10

Scoring Functions for De Novo Protein Structure Prediction Revisited

Shing-Chung Ngan, Ling-Hong Hung, Tianyun Liu, and Ram Samudrala

Summary

De novo protein structure prediction methods attempt to predict tertiary structures from sequences based on general principles that govern protein folding energetics and/or statistical tendencies of conformational features that native structures acquire, without the use of explicit templates. A general paradigm for de novo prediction involves sampling the conformational space, guided by scoring functions and other sequence-dependent biases, such that a large set of candidate ("decoy") structures are generated, and then selecting native-like conformations from those decoys using scoring functions as well as conformer clustering. High-resolution refinement is sometimes used as a final step to fine-tune native-like structures. There are two major classes of scoring functions. Physics-based functions are based on mathematical models describing aspects of the known physics of molecular interaction. Knowledge-based functions are formed with statistical models capturing aspects of the properties of native protein conformations. We discuss the implementation and use of some of the scoring functions from these two classes for de novo structure prediction in this chapter.

Key Words: De novo; physics-based; knowledge-based; potential; protein folding.

1. Introduction

The success of large-scale genome sequencing efforts has spurred structural genomic initiatives, with the goal of determining as many protein folds as possible ***(1–4)***. At present, structural determination by crystallography and nuclear magnetic resonance (NMR) techniques are still slow and expensive in terms of manpower and resources, despite attempts to automate the

From: *Methods in Molecular Biology, vol. 413: Protein Structure Prediction, Second Edition*
Edited by: M. Zaki and C. Bystroff © Humana Press Inc., Totowa, NJ

processes. Computational structure prediction algorithms, while not providing the accuracy of the traditional techniques, are extremely quick and inexpensive and can provide useful low-resolution data for structure comparisons ***(5)***. Given the immense number of structures that the structural genomic projects are attempting to solve, there would be a considerable gain even if the computational structural prediction approach were applicable only to a subset of proteins.

Most current research in protein structure prediction is based on Anfinsen's thermodynamic hypothesis that the native structure of a protein can be determined entirely from its amino acid sequence ***(6)***. The two main categories of methods for predicting protein structure from sequence are comparative and de novo modeling. In the comparative modeling category, the methodologies rely on the presence of one or more evolutionarily related template protein structures that are used to construct a model. Traditionally, the evolutionary relationship can be deduced from sequence similarity ***(7–9)*** or by "threading" a sequence against a library of structures and selecting the best match ***(10,11)***. However, because of the improved sensitivity of the sequence similarity based methods, the threading approach has essentially been supplanted ***(12,13)***. In the de novo category, structure prediction methods attempt to predict tertiary structures from sequences based on general principles that govern protein-folding energetics and/or statistical tendencies of conformational features that native structures acquire, without the use of explicit templates ***(14–16)***. A general paradigm for de novo structure prediction involves sampling the conformational space, guided with scoring functions and other sequence-dependent biases, such that a large set of candidate ("decoy") structures are generated, and then selecting native-like conformations from those decoys using scoring functions and conformer clustering as filters ***(17,18)***. As a final step, detailed energy potentials are sometimes employed to perform high-resolution refinement on these native-like structures. Although the first papers on protein structure prediction appeared some thirty years ago, de novo structure prediction remains a difficult challenge today ***(12,13,19–21)***.

Scoring functions are employed in all stages of de novo structure prediction. For the conformational search stage, a selected combination of scoring functions approximates the energy landscape of the protein conformational space. Search methodologies such as Monte Carlo simulated annealing (MCSA) and molecular dynamics (MD) then generate trajectories leading to the minima of the landscape. As the conformational search process needs to evaluate new conformations encountered at every step, it is computationally intensive, and the scoring functions used in this stage need to be computationally efficient. Because none of the existing scoring functions can faithfully reproduce the

true energy landscape of the conformational space, the search process often leads to many false minima. Thus, one usually repeats the search process many times with many different starting conditions and random seeds and obtains a collection of candidate ("decoy") structures. Then, a second set of (possibly different) scoring functions are used in the decoy selection stage as filter to eliminate non-native structures and retain the native-like ones. Conformer clustering is often used as an additional step to further refine the collection of the native-like conformations, followed by high-resolution refinement of the few remaining candidate structures. Compared to the functions used in the conformational search stage, the functions employed in the decoy selection stage can be algorithmically more complex and more detailed, because the number of candidate conformations to evaluate is much less than the number of conformations encountered during the search process. Scoring functions used in the high-resolution refinement stage are usually computational expensive functions formulated from detailed mathematical models of short-range interactions among atoms, allowing small local perturbations to fine-tune native-like structures.

There are two broad classes of scoring functions. The first class of functions are largely based on some aspects of the known physics of molecular interaction, such as the Van der Waals force, electrostatics, and the bending and torsional forces, to determine the energy of a particular conformation ***(22–27)***. The second class of functions is knowledge-based. Each of these knowledge-based functions tries to capture some aspects of the properties of protein native conformations, for example, the tendencies of certain residues to form contact with one another or with the solvent. These knowledge-based functions are usually compiled based on the statistics of a database of experimentally determined protein structures ***(28–34)***. In essence, the physics-based functions aim at predicting the native structure of a given sequence by mimicking the energetics of protein folding, whereas the knowledge-based functions bypass this intermediate step by directly making statistical inferences on what are observed in the database. Thus, the accuracy of the physics-based functions is determined by how realistic the underlying physical models are, whereas the accuracy of the knowledge-based functions is determined by the quality of the database as well as the validity of the statistical assumptions.

In an earlier edition, we introduced scoring functions for de novo structure prediction ***(35)***. In this chapter, we revisit physics-based and knowledge-based scoring functions in the context of their roles in the current state of the art structure prediction efforts. For the physics-based approach, the often-called Class I force field, which is a common foundation among the widely used

molecular modeling force fields such as AMBER, CHARMM, OPLS, and ENCAD, is discussed. Extensions to this force field and the role of modeling solvent effects are also described. For the knowledge-based approach, we study the Bayesian (conditional) probability formalism, using it to derive the all-atom distance-dependent conditional probability discriminatory function (RAPDF) ***(34)***. As an additional illustration, we delineate how one can combine the Bayesian probability formalism with the neural network methodology to construct neural network-based scoring functions. Then, a few other novel knowledge-based scoring functions from the recent literature are highlighted. Although it is not strictly a physics- or knowledge-based methodology, we briefly discuss the use of conformer clustering to further enhance decoy selection, as this technique has been shown to be useful in de novo structure prediction. Finally, a sophisticated combined physics- and knowledge-based potential used for high-resolution refinement is described.

2. Theoretical Background and Methods

2.1. An Overview of Physics-Based Energy Functions

Using quantum mechanical techniques, highly accurate energies can be calculated for small organic and inorganic molecules ***(36,37)***. However, because of their sizes and flexibility as well as the presence of solvent molecules, proteins are much more difficult systems to model. The polar aqueous environment vastly complicates the calculation of the electrostatic energies. For instance, although there is no dispute that the largest driving force for protein folding is the hydrophobic effect ***(38,39)***, which is associated with the decrease of water entropy upon the solvation of non-polar groups, the exact structural configuration of water molecules hydrating the solute remains unknown.

Although a full quantum mechanical treatment for a complete protein is not feasible, approximations and simplifications can be made to derive empirical physics-based energies. For example, hydrogen bond geometries that are applicable to those found in proteins can be determined from quantum mechanical calculations of simple systems ***(40)***. Electrostatics calculations can be approximated using classical point charges and modifying the dielectric constant to approximate the polarizability of the protein and the solvent. Van der Waals interactions are often approximated by Lennard–Jones potentials. The first use of these approximate functions was in MD simulations, where fast and easily calculated energies were required to determine the force fields. Some prototypes for these types of energies are AMBER ***(41)***, CHARMM ***(42)***, OPLS ***(24)***, and ENCAD ***(43)***. Parameters for these energies have been obtained by fitting equations and results of computer simulations to data from experiments and

from quantum mechanical calculations. These physics-based energies perform adequately for perturbations around a known native conformation ***(44,45)***, because the electrostatic and solvent-dependent information is implicit in the initial conformation itself. In combination with experimental NMR constraints ***(46,47)***, these force fields enable the determination of accurate structures, so long as there are enough constraints to define the fold. Unfortunately, in isolation, the solvent and electrostatic modeling is insufficient for full and reliable simulation of protein folding. As a result, producing accurate protein folding simulations from physics-based energies alone is still a very challenging and active area of research.

2.1.1. Class I Physics-Based Scoring Function and Its Possible Extensions

As we have mentioned, AMBER ***(41)***, CHARMM ***(42)***, OPLS ***(24)***, and ENCAD ***(43)*** are some examples of the widely used physics-based force fields in protein-folding simulation. These force fields share a lot of commonalities in terms of the underlying physical models used and the mathematical approximations assumed. As an illustration, the AMBER force field, which was first developed under the direction of Professor Peter Kollman, has the following form:

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{non-bond}} \tag{1}$$

Here, V_{total} is the total potential energy, V_{bond} is the bond stretching energy, V_{angle} the angle bending energy, and V_{torsion} the angle torsional energy. Together, V_{bond}, V_{angle}, and V_{torsion} are denoted as the bonded interactions terms. $V_{\text{non-bond}}$ is the energy for non-bonded interactions, consisting of a Van der Waals energy term V_{vdW} and an electrostatics term V_{elec}. Other widely used force fields such as CHARMM and OPLS employ similar bonded and non-bonded terms in their formulations, and Eq. 1 is often denoted as the Class I force field.

The bond-stretching energy (*see* **Fig. 1A**) is modeled by treating the bond as an idealized spring and using a simple quadratic function derivable from the Hooke's law.

$$V_{\text{bond}} = k_{\text{bond}}(r - r_{\text{o}})^2 \tag{2}$$

where k_{bond} is the bond-stretching constant, controlling the stiffness of the bond spring, and $(r–r_{\text{o}})$ is the deviation of the bond length from its equilibrium distance. Unique numerical values for k_{bond} and r_{o} are assigned each pair of atom types.

