Knowledge-based potentials for proteins

Manfred J Sippl

University of Salzburg, Salzburg, Austria

Knowledge based potentials and energy functions are extracted from a number of databases of known protein structures. Recent developments have shown that this type of potential is successful in many areas of protein structure research. Among these are quality assessment and error recognition of folds and the prediction of unknown structures by fold-recognition techniques.

Current Opinion in Structural Biology 1995, 5:229–235

Introduction

The development of energy functions and force fields for studying the behaviour of molecular systems is a major goal in physical chemistry. Prediction of native structures of proteins from amino acid sequences, simulation of the folding process, and calculation of protein stabilities are among the most ambitious goals of contemporary research in biomolecular theory [1].

Research on these topics already has a respectable history, and the difficulties encountered over the past two decades seemed to indicate that they might be intractable because of our lack of a suitable theory of molecular interactions, and because of the computational complexities involved. We now, however, have computational tools at hand that enable the recognition of errors in experimentally determined and model structures. Furthermore, fold-recognition techniques are enabling molecular architectures of proteins to be correctly predicted, before their experimentally determined structures are known.

The recent progress has been achieved by new approaches in the development of energy functions. These so-called knowledge-based force fields, or mean force potentials, are derived from experimental data. The basic idea is not new. The molecular structures observed by X-ray analysis or NMR contain a large amount of information on the stabilizing forces within proteins, and statistical analysis has the potential to reveal the underlying rules governing protein stability. In combination with statistical mechanics, the stastistical analysis of known three-dimensional structures of proteins is indeed a powerful tool to extract potential functions from a database of known structures.

Several recent reviews on knowledge-based potentials are available (e.g. [2–7,8•,9•]). Here I concentrate on the characteristics of mean force potentials and the techniques used to investigate their performance and predictive power.

Design principles for molecular energy functions

Construction of energy functions for molecular systems is usually based on several assumptions: first, it is supposed that the behaviour of molecular systems can be captured by a free energy function; second, the potential energy part of the system can be approximated by two-body interactions; and third, molecular structures observed with high frequency correspond to low-energy states. The last statement is a consequence of Boltzmann's principle, which essentially states that probability densities and energies are closely related quantities. Specifically, in the case of protein–solvent systems, the native structures are thought to correspond to the lowest energy states accessible in equilibrium and the assumption is that it is possible to construct energy functions on the basis of intramolecular and intermolecular atomic pair interactions, whose global minima correspond to native folds.

The construction of potential functions for pair interactions has been attempted from different directions, (see [7] for example). The approach from first principles starts from basic physical laws. Examples include potentials for the electrostatic interactions based on Coulomb's law, or the Lennard-Jones type potential, whose functional form follows from quantum mechanical calculations. The second type of approach exploits the ever increasing knowledge base of experimentally determined structures relying explicitly or implicitly on Boltzmann's principle: frequently observed states correspond to low energy states of the system.

The idea to exploit databases for structure prediction has a long tradition [10]. Prediction of secondary structures by observed preferences of amino acids to adopt helix or strand conformations is a popular example. Similarly, statistical potentials have a respectable history. Early attempts to derive potentials from a database of structures were reported almost twenty years ago by Tanaka and

Abbreviations

a—atom *a*; ***b***—atom *b*; $\bar{E}$—average energy; ***E*(*r*)**—energy at *r*; ***f*(*r*)**—probability density at *r*; ***i***—protein; ***k***—Boltzmann's constant; ***r***—distance between two atoms; ***s***—separation of amino acids; **σ**—standard deviation; ***T***—absolute temperature.

© Current Biology Ltd ISSN 0959-440X

Scheraga [11], and many others have reported subsequent attempts in the intervening period (see e.g. [12–16], and [17–24] for more recent developments).

