Bonneau R, Strauss CEM, Baker D: Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* 2001, 43:1-11.

20. Bryngelson JD, Wolynes PG: Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J Phys Chem* 1989, 93:6902-6915.

21. Hao M, Scheraga HA: Designing potential energy functions for protein folding. *Curr Opin Struct Biol* 1999, 9:184-188.

22. • Hardin C, Eastwood MP, Luthey-Schulten Z, Wolynes PG: Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. *Proc Natl Acad Sci USA* 2000, 97:14235-14240.
The authors present modifications to the original associative memory potentials for *ab initio* folding.

23. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D: Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 1999, 34:82-95.

24. Rosen JB, Phillips AT, Oh SY, Dill KA: A method for parameter optimization in computational biology. *Biophys J* 2000, 79:2818-2824.

25. Go N: Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 1983, 12:183-210.

26. Simons KT, Kooperberg C, Huang E, 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.

27. Bystroff C, Thorsson V, Baker D: HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* 2000, 301:173-190.

28. Goldstein RA, Luthey-Schulten Z, Wolynes PG: Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA* 1992, 89:4918-4922.

29. Koretke KK, Luthey-Schulten Z, Wolynes PG: Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc Natl Acad Sci USA* 1998, 95:2932-2937.

30. Koretke KK, Luthey-Schulten Z, Wolynes PG: Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci* 1996, 5:1043-1059.

31. Sasai M, Wolynes PG: Unified theory of collapse, folding, and glass transitions in associative-memory Hamiltonian models of proteins. *Phys Rev A* 1992, 46:7979-7997.

32. • Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J: Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 2001, 44:133-149.
GENECOMP is a prediction procedure that combines multiple approaches and is mechanistically the same regardless of the similarity of the target sequence to sequences of known structure.

33. Skolnick J, Kihara D: Defrosting the frozen approximation: PROSPECTOR — a new approach to threading. *Proteins* 2001, 42:319-331.

34 Haney P, Konisky J, Koretke KK, Luthey-Schulten Z, Wolynes PG: Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus ***Methanococcus***. *Proteins* 1997, 28:117-130.

35. Reva BA, Skolnick J, Finkelstein AV: Averaging interaction energies over homologs improves protein fold recognition in gapless threading. *Proteins* 1999, 35:353-359.

36. Eyrich VA, Standley DM, Friesner RA: Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol* 1999, 288:725-742.

37. Yue K, Dill K: Constraint-based assembly of tertiary protein structures from secondary structure elements. *Protein Sci* 2000, 9:1935-1946.

38. 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.

39. 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.

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

41. • Pillardy J, Czaplewski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll DR, Arlukowicz P, Oldziej S, Arnautova EA, Scheraga HA: Development of physics-based energy functions that predict medium-resolution structures for proteins of **α**, **β**, and **α**/**β** structural classes. *J Phys Chem B* 2001, 105:7299-7311.
The authors discuss changes in the UNRES force field needed to fold α/β proteins and analyze which terms are most sensitive to nonlocal interactions in β-sheet structures.

42. • Crivelli S, Byrd R, Eskow E, Schnabe R, Yu R, Philip TM, Head-Gordon T: A global optimization strategy for predicting **α**-helical protein tertiary structure. *Comput Chem* 2000, 24:489-497.
The approach described in this paper is one of the few attempts to use a full-atom force field in conjunction with a modeled solvation potential and predicted secondary structure to fold proteins.

43. MODELLER, a program for protein structure modeling, release 6 on World Wide Web URL: http://guitar.rockefeller.edu

44. Holm L, Sander C: Database algorithm for generating protein backbone and side-chain coordinates from a C_{α} trace: application to model building and detection of coordinate errors. *J Mol Biol* 1991, 218:183-194.

45. • Samudrala R, Huang ES, Koehl P, Levitt M: Constructing side chains on near-native main chains for ***ab initio*** protein structure prediction. *Protein Eng* 2000, 13:453-457.
The authors show that naive statistics-based rotamer selection for sidechain placement performs as well as more sophisticated methods.

46. Dominy BN, Brooks CL III: Identifying native-like protein structures using physics-based potentials. *J Comput Chem* 2002, 23:147-160.

47. Petrey D, Honig B: Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci* 2000, 9:2181-2191.

48. Bonneau R, Tsai J, Ruczinski I, Baker D: Functional inferences from blind ***ab initio*** protein structure predictions. *J Struct Biol* 2001, 134:186-190.

49. Simons KT, Strauss C, Baker D: Prospects for ***ab initio*** protein
• structural genomics. *J Mol Biol* 2001, 306:1191-1199.
The authors explore the feasibility of the large-scale prediction of function from sequence using low-resolution predicted structures.

50. Fetrow JS, Godzik A, Skolnick J: Functional analysis of the ***Escherichia coli*** genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J Mol Biol* 1998, 282:703-711.

51. O'Donoghue P, Amaro RE, Luthey-Schulten Z: On the structure of hisH: protein structure prediction in the contest of structural and functional genomics. *J Struct Biol* 2001, 134:257-268.

52. Yang AS, Honig B: An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignment. *J Mol Biol* 2000, 301:691-711.

53. Dann CE, Hsieh JC, Rattner A, Sharma D, Nathans J, Leahy DJ: Insights into Wnt binding and signaling from the structures of two frizzled cysteine-rich domains. *Nature* 2001, 412:86-90.