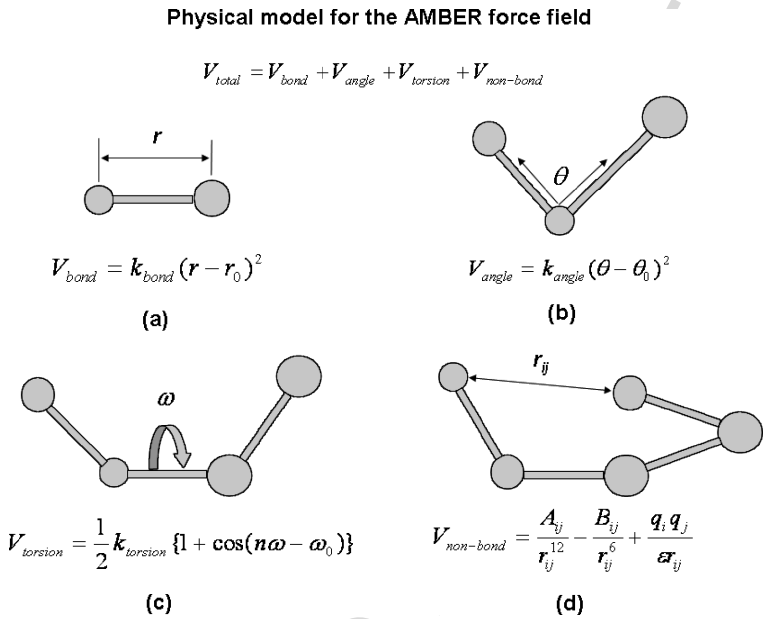


Fig. 1. The physical models for the AMBER molecular mechanics force field. Atoms and bonds are shown. **(A)** The physical model for bond stretching, **(B)** the model for angle bending, **(C)** the model for angle torsional energy, and **(D)** the model for electrostatics and Van der Waals forces.

The angle bending energy (*see* **Fig. 1B**) is similarly modeled by the Hooke's law.

$$V_{\text{angle}} = k_{\text{angle}}(\theta - \theta_{\text{o}})^2 \tag{3}$$

where k_{angle} is the angle bending constant, controlling the stiffness of the angle spring. θ is the angle formed by the atom of interest with its two covalently bonded neighbors, and $(\theta - \theta_{\text{o}})$ is the deviation of the angle from its equilibrium value in radians. Again, unique values for k_{angle} and θ_{o} are determined for each bonded triplet of atom types.

The torsional energy (*see* **Fig. 1C**) is represented by an n-fold periodic function:

$$V_{\text{torsion}} = \frac{1}{2}k_{\text{torsion}}[1 + \cos(n\omega - \omega_0)] \tag{4}$$

Here, the torsional angle ω is the dihedral angle defined by a quartet of bonded atoms, and ω_0 is the reference angle. k_{torsion} is a constant for the

n-fold periodic interaction. *n* represents the periodicity of the torsional barrier, reflecting the intrinsic symmetry in the dihedral angle for the quartet of the bonded atoms. Unique values of k_{torsion}, n, and ω_0 are assigned to each bonded quartet of atom types. In practice, parameterization of torsional energies also corrects for bonding energy terms unaccounted for by the simple bending and stretching models. Additional torsional energy terms (denoted as "improper torsions" in the literature) can be added to ensure that subtle properties such as chirality and planarity are preserved.

For the non-bonded interactions, AMBER and other commonly used force fields employ a 6–12 Lennard–Jones potential to represent the Van der Waals interactions between two non-bonded atoms, and the Coulomb's law to model the interactions of two charged atoms (*see* **Fig. 1D**):

$$V_{\text{non-bond}} = \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \left(\frac{q_i q_j}{\varepsilon r_{ij}}\right) \tag{5}$$

The Van der Waals interaction consists of two components, a short-range attractive force that quickly vanishes when the distance between the interacting atoms, r_{ij}, is greater than a few Angstrom and an even shorter-range repulsive force that dominates when r_{ij} is less than the sum of their individual atomic radii. B_{ij} and A_{ij} in Eq. 5 control the attractive and the repulsive components of the steric potential. A_{ij} can be calculated from quantum mechanics considerations or measured from atomic polarizability experiments, and B_{ij} can be calculated from crystallographic data. For the eletrostatics, interacting atoms are treated as point charges of q_i and q_j. The value of the dielectric constant ε accounts for the attenuation of electrostatic interaction by the polar environment. In more sophisticated solvent models, which are discussed later, the constant ε is replaced by a function dependent on r_{ij}. Earlier versions of AMBER had an explicit term to take into account hydrogen bonding. The latest versions incorporate hydrogen-bonding effects into the parameterization of the electrostatic and van der Waals terms, as these two terms are found to be able to sufficiently represent the distance and angle dependencies of hydrogen bonds in molecular mechanics modeling ***(48)***.

Currently, except in the high-resolution refinement stage, idealized backbone and side-chain bond lengths and angles are often used in de novo structure prediction. Hence, the energy associated with the bonded interactions terms V_{bond}, V_{angle}, and V_{torsion} can be regarded as constant. Improvement in structure prediction can conceivably be achieved by enhancing the physical models for the non-bonded terms. For example, one can replace the Van der Waals terms in Eq. 5 by a buffered 14–7 potential ***(49,50)***, by the Morse function ***(51)***,

or by the Buckingham–Fowler potential ***(52)***. The goal is to reduce the Pauli exclusion barrier so as to allow sufficient sampling of conformations in the neighborhood of the native structure during molecular mechanics or Monte Carlo simulations.

For the electrostatic term, the physical model of fixed charges at atom centers is found to be insufficient to describe charge polarization in the aqueous environment. Examples of the more sophisticated electrostatics models involve generalizing the point charge model with multi-center multi-pole expansion. This can be done through the cumulative atomic multi-pole moment method, the distributed multi-pole analysis, or an atoms-in-molecules-based multi-pole moment method ***(53–55)***. Even though these types of model improvement are computationally expensive, several groups have been making significant progress in incorporating polarizable force fields for MD simulation of proteins. For example, see **refs.** ***56–58***.

2.1.2. Protein Structures in Aqueous Environment

Protein structures are formed in the presence of aqueous environment, and therefore, in order for the search of energy-minimized protein conformation to be accurate, the effect of the solvent must be taken into account. Explicit solvent models that simulate individual water molecules [for example, TIPS ***(59,60)***, SPC ***(61)***, and F3C ***(62)***] are too slow to be practicable for protein structure prediction. Truncation of the non-bonded potentials such that interactions beyond a fixed cutoff distance are ignored can improve speed. However, it often leads to undesirable artifacts and reduced accuracy ***(63)***. Combining Ewald's approach with fast Fourier transform, Darden and his colleagues have developed the particle mesh Ewald method to describe long-range interactions more efficiently ***(64)***. However, direct simulation with explicit water is still highly computational expensive even with this and other advances. On the contrary, the effect of solvation can be modeled implicitly by averaging solvent-solute interaction using mean field formulation and by decomposing the solvation energy into an electrostatic component and a so-called non-polar component, which accounts for everything else. For electrostatics, Poisson–Boltzmann ***(65,66)*** models extend the simple Coulombic potential by allowing charge distributions within the solute and having separate dielectrics for the solvent and solute. Unfortunately, there are no general analytical solutions for the Poisson–Boltzmann equation for irregular protein shapes and precise numerical solutions (for example, by finite differences using GRASP/Delphi ***(67)***) can be very computationally expensive. Faster solutions can be obtained

using generalized-Born (GB) approximations ***(68)***, which have been incorporated into MD simulations. For the non-polar term, which includes hydrophobic interactions, the energy is usually modeled as a simple linear function of solvent accessible area. The resulting generalized-Born/surface-area (GBSA) models are more accurate than the simple non-bonded interaction terms and can rival knowledge-based functions for scoring small loops in accuracy ***(69)***. However, the amount of parameterization involved in GBSA models also rivals that of knowledge-based energies. Recently, other approximate methods for solving the Poisson–Boltzman equation may prove to be as or more accurate with less parameterization ***(70)***. Besides the Poisson–Boltzmann and generalized Born-type approaches, another category of implicit models describes the solvent effect in terms of the dielectric screening of electrostatic interaction within the protein molecule. For example, this can be done by defining the dielectric coefficient as a simple function of distance ***(71,72)*** and as a more detailed function involving solvent-excluded volume ***(73)***, the distance of a charge from the protein surface, and the degree of exposure of a charge point to the solvent ***(74)***.

In summary, the implicit solvent models are computationally much more efficient than the explicit models. The tradeoff is the inability to represent the detailed interaction structures between the solvent and the solute, which can be essential in determining the overall energy landscape. Furthermore, the lack of polarizability in the continuum solvent treatments precludes a flexible description of charge distributions in the aqueous environment.

2.2. An Overview of the Knowledge-Based Scoring Functions

The physics-based functions are formulated from underlying approximate physical models. In contrast, knowledge-based functions are derivable directly from properties observed in known folded proteins ***(75)***. Although the basis of the knowledge-based propensities is still physical, the statistical "black-box" approach to the weighting of physical effects has proved to be more effective than explicitly specifying the form and calculating the coefficients in traditional physics-based energies. As a result, almost all of the most successful de novo structure prediction techniques have both physics-based and knowledge-based components.

The hydrophobic moment ***(76)*** is an example of a simple heuristic energy function. It is analogous to the physical moment of inertia except that the mass term is replaced by a measure of the hydrophobicity of the residue. Minimization of this function leads to compact structures with hydrophobic residues in the core. In general, any property that is differentially observed in

folded proteins and unfolded proteins can be converted into an energy function. Hidden Markov models (HMM), neural nets, support vector machines (SVM), and trial and error have been used to find such properties. A particularly useful class of knowledge-based functions is the pairwise distance preferences ***(11,34,77)***, which reflect proper packing. Consequently, the pairwise distance preference scoring functions can be found in many of the top-performing de novo methods, for example, ROSETTA ***(16)***, FRAGFOLD ***(78)***, TASSER ***(79)***, CABS ***(80)***, and PROTINFO ***(81)***.