Potentials of mean force

In the following I focus on mean force and related potentials. The general definition of database-derived mean force potentials is [16]

$$E(r) = -kTln[f(r)]$$

where r is the distance (or some other parameter, like dihedral angles) between two atoms, $E(r)$ is the energy at r, $f(r)$ is the probability density at r, k is Boltzmann's constant and T is the absolute temperature. Besides r, a particular pair interaction depends on the atom types a and b involved (e.g. an interaction between the Cβ atom of a valine residue and the Cα atom of a glycine), and the separation s of the respective amino acids along the amino acid sequence [16] (this parameter is important for small separations, for example $s<10$; for $s \geq 10$, atoms can be considered as free particles):

$$E^{abs}(r) = -kTln\left[f^{abs}(r)\right].$$

$f(r)^{abs}$ is approximated by relative frequencies obtained from a data base of known structures.

Mean force potentials incorporate all forces (electrostatic, van der Waals, etc.) acting between atoms as well as the influence of the surrounding medium on the interaction. In this form individual potentials contain more or less the same information, but we need the specific information contained in a particular potential that distinguishes it from an average interaction in the system being studied. The redundant information can be captured by a suitably defined reference state. In the case of protein intramolecular interactions a convenient choice for the reference state is [7,16]

$$E^{s}(r) = -kTln[f^{s}(r)] \quad \text{where} \quad f^{s}(r) = \sum_{ab} f^{abs}(r)$$

which is an average over all atom and residue types. Subtracting this redundant information we obtain the specific interaction

$$\Delta E^{abs}(r) = E^{abs}(r) - E^{s}(r) = -kTln\left[\frac{f^{abs}(r)}{f^{s}(r)}\right]$$

The reference state is a critical feature. Successful applications of mean force potentials largely depend on a suitable reference state.

A characteristic feature of molecular force fields

The detailed features of molecular energy functions that govern the folding and stability of proteins are unknown, but some general principles follow from basic physical considerations. Consider a particular protein defined by its amino acid sequence. All possible conformations have associated energy values, and the energy density $N(E)$ (i.e. the number of conformations per energy interval) characterizes the energy distribution for this protein. By the law of large numbers we might guess that the energy density resembles a Gaussian distribution defined by the average energy $\bar{E}$ and standard deviation σ (Fig. 1). In fact, we do not know the shape of this distribution, but every distribution has an average and a standard deviation, and we can use these numbers to normalize energy values [7,25••,26••], $E \rightarrow (E-\bar{E})/\sigma = z$ (these normalized numbers are called z-scores). $\bar{E}$ and σ are sequence-specific parameters, but they are independent of any particular conformation, as they are average quantities in conformation space.

From the principles of statistical mechanics we know that the energy of the native structure has to be much lower than the average energy and it has to have the minimum energy among all (accessible) conformations. In other words, a native fold has a large negative z-score [7,25••,26••]. Energy functions designed for protein–solvent systems must have the same property. In fact, this principle is very useful in judging the quality of molecular force fields, as the z-scores depend on the energy function employed [25••].

The structures of several hundred proteins are known and the energy E_i , as defined by some energy function, can be calculated for each protein i. What is the significance of the energy of a particular conformation and what does it tell us about the quality of the energy function? To get an answer we have to relate E_i to the energy distribution of the respective sequence in conformation space [7]. This can be done by estimating $\bar{E}_i$ and σ_i so that energies can be transformed to z-scores.

As $\bar{E}_i$ and σ_i are average values, they can be estimated from a small sample of conformation space. A popular sampling strategy is to derive fragments from the known protein structures [18,27,28,29••,30•,31•]. This procedure has several drawbacks (e.g. the sample space for larger proteins is very small), which can be avoided when the structures are joined to a polyprotein [25••,26••].

As an example, Fig. 1 shows the mean force energy density of lysozyme 1LZ3 (PDB code) obtained from a polyprotein. The average performance of an energy function over the set of known protein structures can be expressed by the average score

$$\bar{z} = \sum z_i/n,$$

over the individual proteins, where n is the number of proteins in the test set.

