PROTEINS: Structure, Function, and Genetics 43:161–174 (2001)

Lattice Protein Folding With Two and Four-Body Statistical Potentials

Hin Hark Gan,[1] Alexander Tropsha,[2] and Tamar Schlick[1]*

[1]*Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, New York, New York*

[2]*Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina*

ABSTRACT **The cooperative folding of proteins implies a description by multibody potentials. Such multibody potentials can be generalized from common two-body statistical potentials through a relation to probability distributions of residue clusters via the Boltzmann condition. In this exploratory study, we compare a four-body statistical potential, defined by the Delaunay tessellation of protein structures, to the Miyazawa–Jernigan (MJ) potential for protein structure prediction, using a lattice chain growth algorithm. We use the four-body potential as a discriminatory function for conformational ensembles generated with the MJ potential and examine performance on a set of 22 proteins of 30–76 residues in length. We find that the four-body potential yields comparable results to the two-body MJ potential, namely, an average coordinate root-mean-square deviation (cRMSD) value of 8 Å for the lowest energy configurations of all-α proteins, and somewhat poorer cRMSD values for other protein classes. For both two and four-body potentials, superpositions of some predicted and native structures show a rough overall agreement. Formulating the four-body potential using larger data sets and direct, but costly, generation of conformational ensembles with multibody potentials may offer further improvements. Proteins 2001;43:161–174.**

© 2001 Wiley-Liss, Inc.

Key words: lattice model; statistical potential; multibody potentials; chain growth algorithm; Monte Carlo

INTRODUCTION

Statistical potentials derived from simplified representations of protein residues are widely used in protein structure prediction. Because of rapid growth of protein structural databases, many of the current residue potentials are either derived directly from the protein database[1–3] or indirectly by having the parameters of the chosen functional forms determined by the crystal structures.[4] These statistical potentials have found wide-ranging applications in the evaluation of structure/sequence compatibility of proteins,[5,6] homology modeling,[7] and protein folding simulations.[8–10] Currently, most statistical potentials are two-body, such as the Miyazawa–Jernigan[1,2] (MJ) and Sippl[3] potentials. As a larger structural database emerges, the development of multibody potentials—to model the cooperative interactions in native proteins—becomes feasible.

While two-body potentials are adequate for most condensed systems, the importance of multibody potentials increases for dense molecular systems such as compact native protein structures. Multibody potentials may help improve our understanding of the cooperativity of protein folding process and the regularity of protein structures. Recent protein folding studies with multibody potential terms show that they play a role in stabilizing protein folds and in enhancing cooperativity of the folding/unfolding process. For example, Liwo et al.[4,11] have introduced three and four-body correlation terms in their united-residue potentials, which arise from the expansion of mean potentials. Four-body hydrogen bonding correlation terms have also been introduced phenomenologically by Kolinski and Skolnick[8] in their lattice Monte Carlo protein structure prediction algorithm.

To gain a better understanding of multibody potentials, we formulate multibody mean potentials in terms of the potentials of mean force in statistical mechanics.[12] This formulation implies that the multibody contact energy and probability of observing clusters of residues are generally related by the Boltzmann condition. Furthermore, lower-order mean potentials can be derived from the higher-order expressions. Following the methodology of statistical functions,[3,13] such mean potentials can be approximately derived from the protein structural database. Since they yield a greater amount of information about many-body correlations in compact molecular systems, multibody potentials may better characterize native proteins. However, deriving these potentials requires large protein structural databases for accurate formulations.

In this work, we examine a four-body potential developed by Tropsha and Vaisman and coworkers.[14] By using a threading technique, these researchers showed that their statistical potential can discern correct sequence/structure

Grant sponsor: National Institutes of Health; Grant number: GM55164; Grant number: RR08102; Grant sponsor: National Science Foundation; Grant number: BIR-94-23827ER; Grant number: ASC-9704681.

*Correspondence to: Tamar Schlick, Department of Chemistry and Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 251 Mercer Street, New York, NY 10012. E-mail: schlick@nyu.edu

Received 4 August 2000; Accepted 8 December 2000

© 2001 WILEY-LISS, INC.

matches better than two-body statistical potentials.[15,16] In the present investigation, we discuss the evaluation of this potential and its implementation for lattice protein folding using a chain growth algorithm. The evaluation is performed by means of a statistical geometrical method in which the four-body residue neighbors are systematically enumerated using the Delaunay tessellation technique. Four-body energies are then related to the frequencies of observed four-residue clusters in protein structures via the Boltzmann relation. Our evaluation shows that while most terms have attained their saturation values, some have not converged, indicating that a larger structural database is needed to determine this potential more accurately.