2.2.1. Deriving Knowledge-Based Scoring Functions from the Bayesian Probability Formalism

A majority of the knowledge-based scoring functions have their theoretical foundations rooted in the Bayesian (conditional) probability formalism. In such a formalism, we view a given set of conformations for a protein sequence as comprising a subset of correct conformations {C} and a subset of incorrect conformations {I}. Furthermore, we consider a set of conformational properties, which can be any feature of protein structure that differs significantly between the subset of incorrect conformations and the subset of correct conformations. Examples are the preferences of some amino acid subsequences to exhibit certain torsion angles, to form contacts with other amino acid types, and so on. In this subheading, for the purpose of illustration, we focus on the set of interatomic distances within a structure $\{d_{ab}^{ij}\}$, where d_{ab}^{ij} is the distance between atoms numbers i and j, of type a and b. We want to determine $P(C|\{d_{ab}^{ij}\})$, the probability that the structure is a member of the "correct" subset, given that it contains the distances $\{d_{ab}^{ij}\}$. A standard way to achieve this is to express $P(C|\{d_{ab}^{ij}\})$ in terms of probabilities derivable from experimental structures, through the Bayes' theorem:

$$P(C|\{d_{ab}^{ij}\}) = P(C) \times \frac{P(\{d_{ab}^{ij}\}|C)}{P(\{d_{ab}^{ij}\})} \tag{6}$$

Here, $P(\{d_{ab}^{ij}\}|C)$ is the probability of observing the set of distances $\{d_{ab}^{ij}\}$ in a correct structure. $P(\{d_{ab}^{ij}\})$ is the probability of observing such a set of distances in any correct or incorrect structure, and $P(C)$ is the probability that any structure picked at random belongs to the correct subset. $P(\{d_{ab}^{ij}\}|C)$ is regarded as a posterior probability in the sense that the underlying population for the probability distribution consists of structures that are already known to belong to the "correct" subset. On the contrary, $P(\{d_{ab}^{ij}\})$ is regarded as a prior probability in the sense that its underlying population is composed of

structures whose class memberships have not yet been determined. We should note that both $P(\{d_{ab}^{ij}\}|C)$ and $P(\{d_{ab}^{ij}\})$ are highly difficult to compute, because the input arguments to these probability functions are the multitude of distance variables. A full model capturing the dependency among these variables would be extremely complex and would require a huge amount of training data to determine all the implicit parameters. Hence, to ensure computational feasibility of Eq. 6, one often makes the simplifying, albeit not strictly correct, assumption that the distances are statistically independent of one another, that is:

$$P(\{d_{ab}^{ij}\}|C) = \prod_{i,j} P(d_{ab}^{ij}|C); P(\{d_{ab}^{ij}\}) = \prod_{i,j} P(d_{ab}^{ij}) \tag{7}$$

Then, combining Eqs. 6 and 7 gives us

$$P(C|\{d_{ab}^{ij}\}) = P(C) \prod_{i,j} \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \tag{8}$$

For a given protein sequence, $P(C)$ is a constant independent of conformation and therefore can be omitted because we are only interested in selecting native-like conformations among decoys for a fixed protein sequence. Equation 8 suggests a scoring function S, which is proportional to the negative log conditional probability that the given structure is correct, given a set of distances.

$$S(\{d_{ab}^{ij}\}) = \sum_{i,j} s(d_{ab}^{ij}); s(d_{ab}^{ij}) = -\log\left(\frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})}\right) \tag{9}$$

An advantage of using Eq. 9 instead of Eq. 8 as a scoring function is that in the logarithm form, the pitfall of repeated multiplication of small numbers is eliminated, and therefore, it is easier to be implemented on the computer.

One can replace the set of distances $\{d_{ab}^{ij}\}$ with another type of conformational property, say for example $\{m_a^i\}$, where m_a^i represents the value of that conformational property attained by residue number i of amino acid type a. This leads to another scoring function:

$$S(\{m_k\}) = -\sum_k \log\left(\frac{P(m_k|C)}{P(m_k)}\right) \tag{10}$$

To gain an intuitive understanding of the scoring function, we note that if the chosen conformational property does not differ significantly between the subset of incorrect conformations and the subset of correct conformations, then the

values of $P(m_k|C)$ and $P(m_k)$ will tend to be close to each other. The resulting score S will always be close to 0 and is not an informative measure for decoy discrimination. On the contrary, if the conformational property is well chosen, that is, it differs significantly between incorrect and correct conformations, then for a native-like structure, $P(m_k|C)$ will tend to dominate $P(m_k)$, yielding a negative (good) score for S. On the contrary, for a non-native structure, the opposite occurs, yielding a positive (bad) score.

2.2.2. Compilation of the Probabilities

Before one can use Eq. 9 as a scoring function, the statistics for the posterior probability $P(d_{ab}^{ij}|C)$ and the prior probability $P(d_{ab}^{ij})$ need to be compiled. To compile the statistics for $P(d_{ab}^{ij}|C)$, we can tabulate the intra-molecular distances observed in a database of experimentally determined conformations. Such a database is usually extracted from the Protein Data Bank (PDB) ***(82,83)***. For example, one can proceed to select all the proteins from the PDB that also appear in the e-value filtered ASTRAL SCOP genetic domain sequence subset list with the threshold e-value set at 10^{-4} ***(84)***. Such an e-value is chosen, so that sampling bias (i.e., including too many homologous proteins) can be avoided. We then evaluate the quantity

$$P(d_{ab}^{ij}|C) \equiv \frac{N(d_{ab})}{\sum_d N(d_{ab})} \tag{11}$$

where $N(d_{ab})$ is the number of occurrences of atom types a and b in a distance bin d in the database.

To compile the statistics of the prior probability $P(d_{ab}^{ij})$, we apply a formula similar to Eq. 11. But the question is: What would be an appropriate database from which to tabulate the counts? Samudrala and Moult ***(34)*** argued that methods employed for structure prediction usually produce compact models, whether the result is topologically correct or not. Thus, they consider a good choice of prior distribution to be found in the set of possible compact conformations and assume that averaging over different atom types in experimental conformations is an adequate representation of random arrangements of these atom types in any compact conformation. The probability $P(d_{ab})$ of finding atom types a and b in a distance bin d in any native-like or non-native compact conformation is thus approximated by:

$$P(d_{ab}) = \frac{\sum_{ab} N(d_{ab})}{\sum_d \sum_{ab} N(d_{ab})} \tag{12}$$

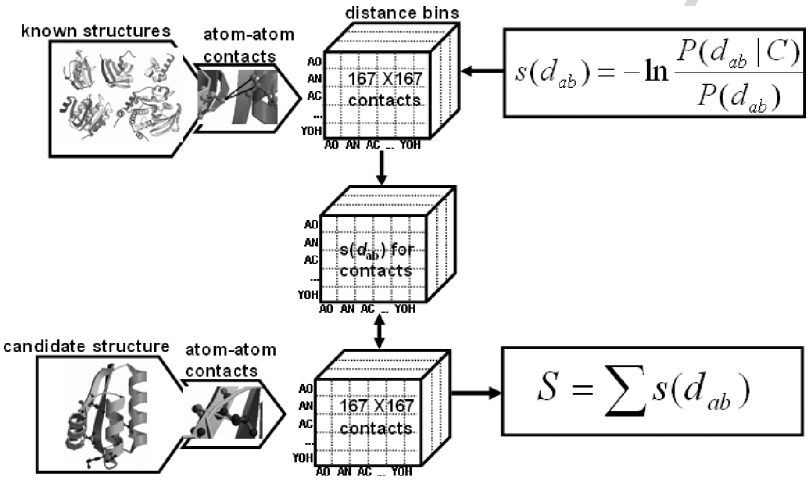
where $\sum_{ab} N(d_{ab})$ is the total number of contacts between all pairs of atom types in a particular distance bin d, and the denominator is the total number of contacts between all pairs of atom types summed over the distance bins d. The pairwise distance preference function described in **Subheading 2.2.1.**, Eq. 9, together with Eq. 11 and the prior distribution assumption of Eq. 12, is termed the RAPDF in ***(34)***. **Figure 2A** highlights the essential components of this scoring function.

Besides the above method of estimating prior distributions, various other approaches have also been suggested. Subramaniam et al. ***(85)*** assumed that all distances are equally probable, and Avbelj and Moult ***(86)*** considered the set of distances observed in some random coil model as appropriate. Lu and Skolnick ***(87)*** employed a quasi-chemical approximation. Alternatively, Zhou and Zhou ***(88)*** assumed that the residues follow uniform distribution everywhere in the protein and developed a new reference state termed "distance-scaled, finite ideal-gas reference state."

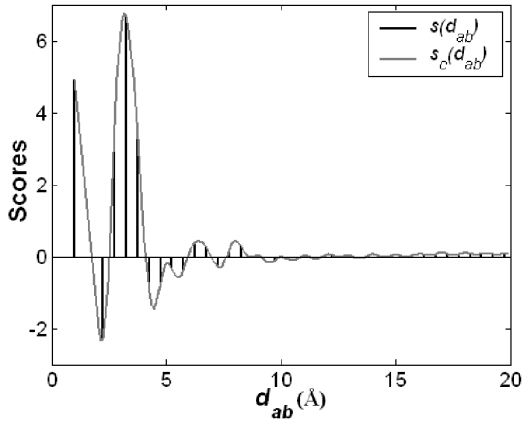
2.2.3. A Pairwise Distance Scoring Function in Continuous Form

The RAPDF scoring function uses discrete distance bins to compile the probability scores. Specifically, contact distances between 0 and 3 Å are grouped into bin 1, 3 and 4 Å into bin 2, 4 and 5 Å into bin 3, and so on up to the 20 Å cutoff. As a result, the score for observing any distance within a bin width is the same for a given pair of atom types. However, the distance preferences between atom types should vary in a continuous manner as the distances between the contacts vary. We can seek a function to interpolate between the scores across the discrete bins such that the score for a given distance can be uniquely defined. Several methods for interpolating discrete points, including linear, polynomial, cubic spline, and band-limited interpolations, have been tested for their efficacy to improve the discriminatory power of RAPDF. The best among the tested methods is band-limited interpolation, derivable from the Fourier Theorems. It assumes that the variation of the log-likelihood scores fluctuates slowly enough such that the scores for any given distance can be exactly reconstructed from the scores across the discrete bins.

Given a pair of atom types a and b at a particular distance, a "continuous" log-likelihood score $s_c(d_{ab})$ can be calculated by interpolating between the scores across the discrete bins of $s(d_{ab})$ through the Shannon's sampling theorem, resulting in a smooth curve ***(89)***. (*see* **Fig. 2B** for illustration.) Given an amino acid sequence in a particular conformation, $s_c(d_{ab})$ of all contacts between pairs of atom types at any distance within the 20 Å cutoff is summed to yield the total



(a)



(b)

Fig. 2. The all-atom distance-dependent conditional probability discriminatory function (RAPDF) and its extension, the interpolated RAPDF function. **(A)** The essential feature of the RAPDF scoring function. A matrix giving the log-likelihood scores for pairwise contact among different atom types at various discrete distance bins is computed using a database of known experimental structures. Then, given a candidate ("decoy") structure, appropriate entries in the matrix can be extracted and summed to give a log-likelihood score for the structure. **(B)** The application of band-limited

log-likelihood score to evaluate whether the conformation is native-like or not. The interpolated RAPDF (IRAPDF) has been evaluated by various decoy sets. Comparison between the IRAPDF and the RAPDF shows that the band-limited interpolation leads to an improved discriminatory power.

2.3. Neural Network Knowledge-Based Scoring Functions

Rather than predicting whether an entire structure is native-like or not, neural network algorithms are often used to predict the likelihood of occurrence of a certain conformational property for each residue along a given protein sequence. Examples of the properties are the tendencies of an amino acid to be exposed or buried relative to the solvent ***(90–92)***, to be part of the helix, strand, or coil local structures ***(93–95)***, the expected number of contacts a residue makes with other residues ***(96–99)***, and so on. Usually, the conformational property of interest is discretized into a number of states, and a neural network algorithm returns numerical values which correlate with the probabilities of occurrences of those states.