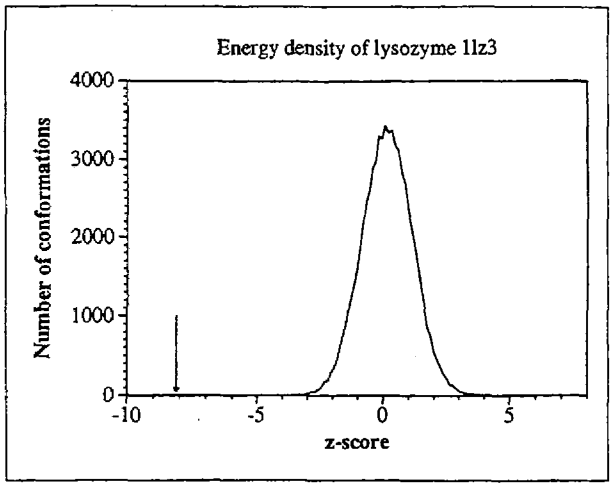

Fig. 1. Energy density of the sequence of lysozyme (PDB code 1lz3) derived from a polyprotein. The total number of conformations is ≈50 000. The arrow marks the position of the native fold of 1lz3.

Extraction of mean force potentials from a data base

Mean force potentials are compiled by extracting relative frequencies from a database of protein structures. Several problems are involved in this process. Perhaps the most serious one arises from low counts. In the case of potentials depending on s, the average number of counts is ≈100. For rare amino acid pairs like methionine and tryptophan this number is close or equal to zero. Approximation of functions on the basis of such a small number of counts is difficult or impossible in general, but methods have been developed to approximate $f^{abs}(r)$ in cases of extremely low counts [16].

The performance and predictive power of mean force potentials depend on a few critical parameters. For example, for distances r >30 Å, frequencies are dominated by the large proteins in the data base. What is a useful cut-off distance at which to truncate potentials? A reasonable value can be found by calculating the average score $\bar{z}$ as a function of cut-off distance. For Cβ–Cβ interactions $\bar{z}$ increases (in absolute value) up to 20 Å [25••]; in other words, the information content is maximized when potentials are compiled up to distances of 20 Å.

The dependence of $\bar{z}$ on other interesting parameters can also be determined in this way [25••,31•]. How does the quality of potentials depend on the number of proteins in the data base? $\bar{z}$ as a function of database size follows an exponential saturation. Increasing the database from 50 to 100 proteins yields a 30% increase in $\bar{z}$, whereas between 200 and 250 proteins the gain is only 5% [25••]. How are intramolecular pair interactions related to protein–solvent interactions? The scores $\bar{z}$ obtained for the individual terms are of comparable size (–6.8 for Cβ pairwise and –6.2 for protein–solvent interactions), but when the two terms are combined the scores increase significantly to –9.66 (M Jaritz, MJ Sippl, unpublished data). In other words, the information contained in intramolecular pair interactions is quite different from protein–solvent interactions and both components are important. Another problem concerns the information contained in pair interactions compiled from different types of atoms. The score of –5.0 for Cα–Cα potentials is less significant as compared to –6.76 for Cβ–Cβ interactions. Their combination scores at an intermediate value of –6.2, in other words, the information contained in these terms is highly redundant (M Jaritz, MJ Sippl, unpublished data).

Our current implementation of mean force potentials consists of pair interactions among all backbone atoms (N, Cα , C, O) and Cβ , and an explicit term for protein–solvent interactions [7,8•]. As discussed above, the polyprotein technique has been extensively used to optimize the performance of this energy function [25••,31•].

Applications

Mean force potentials have been successfully applied to various problems in structural biology, such as recognition of errors in experimentally determined structures or the prediction of protein folds by sequence/structure combination.

Detection of errors in protein structures

In several cases errors have been detected in experimentally determined structures [32,33]. Some of the faulty structures have been deposited with the Brookhaven data bank [34] where the faults remained undetected for years, indicating that the criteria used to judge the quality of experimentally determined structures failed in these cases. With the advent of several new programs the situation has improved considerably [26••,35–37]. Native folds have mean force z-scores in a characteristic range, and the energy distribution within native folds shows a characteristic pattern. Erroneous and deliberately misfolded structures are detected by their poor scores and unusual energy distributions [26••]. Because mean force calculations can be done on reduced sets of atoms it is possible to analyze structures where only the Cα backbone is available.