We compare the performance of the four-body potential in protein structure prediction to the two-body MJ potential using a lattice C_α protein model. On the (311) cubic lattice, we generate configurational ensembles for proteins with our recently implemented chain growth algorithm.[17] We have shown that this variant is effective for calculating conformational and thermodynamic properties of several test proteins.[17] In this article, we examine the quality of the two and four-body potentials using coordinate root-mean-square deviation (cRMSD) between native and predicted structures and energy/cRMSD scatter plots. As the computational cost in generating ensembles using the four-body potential is currently prohibitive, we can only apply it to the conformational ensembles generated with the two-body potential. We find that the shapes of the energy/cRMSD plots for the two and four-body potentials are correlated for low-energy and low cRMSD configurations. The two and four-body energies are generally weakly correlated.

For a set of 22 proteins, the predicted cRMSD values by the MJ potential are about 8 Å for α proteins and somewhat poorer, around 9 Å, for β and α/β protein classes. Superpositions of predicted and native structures show rough overall agreements for some proteins. Our four-body potential yields comparable results. Improving the four-body potential requires larger protein data sets than used in the present study. Furthermore, a more sophisticated protein model that uses finer representation for each residue may be required. Thus, although the present four-body potential may be adequate for protein fold recognition, further improvements are necessary for more accurate ab initio structure prediction.

STATISTICAL POTENTIALS

We begin by presenting the modified MJ potential, the methodology of the four-body statistical potential, and a theoretical formulation of multibody potentials in general. The derivation shows that multibody energies are generally related to the probability distributions of observing proximity of residues via the Boltzmann relation. It also shows how different levels of multibody potentials are related.

Multibody Potentials of Mean Force

Statistical potentials are derived based on the assumption that "contact" energies between amino acid residues in native proteins are related to their observed frequency in a representative structural database. Since energies are computed from a set of folded structures, they are more appropriately interpreted as a mean, rather than as bare interaction energies. These mean energies are related to potentials of mean force in statistical mechanics, obtained by ensemble averaging over equilibrium states.[13,18] In structure-derived potentials, ensemble averaging is effectively replaced by averaging over a set of representative protein structures. Critical assessment of this interpretation using a simple two-residue hydrophobic–hydrophilic (HP) protein model shows that it yields a correct ranking of contact energies, but their absolute values are imprecise.[18] The following discussion presents two-body and multibody potentials as potentials of mean force.

For a protein chain with N residue interaction centers, we define the n-body potential of mean force $w^{(n)}$ for the residue-cluster $(i_1 i_2 \ldots i_n)$ as

$$w^{(n)}_{i_1 i_2 \ldots i_n} = -k_B T \ln[F^{(n)}_{i_1 i_2 \ldots i_n} / R_{i_1 i_2 \ldots i_n}] \quad (1)$$

where $R_{i_1 i_2 \ldots i_n}$ is the reference state (defined below) and $F^{(n)}$ is the probability density of finding the residues in a cluster:

$$F^{(n)}_{i_1 i_2 \ldots i_n} = \int dV^{(N-n)} \exp(-\beta E_N) \Big/ \int dV^{(N-1)} \exp(-\beta E_N) \quad (2)$$

where the temperature parameter $\beta = 1/(k_B T)$, the volume element $dV^{(N-n)} = (dV_1 dV_2 \ldots dV_N)/(dV_{i_1} dV_{i_2} \ldots dV_{i_n})$, and E_N is the total energy of the protein. The n-body potential is defined with respect to a "reference state" $R_{i_1 i_2 \ldots i_n}$ determined by the specific problem of interest. Moreover,

$$F^{(n)}_{i_1 i_2 \ldots i_n} = R_{i_1 i_2 \ldots i_n} \quad (3)$$

when the mean potential $w^{(n)}_{i_1 i_2 \ldots i_n} = 0$.