One can combine the neural network algorithms for predicting conformational properties with the Bayesian probability formalism that has been used to construct various knowledge-based functions. This leads to a class of scoring functions that give log-odd scores, indicating whether a given structure is native-like or not, and that have in their core a neural network component. In the following subheadings, we review a standard formulation of the neural network algorithm that is used to predict conformational properties of residues in a protein sequence. We then describe how the neural network and the Bayesian frameworks are combined to form several neural network-based scoring functions.

2.3.1. Neural Network Algorithms for Predicting Local Structures

For concreteness, we consider the prediction of the degree of solvent accessibility of individual residues along a given protein sequence, with the degree discretized into three states: low, medium, and high. The now standard approach, introduced in **ref. *93*** and improved upon in **ref. *94***, uses a feed-forward neural network. The input to the network is a window of sequence

Fig. 2. interpolation to the discrete distance bins of the RAPDF function. The score $s_c(d_{ab})$ of a given pair of atom types at any distance within the 20 Å cutoff can be uniquely defined by interpolating across the discrete bins of $s(d_{ab})$. The resulting scoring function is termed as the interpolated RAPDF (IRAPDF).

profile corresponding to a consecutive sequence of residues. Such a windowed sequence profile can be obtained by following a procedure described in **ref. *94***. The protein sequence of interest is employed as input to PSI-BLAST ***(100)***, which generates a position-specific scoring matrix (PSSM) associated with that sequence. The PSSM consists of $20 \times M$ entries, where M being the length of the sequence, and each entry in a column gives the log-likelihood for one of the twenty possible amino acid substitutions for the residue position of interest. The standard logistic transform is then applied to each entry of the PSSM, so that these values are rescaled to the 0–1 range, appropriate to serve as neural network inputs. The neural network itself can consist of one or more hidden layers, and its output layer comprises three output units, representing the low, medium, and high solvent accessibility states, respectively. Training of the network is done with back-propagation ***(101)***, using the database of experimentally determined protein structures we have already described in **Subheading 2.2.2.** Given a window of sequence profile of the residue of interest (i.e., the sequence profile of the residue as well as those of the neighboring residues), the resulting neural network returns a numerical value in each output unit correlating with the probability with which the residue assumes the corresponding state.

2.3.2. Combining the Neural Network Algorithms with the Bayesian Probability Formalism

To describe how one combines the Bayesian and the neural network frameworks to construct new scoring functions, for concreteness, suppose once again that the conformational property of interest is the degree of solvent accessibility. Using the language of the preceding subheadings, we want to calculate the probability that a given structure belongs to the subset of correct structures, given the associated conformational string $\{q_a^i\}$. Here, $q_a^i \in \{l, m, h\}$, where l represents low solvent accessibility state, m medium, and h high, i is the residue number, and a is the amino acid type. A scoring function described in Eq. 10 now takes the following form:

$$S(\{q_a^i\}) = -\sum_i \log\left[\frac{P(q_a^i|C)}{P(q_a^i)}\right] \tag{13}$$

$P(q_a^i|C)$ is simply the (posterior) probability of residue i taking on a particular solvent accessibility state q_a^i in a native structure. With an additional processing step involving the nearest-neighbor approach of Yi and Lander ***(102)*** to be discussed in detail in the next subheading, this probability can be estimated by using the neural network algorithm previously described. $P(q_a^i)$, on the contrary, is the (prior) probability that the residue is observed to assume the

solvent accessibility state q_a^i in any native-like or non-native structure. It can be estimated using the formula

$$P(q_a) \equiv \frac{N(q_a)}{\sum_{q \in \{l,m,h\}} N(q_a)} \tag{14}$$

where $N(q_a)$ is the number of occurrences of the amino acid type a taking on the solvent accessibility state q in some database of structures, and $\sum_{q \in \{l,m,h\}} N(q_a)$ is the total number of occurrences of the amino acid type a in that database. Again, the question is: What is an appropriate database from which to tabulate the counts? We can use the same approach adopted by Samudrala and Moult in **ref. *34***, arguing that the set of possible compact conformations is a good choice of prior distribution. Then, the database to use will simply be the database of the experimentally determined structures. Alternatively, we can employ a database of decoy structures. Such a database can be created by applying a de novo conformational space sampling protocol to generate n decoy structures (for example, $n = 10$) for each protein sequence that appears in the database of the experimentally determined structures and then gathering the resulting decoys.

We note that as $P(q_a^i|C)$ is estimated by the neural network algorithm with a window of sequence profile as its input, the influence of the neighbors of residue i on its conformation is automatically taken into account. Thus, the posterior probability that residue i assumes a particular conformation is calculated in the context of its surrounding environment. In contrast, the probability distribution $P(q_a)$ is compiled on a "single-residue" basis. Thus, $P(q_a)$ can be viewed as the tendency of the amino acid type a to adopt a certain conformation averaged over the various types of neighborhood environments.

For further illustration, we generate a neural network-based Bayesian scoring function for each of the following conformational properties: the virtual torsion angle, the virtual bending angle, and the degree of solvent accessibility. The virtual torsion angle and the virtual bending angle are calculated by the DSSP program ***(103)***. Specifically, given a residue i of interest, the virtual torsion angle for i is the dihedral angle defined by the C_α atoms of residues $i-1$, i, $i+1$, and $i+2$. The virtual bending angle is the bending angle defined by the C_α atoms of residues $i-2$, i, and $i+2$. Solvent accessibility is the residue water exposed surface in $Å^2$. To implement the scoring functions, the virtual torsion angle are manually divided into two discrete states, whereas the virtual bending angle and the degree of solvent exposure are each manually divided into three discrete states.

2.3.3. Training and Post-Processing of the Neural Network

The Stuttgart Neural Network Simulator ***(104)*** is a versatile and convenient tool to configure and train the neural networks for predicting the various conformational properties. The network configurations follow the description given in **Subheading 2.3.1.** The input layer receives a window of sequence profile. The window size typically ranges from 1 to 17 consecutive residues. The network has a single hidden layer and an output layer of two or three units representing two or three discrete states. See **Fig. 3** for an illustration.

We divide the database of experimentally determined structures into two equal subsets A and B, which are alternately used as the training and the test sets. The neural network training is done in batch mode using standard back-propagation, and the cycle of batch-mode training is repeated until the test error reaches a minimum. We note that two neural networks are obtained at the conclusion of the training—one (denoted as NN_A) trained with subset A and tested with subset B and another one (denoted as NN_B) trained with subset B and tested with subset A.

Given a residue of interest together with its windowed sequence profile, it is desired to extract from NN_A and NN_B the posterior probabilities with which the residue assumes each of the three states, say in the case of solvent accessibility prediction (two states in the case of virtual torsion angle prediction and three states in the case of virtual bending angle prediction). To this end, the nearest-neighbor approach of Yi and Lander ***(102)*** is employed: The output layer of NN_A gives a 3-tuple vector (s_{lA}, s_{mA}, s_{hA}). The closeness of this vector with respect to vectors corresponding to all instances in the test set can be calculated through the Euclidean measure

$$\left((s_{lA} - s^g_{lA})^2 + (s_{mA} - s^g_{mA})^2 + (s_{hA} - s^g_{hA})^2\right)^{1/2} \tag{15}$$

where g stands for instance g in the test set. The k-nearest neighbors [e.g., the closest 5% of all instances in the test set with respect to (s_{lA}, s_{mA}, s_{hA})] are then determined, and the actual solvent accessibility states of those nearest neighbors are tabulated, yielding the counts (c_{lA}, c_{mA}, c_{hA}). The same procedure is repeated with NN_B. The probability that the residue of interest takes on each of the three states is thus estimated by

$$P(s_q) = \frac{c_{qA} + c_{qB}}{\sum\limits_{r \in \{l,m,h\}} c_{rA} + c_{rB}} \tag{16}$$

where q stands for low, medium, or high accessibility state. Equation 16 supplies the posterior probabilities required in Eq. 13 for score calculation.

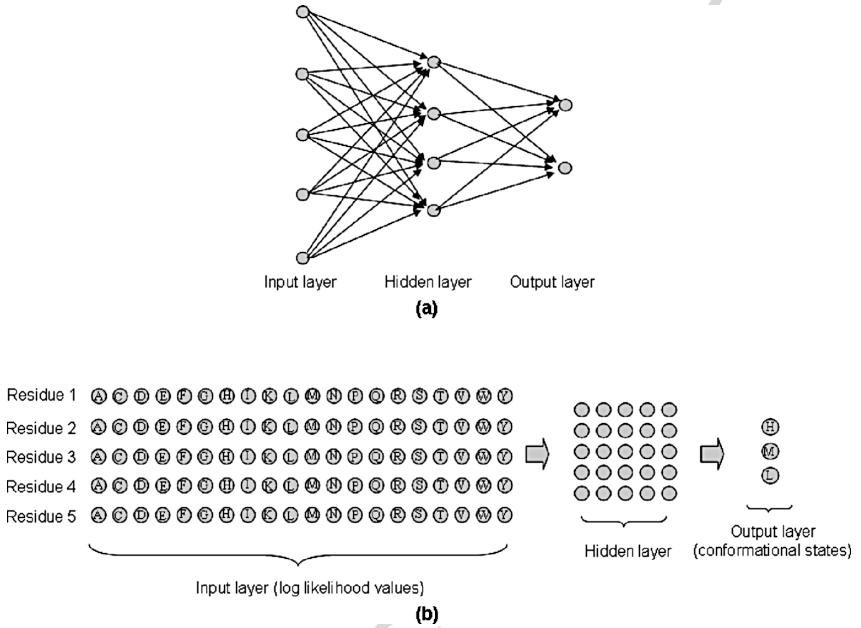


Fig. 3. Schematic diagrams of the neural networks used to predict conformational property given a sequence profile. **(A)** A fully connected neural network with input (5 units), hidden (4 units), and output (2 units) layers. Every unit in the input layers is connected with every unit in the hidden layers. The same holds true for the hidden and the output layers. **(B)** The typical size of a neural network we use for constructing the knowledge-based functions. In this example, the window size of the input sequence profile is five residues. Each residue provides twenty input units, representing the log-likelihood values for the twenty possible amino acid substitutions for that residue position. The hidden layer consists of 25 units. The output layer has three units. In the case of solvent accessibility prediction, these output units correspond to low, medium, and high solvent accessibility states, respectively. The input and the hidden layers, and the hidden and the output layers, are fully connected as in **(A)**, but for simplicity, the connections are not shown.