Fig. 2 shows the energy graphs of two experimentally determined structures, photoactive yellow protein and lysozyme (PDB codes 1phy and 2lzh, respectively). Only the Cα coordinates of these structures were deposited with the Brookhaven protein data bank. The z-score for 2lzh of –8.2 is typical for native structures, but the z-score for 1phy (–1.6) points to an erroneous fold, and the energy graph of 1phy shows that the interactions in this molecule are unfavourable. The z-scores and energy graphs were calculated using the program PROSA-II [26••] (PROtein Structure Analysis), which is available from gundi.came.sbg.ac.at by anonymous file transfer protocol (ftp).

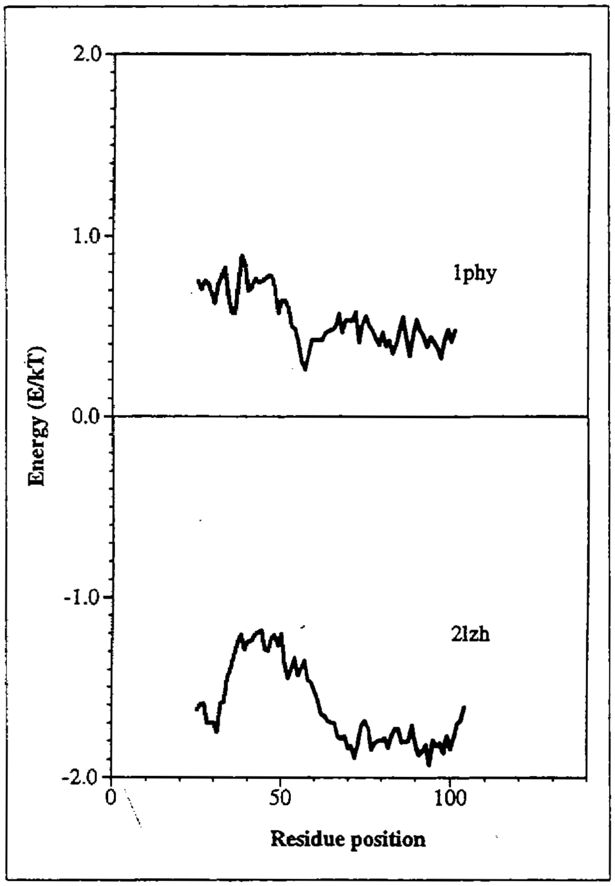

Fig. 2. Energy graphs for (a) photoactive yellow protein (PDB code 1phy) and (b) lysozyme (PDB code 2lzh) generated by PROSA-II [26••]. The graph of 2lzh is typical for native folds, but that for 1phy is problematic.

Quality assessment is an important tool for the judgement of experimental structures and it is a fundamental prerequisite for protein structure theory in general. Prediction of protein structures from amino acid sequences will necessarilly fail if faulty structures cannot be distinguished from genuine native folds [38].

Fold recognition

Proteins frequently have similar three-dimensional folds even in cases where no homology is discernible on the sequence level [39••,40••]. It is expected that in a substantial number of cases the unknown structures of protein sequences resemble some known fold. By combining sequences with structures it should be possible to identify such coincidences [41], and there has been considerable progress in the development of fold recognition techniques over the past few years [41–49,50••].

There are three critical components of fold recognition techniques: first, energy functions or parameter sets providing a reasonable description of protein–solvent systems; second, techniques producing useful alignments between sequences and structures; and finally, criteria for identifying native-like sequence/structure combinations [49].

The performance of fold recognition techniques is documented by several detailed case studies [41–49,50••] and several methods are able to recognize distant relationships, such as the similarity between the ADP-ribosylation factor and Ras p21 [49]. Using a database of 150 pairs of proteins related in structure, but unrelated or distantly related in sequence, our implementation based on mean force potentials successfully identifies the related fold in roughly one out of three cases (MJ Sippl, unpublished data).