It is useful to separate correlated from uncorrelated aspects of many-body interactions. The reference state is often chosen to be the uncorrelated state where there are no interactions between residues. If the interaction energy E_N vanishes, $F^{(n)}_{i_1 i_2 \ldots i_n}$ in eq. 2 becomes a product of the uncorrelated, single residue properties. Mathematically, we write $R_{i_1 i_2 \ldots i_n} \sim \Pi^n_\alpha R_{i_\alpha}$, where R_{i_α} is the frequency of occurrence of individual residues. Consequently, the mean potential $w^{(n)}_{i_1 i_2 \ldots i_n}$ is interpreted as a measure of the nonrandom nature of residue distributions, or contacts, in protein structures. The correlations among the residues in the cluster $(i_1 i_2 \ldots i_n)$ are determined by the temperature and energy function E_N through the canonical average. A major simplifying assumption of statistical potentials is that the probability density $F^{(n)}_{i_1 i_2 \ldots i_n}$ is determined by the observed frequencies of the residue cluster $(i_1 i_2 \ldots i_n)$ for $n = 2, 3, 4, \ldots$, instead of evaluating the expression in eq. 2.

Multibody mean potentials contain information about correlations between residue pairs, triplets and quadruplets, and so on, in the system. In general, mean potentials $w^{(n)}_{i_1 i_2 \ldots i_n}$ for different n are related through probability densities:

$$F^{(n)}_{i_1 i_2 \ldots i_n} = \int dV_{n+1} F^{(n+1)}_{i_1 i_2 \ldots i_{n+1}} \quad (4)$$

By using the definition for the distribution functions, we can relate the different levels of multibody mean potentials, $w^{(n)}_{i_1 i_2 \ldots i_n}$ for $n = 2, 3, \ldots$, as follows:

$$w^{(n)}_{i_1 i_2 \ldots i_n} = -k_B T \ln\left[\int dV_{n+1} \frac{R_{i_1 i_2 \ldots i_{n+1}}}{R_{i_1 i_2 \ldots i_n}} \exp(-\beta w^{(n+1)}_{i_1 i_2 \ldots i_{n+1}})\right] \quad (5)$$

This formula provides the recipe for determining lower-order potentials in terms of higher-order potentials; e.g., $w^{(3)}_{i_1 i_2 i_3}$ is derived from $w^{(4)}_{i_1 \ldots i_4}$ and $w^{(2)}_{i_1 i_2}$ from $w^{(3)}_{i_1 i_2 i_3}$. Because of molecular packing, two and three-body correlations are significant even for disordered monatomic liquids. Higher-order (correlated) potentials are expected to be important for folded protein structures whose packing density resembles that of crystals. Still, they are more difficult to evaluate than two-body potentials. From a practical viewpoint, estimating lower-order potentials directly from structure database may lead to more accurate results than using eq. 5, especially when the higher-order potentials cannot be accurately determined from databases.

Scheraga and coworkers recently considered multibody terms in their interaction potentials for residues derived as a mean potential (cumulant) expansion.[11] These investigators argued that, in the leading approximation, the four-body contributions are similar to the four-body cooperative hydrogen bonding interactions introduced by Kolinski and Skolnick.[8] Multibody terms derived in this way are related to, but different than, the mean potentials $w^{(n)}_{i_1 i_2 \ldots i_n}$ derived here. In the following discussion, we consider the use of two- and four-body potentials of mean force estimated from protein structural databases.

Two-Body Potential

From eq. 1, the two-body contact energies $\{\epsilon_{ij}\}$ are related to the frequency of residue pairs i, j in the protein structural database via the Boltzmann's condition:

$$\epsilon_{ij} = -k_B T \ln[F_{ij}/R_{ij}] \quad (6)$$

where F_{ij} is now the observed contact frequency for the pair i, j in protein database and R_{ij} is its corresponding reference state.[1,3,19] Contact energies $\{\epsilon_{ij}\}$ calculated with the random reference state reflect the residual nonrandom preferences for residue/residue contacts in proteins; energies calculated with respect to a solvent-mediated reference state are related to the preferences for residue/residue over residue/solvent contacts. The MJ energies are derived using a solvent-mediated reference state and are correlated with experimental hydrophobicities of residues.[1]