2.3.4. Decoy Sets and Evaluation of the Knowledge-Based Scoring Functions

One evaluates the usefulness of a scoring function by examining the ability of the scoring function to distinguish native-like conformations from non-native ones. This is achieved through generating test decoy sets and testing

the performance of the function on those sets. There are various approaches to generate test decoys. For example, they can be created by sampling discrete-state models starting from a native conformation ***(105)***, having amino acid sequences with known folds mounted onto different folds ***(106,107)***, and using crystal structures of various resolutions ***(85)***. Databases of test decoy sets have been created to enable the evaluation of scoring functions on multiple types of decoys ***(108–110)***. An approach most relevant to evaluating scoring functions for de novo structure prediction is to create test decoys through de novo conformational space sampling. A typical de novo conformational space sampling protocol consists of an MCSA search procedure guided by a set of energy functions, with move set based on lattice models ***(111,112)***, fragment substitution ***(113,114)***, or continuous torsional distributions ***(81)***.

There are several commonly used measures for evaluating the usefulness of scoring functions. The $\log P_{B1}$ measure is the log probability of selecting the lowest C_α root mean square deviation (RMSD) conformation in a test decoy set, calculated with the formula

$$\log P_{B1} = \log_{10}\left(\frac{R_i}{n}\right) \tag{17}$$

Here, R_i is the C_α RMSD rank of the best scoring conformation in the test set of n decoys. The $\log P_{B10}$ measure is the log probability of selecting the lowest C_α RMSD conformation among the top-10 best-scoring conformations, that is, instead of using the RMSD rank of the best-scoring conformation, the best RMSD rank achieved among the top-10 best-scoring conformations is used as R_i in Eq. 17. The CC measure is the correlation coefficient between the C_α RMSDs and the scores generated by the scoring function. The enrichment ratio measure is the fraction enrichment of the top 10% lowest RMSD conformations in the top 10% best scoring conformations. Specifically, after a scoring function is applied to a test decoy set, we count the number of decoys (denoted as a), which are in the top 10% in terms of both their scores and their C_α RMSDs relative to the native structure. The expected number in a random distribution is 10% × 10% × (number of decoys in the set) (denoted as b). The enrichment ratio is a/b. A value above 1 indicates enrichment over the random distribution. The four evaluation measures are illustrated in an example in **Fig. 4**.

To examine the utility of the knowledge-based scoring functions in decoy discrimination, we apply both the RAPDF and the neural network-based functions to 41 test decoy sets of varying quality generated with de novo conformational space sampling. Each decoy set contains approximately 10,000 decoy conformations. Table 1 summarizes the PDB identifiers and the SCOP

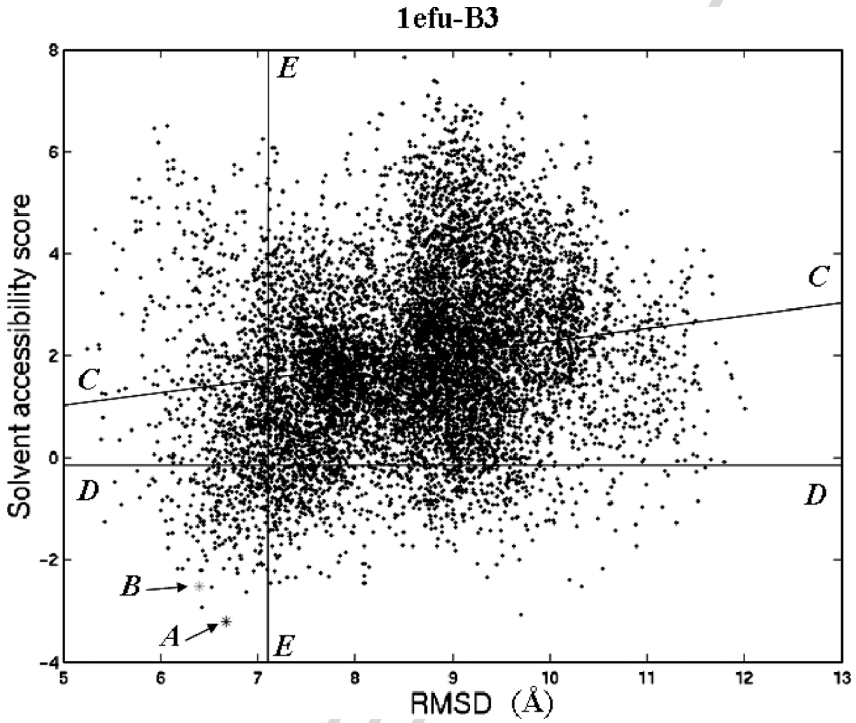


Fig. 4. Measures for evaluating scoring functions. Log P_{B1} is the log probability of selecting the lowest C_α RMSD conformation in a test decoy set (point *A*), which is -1.42 in this example. Log P_{B10} is the log probability of selecting the lowest C_α RMSD conformation among the top 10 best-scoring conformations in a test decoy set (point *B*), which is -1.76 in this example. The correlation coefficient between the C_α RMSDs and the scores is equal to the slope of line *C*-*C* and has the value of 0.25 in the present case. Line *D*-*D* represents the top 10% score cutoff for the decoy set. By counting the number of decoys below this line, which are also within the top 10% RMSD cutoff (left of line *E*-*E*), and dividing this number by the expected value for a random distribution, an enrichment ratio of 2.7 is obtained. Different measures are needed dependent on the specific purposes and roles of the scoring functions.

classifications of the 41 protein sequences used in generating the test decoy sets. Also included is the C_α RMSD of the best decoy relative to the corresponding native structure in each test set. Among them, fifteen test decoy sets have their best structures below 6 Å C_α RMSD relative to their native conformations.

Table 1
List of the Protein Sequences Used in Generating the Test Decoy Sets

Protein	SCOP classifications	Length	Minimum RMSD
1b0n-A2	a.35.1.3 (A:1–68)	68	2.729
1b33-N	d.30.1.1 (N:)	67	7.349
1b34-A	b.38.1.1 (A:)	80	7.943
1b4b-A	d.74.2.1 (A:)	71	5.506
1b79-A	a.81.1.1 (A:)	102	5.29
1ck9-A	d.79.3.1 (A:)	104	7.661
1ctf	d.45.1.1 (–)	68	4.37
1dgn-A	a.77.1.1 (A:)	89	4.482
1dj8-A	a.57.1.1 (A:)	79	5.092
1dtj-A	d.51.1.1 (A:)	74	4.902
1e68-A	a.64.2.1 (A:)	70	3.794
1eai-C	g.22.1.1 (C:)	61	6.914
1edz-A2	c.58.1.2 (A:3–148)	146	9.277
1efu-B3	a.5.2.2 (B:1–54)	54	5.247
1ev0-A	d.71.1.1 (A:)	58	6.641
1f53-A	b.11.1.4 (A:)	84	9.123
1fc3-A	a.4.6.3 (A:)	119	8.184
1fmt-A1	b.46.1.1 (A:207–314)	108	7.385
1g6e-A	b.11.1.6 (A:)	87	7.891
1g7d-A	a.71.1.1 (A:)	106	5.867
1goi-A1	b.72.2.1 (A:447–498)	52	6.111
1gut-A	b.40.6.1 (A:)	67	6.459
1h5p-A	b.99.1.1 (A:)	95	8.223
1h8a-C1	a.4.1.3 (C:87–143)	57	2.941
1ijy-A	a.141.1.1 (A:)	122	7.916
1ira-Y1	b.1.1.4 (Y:1–101)	101	8.317
1iwg-A1	d.58.44.1 (A:38–134)	97	5.7
1jju-A3	b.1.18.14 (A:274–351)	78	6.614
1jos-A	d.52.7.1 (A:)	100	5.302
1jyg-A	a.60.11.1 (A:)	69	3.471
1k2y-X2	c.84.1.1 (X:155–258)	104	6.889
1ktz-B	g.7.1.3 (B:)	106	8.586
1l9l-A	a.64.1.1 (A:)	74	4.041
1msp-A	b.1.11.2 (A:)	124	9.932
1n69-A	a.64.1.3 (A:)	78	6.753
1qu6-A1	d.50.1.1 (A:1–90)	90	8.597
1rie	b.33.1.1 (–)	127	9.548
1sra	a.39.1.3 (–)	151	8.781

1sro	b.40.4.5 (–)	76	6.031
2igd	d.15.7.1 (–)	61	6.508
7gat-A	g.39.1.1 (A:)	66	7.248

Each row lists the Protein Data Bank (PDB) identifier of the sequence, the SCOP classification, the length of the protein sequence, and the C_α RMSD of the best decoy structure relative to the native conformation in the test decoy set. Each test decoy set contains $\sim$ 10, 000 decoys. Fifteen test decoy sets have their best structures below 6 Å C_α RMSD relative to their corresponding native conformations. Twenty-four test decoy sets have their best structures below 7 Å C_α RMSD relative to their corresponding native conformations.

Twenty-four decoy sets have their best structures below 7 Å C_α RMSD relative to their native conformations, and so on. For illustration purpose, we employ the enrichment ratio measure to evaluate the scoring functions. The results are displayed in **Fig. 5**. From the figure, we observe that the RAPDF function gives uniform performance for decoy discrimination across decoy sets of different quality, whereas the neural network-based scoring functions tend to perform better for decoy sets with better quality.

2.4. Some Other Knowledge-Based Scoring Functions in the Recent Literature

In the formulation of the RAPDF scoring function as well as of the other pairwise distance preference functions described in **refs.** ***11,77,87*** and ***(88)***, the solvation effect is not explicitly modeled. However, as we have previously discussed, as protein folding occurs in the aqueous environment, a careful accounting of the solvent effect is important in determining the native conformation. In this regard, McConkey et al. ***(115)*** quantify contact surfaces of atoms by integrating the solvent accessible surface and the inter-atomic contacts into one quantity and construct an all-atom contact potential based on the contact preferences of 167 residue-specific atom types with 168 possible contact types (167 possible atom contact types and one solvent contact). They demonstrate that this all-atom contact potential delivers satisfactory performance for distinguishing native conformations from decoy structures.