The most serious challenge testing our current ability to design fold-recognition techniques was the recent prediction experiment organized by J Moult and others. Sequences of several proteins were collected from laboratories that were close to solving the respective structures. Predictions for the individual targets were accepted until the structure was solved and finally the quality of the predictions was judged by a team of assessors. The folds of several proteins were correctly identified by various groups employing fold recognition and/or multiple sequence alignment techniques ([51••] and T Hubbard, personal communication; an issue of *Proteins* dedicated to this prediction experiment is scheduled for 1995).

In some cases the predictions were close to atomic resolution. For the replication terminating protein (PDB code rtp), no homologous sequence could be found by sequence comparison techniques. Fold recognition using mean force potentials suggested the structure of histone (PDB code 1hst) to be a good model for the fold of this protein. The prediction was correct: replication terminating protein and histone indeed have very similar structures. Moreover, the alignment was of excellent quality. Gaps were inserted correctly and the residues of replication terminating protein were placed at the appropriate positions in the histone structure. For the propiece of subtilisin, fold recognition predicted the structure of ferredoxin (PDB code 2fxd). The two structures can be superimposed to a root mean square error of 3.5 Å.

A critical point in fold recognition concerns the quality assessment of a particular model. Any sequence can be combined with any structure. But the question is whether a particular combination corresponds to a useful model that is to some degree similar to the unknown fold of the respective sequence. An important result is that the energy calculated from a model can be misleading. Rather, the energy of the model has to be compared to the energy distribution in conformation space [7,49]. This can be achieved by calculating z-scores as described above.

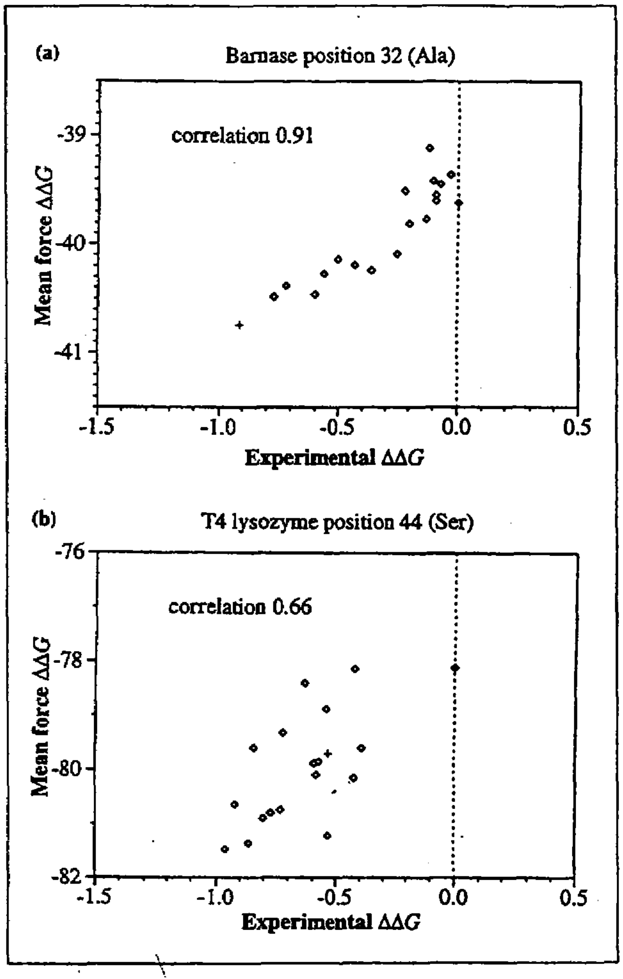

Fig. 3. Calculated $\Delta\Delta G$ values plotted against experimental data. Every point in the diagrams corresponds to a particular substitution. **(a)** In barnase, Ala32 was replaced by all 19 standard amino acids. **(b)** In T4 lysozyme, Ser44 was replaced by all 19 standard amino acids. The correlation between measured and calculated $\Delta\Delta G$ values is 0.91 for barnase and 0.66 for T4 lysozyme.