For this work, we use a slightly modified version[17] of the MJ potential in which the interaction matrix is modified by a simple shift: $\epsilon_{ij} \leftarrow M_{ij} + 2$, where M_{ij} is the MJ interaction matrix[2] as reevaluated in 1996; the energies are expressed in $k_B T_0$ units, where T_0 is the room temperature. Our simple shift weakens the attractive energies between the residues, and they are effectively similar to the interaction energies derived by Skolnick and coworkers.[20,21]

We choose a simple square-well function to parameterize the residue/residue potential. The attractive interactions are represented by the shifted MJ energies $\{\epsilon_{ij}\}$. For each i and j pair representing two residues, with distance separation R_{ij}, our potential has the form

$$u_{ij}(R_{ij}) = \begin{cases} \epsilon_r & \text{if } R_{ij} < 4\text{Å} \\ \epsilon_{ij} & \text{if } 4\text{Å} \le R_{ij} \le 6.5\text{Å} \\ 0 & \text{if } R_{ij} > 6.5\text{Å} \end{cases} \quad (7)$$

where ϵ_r is a residue-independent finite repulsive energy, and ϵ_{ij} is modified MJ energy. The short-range repulsive energy ensures minimal overlap between protein cores. The value of ϵ_r is set simply as: $\epsilon_r = 5 \max_{\{ij\}}|\epsilon_{ij}|$; we found results to be insensitive to the precise value of ϵ_r within a wide range.

Four-Body Potential

As introduced earlier, multibody statistical potentials are conceptually similar to two-body analogues, although the methodology used to evaluate them can vary. For two-body potentials, residues that are within a prescribed radius R_{cut} (e.g., ~7 Å) from the reference residue are typically considered neighbors. This procedure is inadequate for multibody contacts because it leads to over counting of multibody contributions, since within a given cutoff radius the number of possible multibody (interaction) terms is larger than allowed geometric nearest neighbors. Thus, a rigorous definition of contact neighbors is required to relate contact energies to residue neighbors. Tropsha and Vaisman and coworkers[14] have introduced a novel multibody potential derived from computational geometry analysis of protein structures. In their scheme, the united residues (typically represented by C_α atoms or side chain centroids) are used to tessellate protein structures using the Delaunay triangulation technique[22] (Fig. 1). The shape and volume of a tessellated protein structure are defined by the aggregate of tetrahedrons whose vertices are C_α centers. The vertices of the tetrahedrons define unique four-residue clusters, which are the basis for the computation of the four-body statistical potential.

The following discussion presents the four-body potential and the methodology for its determination. We reevaluate the potential with a larger protein data set and show the effects of data set size on the results.

Tessellating protein structures

The closest neighbors for each point in a set of arbitrary points in space can be identified with the aid of Voronoi or Delaunay tessellation technique.[22] A Voronoi tessellation of a set of points or sites in three dimensions defines an aggregate of polyhedra or convex polytopes enclosing the points. The faces of a polyhedron are boundaries defined by planes perpendicular to the lines joining a point and its nearest neighbors; in 2D, the boundaries are lines (Fig. 1). Thus, a region in a Voronoi polyhedron is closest to the point inducing the tessellation than to the points in other