Another possible approach to augment the pairwise distance preference scoring functions is by considering various multi-body geometric properties. In **ref.** ***116***, a four-body SNAPP potential involving the tiling of protein structures with tetrahedra having the center of mass of each amino acid side-chain at each vertex is introduced. This formulation results in 8855 possible tetrahedron types with the corresponding log-likelihoods computed from structural databases. It is found that the SNAPP potential is accurate in predicting the

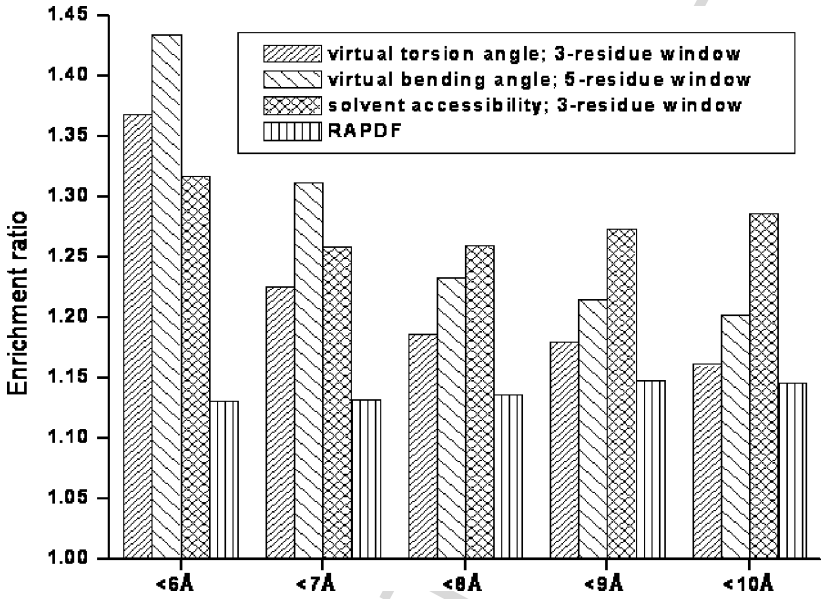


Fig. 5. Performances of the various knowledge-based scoring functions. The functions are evaluated using the average enrichment ratios on test decoy sets of varying quality. For example, the first four bars indicates the average enrichment ratios attained by the individual functions for the test decoy sets that contain structures of less than 6 Å C_α RMSD relative to the native conformations. The following scoring functions are examined in the figure: a neural network-based virtual torsion angle scoring function with a three-residue window; a neural network-based virtual bending angle scoring function with a five-residue window; a neural network-based solvent accessibility scoring function with a three-residue window; and the all-atom distance-dependent conditional probability function.

effects of hydrophobic core mutations. A similar four-body scoring function derived through the Delauney tessellation of side-chain centroids of amino acids is shown to be able to distinguish native conformation from partially unfolded and deliberately misfolded structures ***(117)***. On the basis of the work of Professor Banavar and his colleagues, Ngan et al. ***(118)*** construct a three-body knowledge-based potential involving the radii of curvature formed among triplets of residues in protein conformations. The resulting residue-triplet function is shown to be of utility in discriminating native-like conformations from non-native structures. Finally, Li et al. ***(119)*** introduce a knowledge-based

AQ1

scoring function based on the edge simplices from the alpha shape of the protein structure. Formally, their statistical alpha contact potential is a two-body scoring function, and their definition of contact is when atoms from non-bonded residues share a Voronoi edge, with the edge at least partially contained in the body of the protein. This formulation has the benefit of avoiding spurious contact between two residues when a third residue is between them. The authors have shown that the alpha contact potential performs comparably with other atom-based potentials, while requiring fewer parameters.

In summary, the construction of a knowledge-based scoring function involves the following steps: (1) selection of a conformational property that differs between native-like and non-native structures; (2) compilation of the posterior probability distributions of this conformational property by direct counting or through statistical techniques such as neural network, based on a database of experimentally determined structures; (3) derivation of the prior probability distributions based on a database of decoy structures or through simplifying assumptions such as the averaging-over-atom-types argument of Samudrala and Moult ***(34)***, the quasi-chemical approximation of Lu and Skolnick ***(87)***, or the uniform distribution argument of Zhou and Zhou ***(88)***; and (4) formation of the log-odd scores from the prior and posterior probabilities. Step 1 is perhaps the most critical step and is largely dependent on one's insights into the physical and chemical processes involved in protein folding and by trial and error. In step 2, the selection of appropriate statistical techniques is heavily influenced by the size and quality of the available data set, because these factors have a direct impact on determining whether certain statistical assumptions (e.g., the conditional independence assumption in Eq. 7) are needed.

2.5. The Design of Decoy Filters

As we have discussed, conformational search algorithms produce a multitude of candidate conformations. Various scoring functions can be combined into a filter to distill this vast collection of decoys, to retain those that are native-like. An approach to constructing such a filter is to assign weights to the different scoring functions, such that the resulting linear combination of the scores gives the overall quantitative assessment of a decoy structure of interest. The weights used in the linear combination can be derived by performing logistic regression on test decoy sets. Specifically, native-like decoys (determined by a suitably chosen C_α RMSD cutoff) in each test set are labeled as belonging to class 1, and the rest labeled as class 0. The normalized scores for an individual decoy become the independent variables (x_j; $j = 1 \ldots k$; $k =$ the total number of score

types), whereas its associated class label forms the dependent variable (p), which are then used to fit the following equation to obtain the weights w_js:

$$\log\left(\frac{p}{1-p}\right) = \alpha + w_1 x_{1,i} + \ldots + w_k x_{k,i} \quad (18)$$

Here, α is a constant representing the intercept. i ranges from 1 to N, and N is the total number of decoys. Normalization of a scoring function can be achieved by subtracting its mean and dividing by its standard deviation, where the mean and the standard deviation are computed over all decoys within a test set, or by replacing the raw score of a decoy with its rank and then dividing by the total number of decoys in the test set. Techniques such as leave-one-out cross-validation and forward and backward stepwise regression can be applied to determine which independent variables are helpful in assessing the accuracy of a given decoy structure and which can be discarded. Essentially, functions describing useful orthogonal characteristics of protein native conformations will receive large weights, whereas those that are less useful or containing overlapping information will have smaller or zero weights. Finally, alternative approach to performing logistic regression is also possible, for example, by replacing it with machine-learning techniques such as the neural network or SVM. The decision is again influenced by the size and quality of the available test data.

2.6. Further Enhancement of Decoy Selection Through Conformer Clustering and High-Resolution Refinement

Conformer clustering and high-resolution refinement are often used as additional steps in the decoy selection process to further refine the set of native-like conformations retained by the decoy filter. The idea of conformer clustering is based on the following observation: Conformers with correct folds are in general similar to other conformers with correct folds. On the contrary, it is unlikely that multiple conformers share the same mistake, and therefore, conformers with incorrect folds are in general dissimilar to each other as well as to conformers with correct folds. Hence, the conformers that are most similar to the others, that is, those at the cluster centers of the conformational distribution, will tend to be the correct ones. Various metrics are used to describe the conformational distribution, including pairwise RMSD, pairwise RMSD with cutoffs, and number of neighbors ***(16,120)***. Heuristic schemes such as k-mean clustering, visual inspection following dimensionality reduction, and iterative sampling ***(121)*** can be used to locate these cluster centers.

Figure 6 illustrates the performance of a conformer-clustering algorithm [the density score function available in the RAMP package ***(122)***] in distinguishing

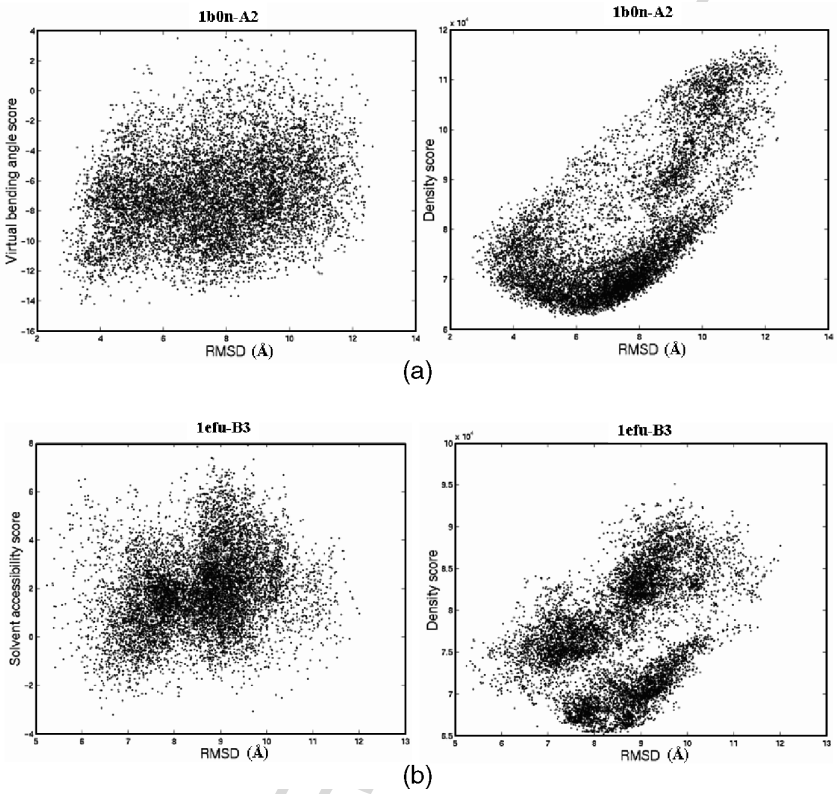


Fig. 6. The comparison of some knowledge-based scoring functions and the density score function in discriminating decoys. In **(A)**, the virtual bending angle scoring function is compared to the density score function, whereas in **(B)**, the solvent accessibility scoring function is compared to the density score function. The diagrams show that the density score function produces improved correlation between the C_α RMSDs and the scores in both cases, suggesting that conformer clustering is useful as a complementary step in decoy selection.

native-like structures from non-native conformations. Compared with the neural network-based virtual bending angle and solvent accessibility scoring functions, the density score function produces results that show improved correlation between the C_α RMSDs and the generated scores. This observation suggests that applying conformer clustering in addition to using scoring functions as filter can enhance the overall ability to select native-like structures from decoys.

The goal of high-resolution refinement is to further optimize the remaining candidate structures that have passed through the decoy filtering and conformer clustering stages. The optimization is carried out by making small perturbations to a candidate structure guided by a highly detailed energy potential. One of the most notable methods is that of Misura et al., which has been shown to be effective in the Sixth Critical Assessment of Techniques for Protein Structure Prediction (CASP-6) ***(123,124)***. It involves applying perturbations to backbone and side-chain torsion angles using an all-atom force field. The force field consists of a standard 6–12 Lennard–Jones potential for Van der Waals packing, the implicit solvation model of Lazaridis and Karplus describing dielectric screening ***(73)***, and a new orientation-dependent hydrogen bonding term ***(125)***. The hydrogen-bonding term is derived based on observed geometrical parameters of hydrogen bonds in high-resolution crystal structures of proteins. Using this combined physics-based and knowledge-based function as part of their prediction protocol, Bradley et al. have reported success in high-resolution structure prediction of less than 1.5 Å for protein domain of less than 85 residues ***(124)***.