Genome sequencing projects produce a large number of new sequences. For approximately half of the new genes discovered a biological role or function can be assigned by sequence comparison. The biological information contained in the remaining sequences is not accessible. In a recent large-scale fold recognition study on the 483 genes found in the central gene cluster of *Caenorhabditis elegans* chromosome III, using a data base of 263 known structures, putative models for the unknown folds of 20 sequences have been obtained (M Braxenthaler, MJ Sippl, unpublished data). In light of the successful blind predictions described above, the study demonstrates that fold recognition generates valuable structural and functional information for otherwise uncharacterizable genes and that large-scale applications are computationally feasible.

Protein stabilities

A vital requirement for rational protein engineering and design is the ability to predict the effect of amino acid replacements on the stability of proteins. In some cases experimental results are well documented. In the case of barnase, Alan Fersht's group has collected a large number of $\Delta\Delta G$ values for various mutations ([52]; $\Delta\Delta G$ is the change in stability between wild-type and mutant protein). As shown in Fig. 3, $\Delta\Delta G$ values calculated from mean force potentials correlate well with measured data (M Hendlich, MJ Sippl, unpublished data).

In other cases the correspondence is less satisfying. Proteins are represented by their Cα backbones only and it is assumed that the Cα backbone is not changed by amino acid replacements. Hence, the present calculations contain assumptions and approximations that may be valid in some situations but may be too crude in others. The applicability of knowledge-based potentials in the design of sequences that fold into predefined structures, a problem closely related to protein stabilities, has been investigated by Jones [53••].

Ab initio prediction

The long-range goal of force-field development is the *ab initio* prediction of protein structures solely from the information contained in amino acid sequences by energy minimization and folding simulations. Some preliminary studies on small proteins have been performed (e.g. [8•,54,55•,56•]). For example, calculations on thymosin β_4 are in good agreement with models obtained from NMR studies [54], and computations on the Antennapaedia peptide, a small three-helix bundle, come close to the observed structure [8•].

Conclusions

Knowledge-based force fields have matured into useful tools, providing the basis for powerful techniques in many areas of research into protein structure. They will help to identify the function of protein sequences and aid the determination of their structures. In spite of these successes, the development of force fields and associated methods, like fold recognition, is still at the beginning and there are several problem areas where improvements are possible.

It is clear that *ab initio* prediction, reliable estimation of $\Delta\Delta G$ values, molecular docking and other problems in structural biology require a more detailed representation of molecular structures and atomic interactions than is currently available. The development of such force fields remains a major challenge.

Acknowledgements

This work was supported by the Fonds Wissenschaftlicher Forschung Austria, grant 9661-MOB and by the Nationalbank of Austria, grant 5158.

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. Lattman EE, Rose GD: **Protein folding — what's the question?** *Proc Natl Acad Sci USA* 1993, **90**:439–441.

2. Wodak SJ, Rooman MJ: **Generating and testing protein folds.** *Curr Opin Struct Biol* 1993, **3**:247–259.

3. Bowie JU, Eisenberg D: **Inverted protein structure prediction.** *Cur Opin Struct Biol* 1993, **3**:437–444.

4. Rost B, Sander C: **Structure prediction of proteins - where are we now?** *Curr Opin Biotechnol* 1994, **4**:372–380.

5. Abagyan RA: **Towards protein folding by global energy optimization.** *FEBS Lett* 1993, **325**:17–22.

6. Jones DT, Thornton JM: **Protein fold recognition.** *J Comput Aided Mol Des* 1993, **7**:439–438.

7. Sippl MJ: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *J Comput Aided Mol Des* 1993, **7**:473–501.

8. Sippl MJ, Weitckus S, Floeckner H: **In search of protein folds.**
• In *The protein folding problem and tertiary structure prediction.* Edited by Merz KH, LeGrand S. Boston: Birkhäuser; 1994:353–407.