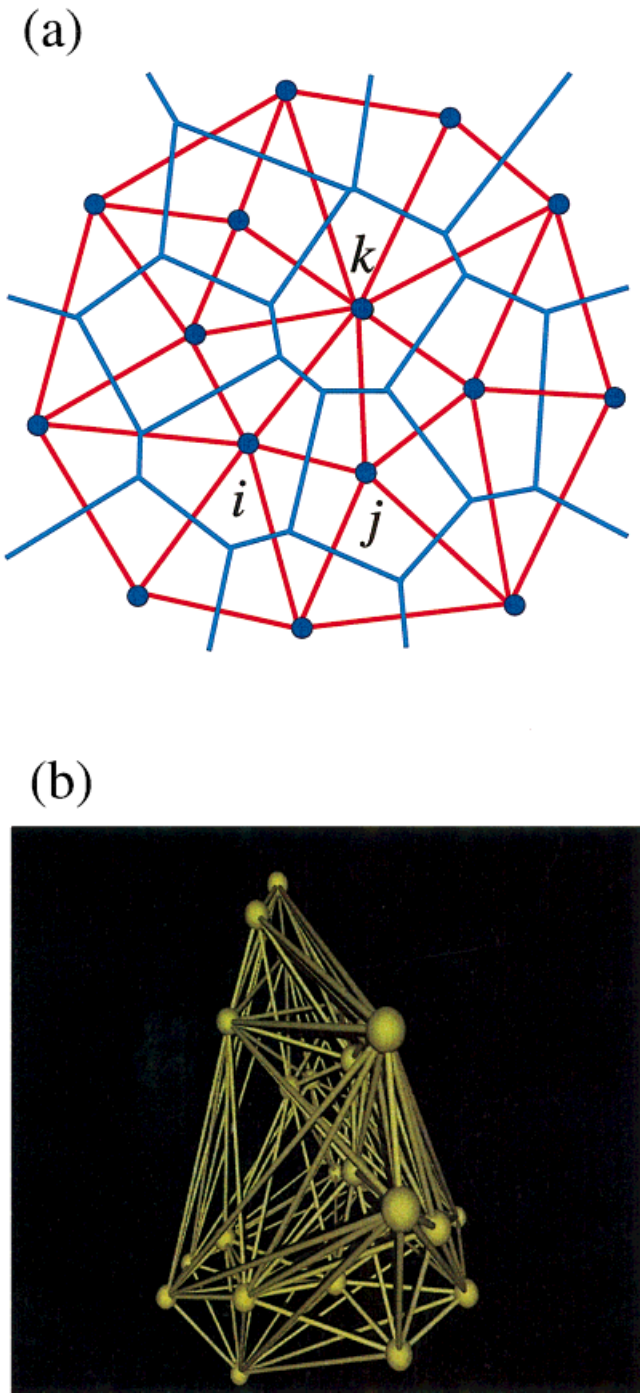


Fig. 1. **a:** Voronoi (blue) and Delaunay (red) tessellations of N points (black dots) in two dimensions where the triangles (ijk) form clusters of near neighbors. **b:** Delaunay tessellation of the protein 2mhu, where the vertices are C_α positions and four-residue (tetrahedral) clusters (ijkl) form near neighbors.

polyhedra. The Delaunay triangulation is defined by a set of triangles formed by lines joining the points that share a boundary. Thus, Delaunay tessellation is mathematically the dual of Voronoi tessellation. Figure 1a illustrates the Voronoi and Delaunay tessellations for a set of points in two dimensions.

In 2D, the Delaunay triangles partition the area occupied by the points, which may be viewed as an aggregate of three-point nearest-neighbor clusters. Similarly, in 3D the Delaunay tessellation produces an aggregate of four-body clusters or tetrahedrons. Information about the distribution of these clusters helps characterize the geometric properties of any system. Thus, tessellation techniques are useful for analyzing irregular structures, such as disordered crystals[23] and proteins.[24] Delaunay tessellation of protein 2mhu using only the C_α is shown in Figure 1b. We only use a simplified C_α representation of protein chains.

The statistical geometrical characterization of protein structures can acquire physical meaning when the frequencies of the residue compositions of the tetrahedrons are related to residue contact energies through the Boltzmann relation (see eq. 1). Our analysis ignores the presence of solvent molecules, metal ions, heme groups, and other molecules complexed with proteins. We use the program developed by Barber et al.[25] to tessellate native proteins, which gives all possible four-body C_α neighbors in protein structures. The upper bound for the algorithmic complexity of generating Voronoi or Delaunay tessellations is estimated to be $N^{(D+3)/2}$, where N is the number of points and D is spatial dimension.[26]

More generally, the tessellation technique as described above can be used to define two-, three-, and four-body statistical potentials. However, the discriminatory power of the two and three-body potentials derived in this manner will have to be tested, for example, using threading or decoys before implementing them in folding simulations. As discussed earlier, we expect the four-body potential to discern native structures better than the two- and three-body potentials. A combination of multibody potentials may work even better.