A summary of the scoring functions discussed in this chapter can be found in Table 2. We should note that there are other means to guide conformational search and decoy filtering besides using scoring functions. For example, filtering schemes based on contact order ***(126)*** and beta sheet topology ***(127)*** have been found to be beneficial in enriching the ensemble quality of decoy structures.

3. Discussion and Conclusion

A main objective of the structural genomic initiatives, spurred by large-scale genome sequencing efforts, is to determine as many protein folds as possible. The need to determine protein structures rapidly and inexpensively in turn leads to an increased interest in computational protein structure prediction, the two main approaches of which being homology modeling and de novo structure prediction.

The key components in de novo protein structure prediction are conformational space sampling and decoy selection. Scoring functions are employed in both the conformational sampling stage and the decoy selection stage. In the first stage, a selected combination of scoring functions approximates the energy landscape of the conformational space, and conformational search algorithms generate trajectories leading to the landscape minima, whereas in the second stage, another set of possibly different scoring functions are used as filter to

Table 2
A list of the scoring functions discussed in Section 2

Scoring function	Subheading	Usage	Description
Class I force field	2.1.1.	Conformational space search	Physics-based force field modeling bonded and non-bonded interactions among atoms
RAPDF	2.2.1.	Conformational space search/decoy filtering	Knowledge-based potential describing atom–atom distance preferences
IRAPDF	2.2.3.	Conformational space search/decoy filtering	Continuous version of the RAPDF function
Neural network knowledge-based functions	2.3.	Conformational space search/decoy filtering	Incorporation of neural network into the Bayesian probability framework to describe various conformational properties
Atom–atom contact scoring function	2.4.	Conformational space search/decoy filtering	Knowledge-based atom–atom contact preference function taking solvent accessibility into account
SNAPP potential	2.4.	Conformational space search/decoy filtering	A four-body knowledge-based function describing tiling of protein structures with tetrahedra
Four-body contact scoring function	2.4.	Conformational space search/decoy filtering	A four-body knowledge-based function based on Delauney tessellation of side chain
Residue triplet scoring function	2.4.	Conformational space search/decoy filtering	A three-body knowledge-based function based on the radii of curvature formed among triplets of residues

(Continued)

Table 2
(Continued)

Scoring function	Subheading	Usage	Description
Alpha contact potential	2.4.	Conformational space search/decoy filtering	A two-body knowledge-based function based on edge simplices from the alpha shape of the protein structure
Structure refinement potential of Misura et al.	2.6.	High-resolution refinement	A combined physics- and knowledge-based function modeling Van der Waals interaction, solvent effects, and hydrogen bonding

Each row gives the name of the scoring function, the subheading in which it is discussed, its usage, and a brief description of its components.

retain a collection of the native-like structures. Conformer clustering and high-resolution refinement can also be used as additional steps to further refine this collection. In this chapter, we have studied some examples of the physics-based and knowledge-based scoring functions. For the physics-based approach, the Class I force field and its extensions as well as solvation modeling were discussed. For the knowledge-based approach, we studied the Bayesian probability formalism and used it to derive the RAPDF ***(34)***. In addition, we detailed the construction of the neural network-based Bayesian scoring functions. The Bayesian probability formalism was combined with the neural network methodology to construct various types of log-likelihood scoring functions. Then, we described some of the new knowledge-based scoring functions from in the recent literature. These functions extend the pairwise distance preference scoring functions in various ways, for example, by explicitly modeling the solvent effects and by considering multi-body geometric arrangements and interactions. Finally, we briefly discussed conformer clustering and described a detailed energy potential used for high-resolution refinement. In general, because of the weaknesses of solvent and electrostatic modeling, simulations attempting to fold proteins de novo from physics-based scoring functions alone do not perform satisfactorily. The statistical models that are used to construct knowledge-based functions provide added flexibilities over direct physical

modeling, and as a result, most of the successful de novo structure prediction protocols have both physics-based and knowledge-based components.

Scoring function design remains a very difficult problem. None of the existing physics-based and knowledge-based functions can faithfully reproduce the true energy landscape of the protein conformational space, and none of them can consistently and reliably select native-like conformations from non-native structures for a broad spectrum of proteins. The difficulty is mainly because the physical and statistical models considered so far in the literature cannot well approximate the quantum mechanical character of intra-molecular and solvent-protein interactions. Furthermore, scoring functions describing truly orthogonal characteristics of protein native conformations are difficult to discover, especially for the knowledge-based functions that are the sum of many constituent effects. Thus, it is of practical interest to continue developing various types of new scoring functions, to exploit their differences, and to capture the cumulative effect of incremental enrichments. Fortunately, the increase in the size of the PDB together with increased computational power means that the construction of more sophisticated knowledge-based scoring functions are now possible. More realistic electrostatics and solvation models are also being developed, increasing the capabilities of the physics-based force fields. These advances will play important roles to improving the state of the art of protein folding simulation and de novo structure prediction.

Acknowledgments

We thank Drs. Enoch Huang and Britt Park for their earlier edition on scoring functions for de novo protein structure prediction and the anonymous reviewer for the many helpful suggestions. This work is supported in part by a Searle Scholar Award, NSF Grant DBI-0217241, an NSF CAREER award, and NIH Grant GM068152 to R.S. and the University of Washington's Advanced Technology Initiative in Infectious Diseases.

References

1. Brenner, S., Levitt, M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
2. Brenner, S.E. (2001) A tour of structural genomics. *Nat. Genet.*, **210**, 801–809.
3. Burley, S.K. (2000) An overview of structural genomics. *Nat. Struct. Biol.*, **7 (Suppl)**, 932–934.
4. Heinemann, U., Illing, G., Oschkinat, H. (2001) High-throughput three-dimensional protein structure determination. *Curr. Opin. Biotech.*, **12**, 348–354.
5. Bonneau, R., Baker, D. (2001) Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173–189.

6. Anfinsen, C.B., Haber, E., Sela, M., White, F.H., Jr. (1961) The kinetics of formation of active ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.*, **47**, 1309–1314.
7. Doolittle, R. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.
8. Sander, C., Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
9. Murzin, A., Bateman, A. (1997) Distance homology recognition using structural classification of proteins. *Proteins*, **29S**, 105–112.
10. Bowie, J., Luthy, R., Eisenberg, D. (1991) Method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
11. Jones, D., Taylor, W., Thornton, J. (1992) A new approach to protein fold recognition. *Nature*, **258**, 86–89.
12. Moult, J., Fidelis, K., Zemla, A. Hubbard, T. (2003) Critical assessment of methods of protein structure prediction (CASP): round V. *Proteins*, **53**, 334–339.
13. Moult, J., Fidelis, K., Rost, B., Hubbard, T., Tramontano, A. (2005) Critical assessment of methods of protein structure prediction (CASP) – round 6. *Proteins*, **61**, 3–7.
14. Lee, J., Liwo, A., Ripoll, D., Pillardy, J., Scheraga, J. (1999) Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, **S3**, 204–208.
15. Samudrala, R., Xia, Y., Huang, E., Levitt, M. (1999) Ab initio protein structure prediction using a combined hierarchical approach. *Proteins*, **S3**, 194–198.
16. Simons, K., Bonneau, R., Ruczinski, I., Baker, D. (1999) Ab initio structure prediction of CASP3 targets using ROSETTA. *Proteins*, **S3**, 171–176.
17. Samudrala, R., Xia, Y., Levitt, M., Huang E.S. (1999) A combined approach for ab initio construction of low resolution protein tertiary structures from sequence, in *Proceedings of the Pacific Symposium on Biocomputing* (Altman, R. B., Dunker, A.K., Hunter, L., Klein, T.E., Lauderdale, K., eds.), World Scientific Press, Singapore, pp. 505–516.
18. Samudrala, R., Levitt, M. (2002) A comprehensive analysis of 40 blind protein structure predictions. *BMC Struct Biol*, **2**, 3–18.
19. Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., Pedersen, J.T. (1997) Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*, **29**, 2–6.
20. Moult, J., Hubbard, T., Fidelis, K., Pedersen, J.T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, **37**, 2–6.
21. Moult, J., Fidelis, K., Zemla, A., Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, **45**, 2–7.
22. Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., Karplus, M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**, 187–217.

23. Weiner, S., Kollman P., Nguyen, D., Case, D. (1986) An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.*, **7**, 230–252.
24. Jorgensen, W., Tirado-Rives, J. (1988) The OPLS potential function for proteins. Energy minimisations for crystals of cyclic peptides and crambin. *J. Amer. Chem. Soc.*, **110**, 1657–1666.
25. MacKerell, A.D., Jr., Bashford, D., Bellott, M., Dunbrack, R.L., Jr., Evanseck, J.D., et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616.
26. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Jr., Fergusson, D.M., Spellmeyer, D.C., Fox, D.C., Caldwell, J.W., Kollman, P.A. (1995) A second generation force field for the simulation of proteins and nucleic acids. *J. Amer. Chem. Soc.*, **117**, 5179–5197.
27. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A. (1992) Energy parameters in peptides: improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, **96**, 6472–6484.
28. Wodak, S., Rooman, M. (1993) Generating and testing protein folds. *Curr. Opin. Struct. Biol.*, **3**, 247–259.
29. Sippl, M. (1995) Knowledge based potentials for proteins. *Curr. Opin. Struct. Biol.*, **5**, 229–235.
30. Gilis, D., Rooman, M. (1996) Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. *J. Mol. Biol.*, **257**, 1112–1126.
31. Jernigan, R.L., Bahar I. (1996) Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.*, **6**, 195–209.
32. DeBolt, S.E., Skolnick, J. (1996) Evaluation of atomic level mean force potentials via inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. *Protein Eng.*, **8**, 637–655.
33. Zhang, C., Vasmatzis, G., Cornette, J.L., DeLisi, C. (1997) Determination of atomic desolvation energies from the structures of crystallised proteins. *J. Mol. Biol.*, **267**, 707–726.
34. Samudrala, R., Moult, J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, **275**, 895–916.
35. Huang, E.S., Samudrala, R., Park, B.H. (2000) Scoring functions for ab initio protein structure prediction. *Methods Mol. Biol.*, **143**, 223–245.
36. Hartree, D.R. (1957) *The Calculation of Atomic Structure*. John Wiley & Sons, New York.
37. Hohenberg, P., Kohn, W. (1964) Inhomogeneous electron gas. *Phys. Rev.*, **136**, 864.