The review presents basic principles of mean force potentials and contains a case study on fold recognition. 24 000 amino acid sequences are combined with an immunoglobulin fold and the high-scoring sequences are analyzed in detail.

9. Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL:
• **Knowledge-based protein modeling.** *Crit Rev Biochem Mol Biol* 1994, **29**:1–68.

A detailed review of techniques for modeling protein structures on homologous templates.

10. Fasman GD: **Development of protein structure prediction.** In *Prediction of protein structure and the principles of protein conformation.* Edited by Fasman GD. New York: Plenum Press; 1989:193–316.

11. 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.

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

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

14. Eisenberg D, McLachlan AD: **Solvation energy in protein folding and binding.** *Nature* 1986, **319**:199–203.

15. Wilson C, Doniach S: **A computer model to dynamically simulate protein folding: studies with crambin.** *Proteins* 1989, **6**:193–209.

16. 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.

17. Holm L, Sander C: **Evaluation of protein models by atomic solvation preference.** *J Mol Biol* 1992, **225**:93–105.

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

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

20. Avbelj F: **Use of a potential of mean force to analyze free energy contributions in protein folding.** *Biochemistry* 1992, **31**:6290–6297.

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

22. Karlin S, Zucker M, Brocchieri L: **Measuring residue associations in protein structures — possible implications for protein folding.** *J Mol Biol* 1994, **239**:227–248.

23. Madej T, Mossing MC: **Hamiltonians for protein tertiary structure prediction based on three-dimensional environment principles.** *J Mol Biol* 1993, **233**:480–487.

24. Zhang KY, Eisenberg D: **The three-dimensional profile method using residue preference as a continuous function of residue environment.** *Protein Sci* 1994, **3**:687–695.

25. Sippl MJ, Jaritz M: **Predictive power of mean force pair poten-**
•• **tials.** In *Protein structure by distance analysis.* Edited by Bohr H, Brunak S. Amsterdam: IOS Press; 1994:113–134.

Protein sequences can adopt an enormous number of folds. A frequent problem is to determine whether or not a given sequence/structure combination corresponds (or is close) to the fold of a native protein. This requires the determination of the energy distribution in conformation space. The technique employed here uses a polyprotein, enabling efficient calculation of average energy and standard deviation for the energy distribution. The parameters obtained are used to transfrom conformational energies to scores expressing the significance of sequence/structure combinations.

26. Sippl MJ: **Recognition of errors in three-dimensional structures**
•• **of proteins.** *Proteins* 1993, **17**:355–362.

A program to recognize erroneous folds or faulty parts of structures obtained by X-ray or NMR analysis or by modeling studies. The program, PROSA-II, is available via anonymous ftp (file tranfer protocol) from gundi.came.sbg.ac.at.

27. 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.

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

29. Kocher JPA, Rooman MJ, Wodak SJ: **Factors influencing**
•• **the ability of knowledge based potentials to identify native sequence-structure matches.** *J Mol Biol* 1994, **235**:1598–1613.

These authors evaluate the predictive power of various knowledge-based potentials and statistically derived parameters. A potential based on side-chain centroids is found to score highest in the evaluation scheme they define.

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

The quality of knowledge-based potentials depends on the number of protein structures available for statistical analysis. Sparse data pose a problem, especially for the rare amino acids. Using a weighting scheme, based on amino acid mutation frequencies, potentials of similar amino acids are combined to a single type.

31. Sippl MJ, Jaritz M, Hendlich M, Ortner M, Lackner P: **Appli-**
• **cations of knowledge based mean fields in the determination of protein structures.** In *Statistical mechanics, protein structure and protein-substrate interactions.* Edited by Doniach S. New York: Plenum Publishers; 1994:297–315.

Mean force pair potentials decribe intramolecular interactions, but explicit protein solvent terms are important components of the system.

When pair potentials are combined with solvent terms the predictive value of the energy function increases almost twofold, demonstrating the complementary nature of these energy terms.