38. Kauzmann, W. (1959) Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, **14**, 1–64.
39. Dill, K.A. (1990) Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
40. Morozov, A.V., Kortemme, T., Tsemekhman, K., Baker, D. (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 6946–6951.
41. Weiner, P.K., Kollman P.A. (1981) AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comp. Chem.*, **2**, 287–303.
42. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, **4**, 187–217.
43. Levitt, M., Hirshberg, M., Sharon, R., Daggett, V. (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Comm.*, **91**, 215–231.
44. Levitt, M. (1983) Molecular dynamics of native protein. I. Computer simulation of trajectories. *J. Mol. Biol.*, **168**, 595–617.
45. Daggett, L.P., Sacaan, A.I., Akong, M., Rao, S.P., Hess, S.D., Liaw, C., Urrutia, A., Jachec, C., Ellis, S.B., Dreessen J, et al. (1995) Molecular and functional characterization of recombinant human metabotropic glutamate receptor subtype 5. *Neuropharmacology*, **34**, 7133–7155.
46. Levitt, M. (1983) Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, **170**, 723–764.
47. Brunger, A.T., Clore, G.M., Gronenborn, A.M., Karplus, M. (1986) Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 3801–3805.
48. Ferguson, D.M., Kollman, P.A. (1991) Can the Lennard-Jones 6-12 function replace the 10–12 form in molecular mechanics calculations? *J. Comput. Chem.*, **12**, 620–626.
49. Halgren, T.A. (1992) Representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.*, **114**, 7827–7843.
50. Halgren, T.A. (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, **17**, 490–519.
51. Hart, J.R., Rappe, A.K. (1992) van der Waals functional forms for molecular simulations. *J. Chem. Phys.*, **97**, 1109–1115.
52. Buckingham, A.D., Fowler, P.W. (1985) A model for the geometries of van der Waals complexes. *Can. J. Chem.*, **63**, 2018.

53. Sokalski, W.A., Shibata, M., Ornstein, R.L., Rein, R. (1993) Point charge representation of multicenter multipole moments in calculation of electrostatic properties. *Theor. Chim. Acta*, **85**, 209–216.
54. Stone, A.J. (1981) Distributed multipole analysis, or how to describe a molecular charge distribution. *Chem. Phys. Lett.*, **83**, 233–239.
55. Kosov, D., Popelier, P.L.A. (2000) Atomic partitioning of molecular electrostatic potentials. *J. Phys. Chem. A*, **104**, 7339–7345.
56. Cieplak, P., Caldwell, J., Kollman, P. (2001) Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comput. Chem.*, **22**, 1048–1057.
57. Kaminski, G.A., Stern, H.A., Berne, B.J., Friesner, R.A., Cao, Y.X., Murphy, R.B., Zhou, R., Halgren, T.A. (2002) Development of a polarizable force field for proteins via ab initio quantum chemistry: first generation model and gas phase tests. *J. Comput. Chem.*, **23**, 1515–1531.
58. Ren, P., Ponder, J.W. (2003) Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B*, **107**, 5933–5947.
59. Jorgensen, W.L. (1981) Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *J. Am. Chem. Soc.*, **103**, 335–340.
60. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
61. Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.*, **91**, 6269–6271.
62. Levitt, M., Hirshberg, M., Sharon, R., Laidig, K.E., Daggett, V. (1997) Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J. Phys. Chem. B*, **101**, 5051–5061.
63. York, D.M., Darden, T., Pedersen, L.G. (1993) The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the Ewald and truncated list methods. *J. Chem. Phys.*, **99**, 8345–8348.
64. Darden, T., York, D., Pedersen, L. (1993) Particle mesh Ewald: an N*log(N) method for Ewald sums in large systems *J. Chem. Phys.*, **98**, 10089–10092.
65. Gouy, M. (1910) Sur la constitution de la charge électrique a la surface d'un électrolyte. *Journ. Phys.*, **9**, 457–468.
66. Gilson, M.K., Honig, B. (1988) Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*, **4**, 7–18.
67. Nicholls, A., Honig, B. (1991) A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comp. Chem.*, **12**, 435–445.

68. Bashford, D., Case, D.A. (2000) Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **51**, 129–152.
69. de Bakker, P.I.W., DePristo, M.A., Burke, D.F., Blundell, T.L. (2003) Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. *Proteins*, **51**, 21–40.
70. Fogolari, F., Brigo, A., Molinari, H. (2003) Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys. J.*, **85**, 159–166.
71. Warshel, A., Levitt, M. (1976) Theoretical studies of enzymic reactions – dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J. Mol. Biol.*, **103**, 227–249.
72. Gelin, B.R., Karplus, M. (1979) Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. *Biochemistry*, **18**, 1256–1268.
73. Lazaridis, T., Karplus, M. (1999) Effective energy function for proteins in solution. *Proteins*, **35**, 133–152.
74. Mallik, B., Masunov, A., Lazaridis, T. (2002) Distance and exposure dependent effective dielectric function. *J. Comp. Chem.*, **23**, 1090–1099.
75. Moult, J. (1997) Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.*, **7**, 194–199.
76. Eisenberg, D., Weiss, R.M., Terwillinger, T.C. (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**, 371–374.
77. Sippl, M.W., S. (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins*, **13**, 258–271.
78. Jones, D.T. (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins*, **45**, 127–132.
79. Zhang, Y., Skolnick, J. (2004) Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins. *Biophys. J.*, **87**, 2647–2655.
80. Boniecki, M., Rotkiewicz, P., Skolnick, J., Kolinski, A. (2003) Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.*, **17**, 725–738.
81. Hung, L.H., Ngan, S.C., Liu, T., Samudrala, R. (2005) PROTINFO: new algorithms for enhanced protein structure predictions. *Nucleic Acids Res.*, **33**, W77–W80.
82. Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
83. Bourne, P.E., Addess, K.J., Bluhm, W.F., Chen, L., Deshpande, N., Feng, Z., Fleri, W., Green, R., Merino-Ott, J.C., Townsend-Merino, W., Weissig, H., Westbrook, J., Berman, H.M. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.*, **32**, D223–D225.
84. Chandonia, J.M., Hon, G., Walker, N.S., LoConte, L., Koehl, P., Levitt, M., Brenner, S.E. (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

85. Subramaniam, S., Tcheng, D.K., Fenton, J. (1996) Knowledge-based methods for protein structure refinement and prediction, in *Proceedings of the Fourth International Conference on Intelligent Systems in Molecular Biology* (States, D., Agarwal, P., Gaasterland, T., Hunter, L. & Simth, R., eds.), AAAI Press, Menlo Park, CA, pp. 218–229.
86. Avbelj, F., Moult, J. (1995) Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, **34**, 755–764.
87. Lu, H., Skolnick, J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, **44**, 223–232.
88. Zhou, H., Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
89. Oppenheim, A.V., Schafer, R.W., Buck, J.R. (1999) *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, Upper Saddle River, NJ.
90. Rost, B., Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.
91. Ahmad, S., Gromiha, M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–824.
92. Kim, H., Park, H. (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins*, **54**, 557–562.
93. Rost, B., Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
94. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
95. Cuff, J.A., Barton, G.J. (1999) Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
96. Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., Brunak, S. (1997) Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.*, **10**, 1241–1248.
97. Pollastri, G., Baldi, P., Fariselli, P., Casadio, R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.
98. Olmea, O., Valencia, A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des.*, **2**, S25–32.
99. Fariselli, P., Casadio, R. (1999) Neural network based predictor of residue contacts in proteins. *Protein Eng.*, **12**, 15–21.
100. Altschul, S.F., Madden, T.L., Schaffer, A.A. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

101. Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.
102. Yi, T.-M., Lander, E.S. (1993) Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, **232**, 1117–1129.
103. Kabsch, W., Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
104. Zell, A., Mamier, G., Vogt, M., et al. (2005) The SNNS users manual version 4.1. Available at http://www-ra.informatik.uni-tuebingen.de/snns.
105. Park, B., Levitt, M. (1996) Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J. Mol. Biol.*, **266**, 831–846.
106. Novotny, J., Bruccoleri, R., Karplus, M. (1984) An analysis of incorrectly folded protein models. Implications for structure predictions. *J. Mol. Biol.*, **177**, 787–818.
107. Holm, L., Sander, C. (1992) Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.*, **225**, 93–105.
108. Samudrala, R., Levitt, M. (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.*, **9**, 1399–1401.
109. Tsai J., B., R., Morozov, A.V., Kuhlman, B., Rohl, C.A., Baker, D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, **53**, 76–87.
110. Park, B.H., Huang, E.S., Levitt, M. (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.*, **266**, 831–846.
111. Hinds, D.A., Levitt, M. (1992) A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 2536–2540.
112. Park, B., Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.*, **249**, 493–507.
113. Simons, K.T., Kooperberg, C., Huang, E., Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
114. Hung, L.H., Samudrala, R. (2003) PROTINFO: secondary and tertiary protein structure prediction. *Nucleic Acids Res.*, **31**, 3296–3299.
115. McConkey, B.J., Sobolev, V., Edelman, M. (2003) Discrimination of native protein structures using atom-atom contact scoring. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 3215–3220.
116. Carter, C.W., Jr., LeFebvre, B.C., Cammer, S.A., Tropsha, A., Edgell, M.H. (2001) Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.*, **311**, 625–638.
117. Krishnamoorthy, B., Tropsha, A. (2003) Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, **19**, 1540–1548.

118. Ngan, S.-C., Inonye, M.T, Samudrala, R. (2006) A knowledge-based scoring function based on residue triplets for protein structure prediction. *Protein Eng.*, **19**, 187–193.
119. Li, X., Hu, C., Liang, J. (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, **53**, 792–805.
120. Wang, K., Fain, B., Levitt, M., Samudrala, R. (2004) Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.*, **4**, 8.
121. Zhang, Y., Skolnick, J. (2004) SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.*, **25**, 865–871.
122. Samudrala, R. (2006). RAMP Howto. Available at http://software.compbio.washington.edu/ramp/ramp.html
123. Misura, K.M.S., Baker, D. (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, **59**, 15–29.
124. Bradley, P., Misura, K.M.S., Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**, 1868–1871.
125. Kortemme, T., Morozov, A.V., Baker, D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.*, **326**, 1239–1259.
126. Bonneau, R., Ruczinski, I., Tsai, J., Baker, D. (2002) Contact order and ab initio protein structure prediction. *Protein Sci.*, **11**, 1937–1944.
127. Bradley, P., Malmstrom, L., Qian, B., Schonburn, J., Chivian, D., Kim, D.E., Meiler, J., Misura, K.M., Baker D. (2005) Free modeling with Rosetta in CASP6. *Proteins*, **61**, 128–134.

QUERIES TO BE ANSWERED (SEE MARGINAL MARKS)

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Chapter-10

Query No.	Page No.	Line No.	Query
AQ1	262	34	The text 'Based on the work of Professor Banavar and his colleagues' has been changed to 'On the basis of the work of Professor Banavar and his colleagues'. Please check if this is OK.