32. Bränden C-I, Jones TA: **Between objectivity and subjectivity.** *Nature* 1990, **343**:687–689.

33. Janin J: **Errors in three dimensions.** *Biochimie* 1990, **72**:705–709.

34. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: **The protein data bank: a computer-assisted archival file for macromolecular structures.** *J Mol Biol* 1977, **112**:535–542.

35. MacArthur MW, Laskowski RA, Thornton JM: **Knowledge based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy.** *Curr Opin Struct Biol* 1994, **4**:731–737.

36. Luethy R, Bowie JU, Eienberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356**:83–85.

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

38. Novotny J, Bruccoleri R, Karplus M: **An analysis of incorrectly folded protein models.** *J Mol Biol* 1984, **177**:787–818.

39. Holm L, Sander C: **Searching protein structure data bases has**
•• **come of age.** *Proteins* 1994, **19**:165–173.

Structures in the Brookhaven data base are compared and a spanning tree is constructed displaying the structural hierarchy among the various folds. Detailed databases of structural relationships among proteins are extremely important for fold recognition tecchniques.

40. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and**
•• **domain superfolds.** *Nature* 1994, **372**:631–634.

This paper describes work similar to that in [39••], but the relationships are obtained by different techniques. In cases of very similar structures both techniques yield the same results. For very distantly related cases, the methods produce complementary data.

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

42. Sippl MJ, Weitckus S: **Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations.** *Proteins* 1992, **13**:258–271.

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

44. Godzik A, Kolinski A, Skolnick J: **Topology fingerprint approach to the inverse folding problem.** *J Mol Biol* 1992, **227**:227–238.

45. Ouzounis C, Sander C, Scharf M, Schneider R: **Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from 3D structures.** *J Mol Biol* 1993, **232**:805–825.

46. Wilmans M, Eisenberg D: **Three-dimensional profiles from residue-pair preferences: identification of sequences with beta/alpha-barrel fold.** *Proc Natl Acad Sci USA* 1993, **90**:1379–1382.

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

48. Johnson MS, Overington JP, Blundell TL: **Alignment and searching for common protein folds using a data bank of structural templates.** *J Mol Biol* 1993, **231**:735–752.

49. Sippl MJ, Weitckus S, Floeckner H: **Fold recognition.** In *Modelling of biomolecular structures and mechanisms.* Edited by Pullman A, Jortner J, Pullman B. Kluwer; 1995 in press.

50. Lathrop RH: **The protein threading problem with sequence**
•• **amino acid interaction preferences is NP-complete.** *Protein Eng* 1994, **7**:1059–1068.

One goal in fold recognition is the calculation of optimal alignments between sequences and structures. In contrast to sequence/sequence comparison, the problem is found to be NP-complete. The consequence is that the problem is computationally intensive, and it is unlikely that fast algorithms can be found that at the same time guarantee an optimal solution.

51. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hid-**
•• **den Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**:1501–1531.

Hidden Markov chains are applied to calculate local structures in proteins. As demonstrated recently by Tim Hubbard in a blind test, results obtained by Hidden Markov chains can be used to identify protein folds.

52. Fersht AR, Serrano L: **Principles of protein stability derived from protein engineering experiments.** *Curr Opin Struct Biol* 1993, **3**:75–83.

53. Jones DT: **De-novo protein design using pairwise potentials**
•• **and a genetic algorithm.** *Protein Sci* 1994, **3**:567–574.

Mean force potentials are used to derive sequences compatible with a given fold. The sequence of a starting protein is forced to change while the conformation is kept fixed. Mutations are only accepted if the stability of the protein does not deteriorate.

54. Sippl MJ, Hendlich M, Lackner P: **Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments. Development of strategies and construction of models for myoglobin, lysozyme and thymosin β-4.** *Protein Sci* 1992, **1**:625–640.

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

Ab initio folding studies are attempted on the myoglobin sequence using knowledge-based potentials.

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

The native topology of four helix bundle proteins can be found by minimizing knowledge-based potentials, when the known secondary structure is preformed.

MJ Sippl, Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob Haringer Straße 1, A-5020 Salzburg, Austria.