Evaluating the four-body potential from representative protein structures

Following Tropsha and Vaisman and colleagues,[14] the four-residue contact energies Q^{α}_{ijkl} are computed using the formula

$$Q^{\alpha}_{ijkl} = -k_B T \ln[f^{\alpha}_{ijkl}/p_{ijkl}] \tag{8}$$

where f^{α}_{ijkl} is the frequency of residue composition $(ijkl)$ in a set of protein structures, p_{ijkl} is the expected random frequency for each combination $(ijkl)$, and superscript α denotes the type of four-body contact used (defined below). This expression corresponds to $n = 4$ in eq. 1. (Tropsha and Vaisman and colleagues define a four-body score using base 10 logarithm instead of the natural logarithm used here; they also ignore the $-k_B T$ factor.) In addition, four-body contacts of adjacent residues along the chain backbone (e.g., consecutive residues $i, i+1, i+2, i+3$) are distinguished from those contacts with nonsequential vertices. We label the different quadruplet types by the superscript $\alpha = 0, 1, 2, 3, 4$, denoting how many residues are consecutive along the chain. Thus, for example, $\alpha = 4$ corresponds to all four residues adjacent along the chain (i through $i+3$), and $\alpha = 0$ represents four indices, none of which are nearest neighbors along the chain. Given a set of

representative protein structures, the observed frequencies f^{α}_{ijkl} and p_{ijkl} in eq. 8 are defined as follows:

$$f^{\alpha}_{ijkl} = \frac{\text{observed occurrences of type } \alpha \text{ } (ijkl) \text{ neighbors}}{\text{total number for } \alpha \text{ type}} \quad (9)$$

and

$$p_{ijkl} = \frac{4!}{\prod_{i=1}^{N_{aa}} t_i!} a_i a_j a_k a_l \quad (10)$$

where

$$a_i = \frac{\text{observed occurrences of amino acid type } i}{\text{total number of residues in data set}} \quad (11)$$

Here, N_{aa} is the number of amino acid groups (20 if each amino acid is a group, but less if subgroups are formulated; see below) and t_i is the number in each type. The observed occurrences of $(ijkl)$ in f^{α}_{ijkl} are derived from Delaunay tessellations of protein structures.

The frequencies f^{α}_{ijkl} are derived from a representative protein data set of structures that spans various different protein families or folds. This ensures that the computed four-body energies are unbiased by over representation of certain protein families. We use the nonredundant protein database designed by Hobohm and Sander,[27] which includes proteins with resolution of ≤2.5 Å, where no two proteins have more than 25% sequence identity. This list, released on January 8, 1999, includes 840 proteins (http://www.sander.embl-heidelberg.de/pdbsel/). We further screened these proteins to exclude chains with unusually large gaps between adjacent C_α (more precisely, with gaps of >4.12 Å); this can occur by limited X-ray resolution for certain residues. The final set of 666 proteins was used to calculate the four-body potential according to eq. 8.

To compute f^{α}_{ijkl}, we include only those tetrahedrons whose edge lengths are less than a specified cutoff value R_{cut} (we use 8 Å here); R_{cut} is dictated by the range of physicochemical interactions (the typical range for two-body potentials is 6–7 Å). We found that smaller cutoffs lead to several quadruplet residue combinations $(ijkl)$ with no representations. As larger protein data sets become available in the near future, lower cutoff values might be implemented. Detailed analysis of the distribution of tetrahedron sizes shows that the percent of tetrahedrons excluded by the distance filter could be substantial[16] (>60% for R_{cut} = 8 Å). Most relatively small tetrahedrons are formed by hydrophobic residues in protein interiors, and they contribute more to the four-body potential than the large tetrahedrons found on the protein surface (Fig. 1b).

Since the available representative protein dataset is not large, we group similar residues to reduce the number of terms in the four-body potential. For two-body potentials, this is not an issue since there are only 210 distinct residue pairs (20 × 19/2 + 20). The four-body potential has 20^4 = 160,000 parameters in comparison. However, if we assume that all four-residue clusters $(ijkl)$ that are related by permutations are equivalent, the number of required parameters is reduced to 8,855. Since this number is still large,[16] Tropsha and Vaisman and colleagues,[15] after Goldstein et al.,[28] have considered reduced residue potentials based on the following six residue types (letters) as defined by George et al.[29] from correlations between residue mutation rates and Dayhoff matrices:

c = {cysteine}

f = {phenylalanine, tyrosine, tryptophan}

h = {histidine, arginine, lysine}

n = {asparagine, aspartic acid, glutamine, glutamic acid}

s = {serine, threonine, proline, alanine, glycine}

v = {methionine, isoleucine, leucine, valine}

The above residue categories have the following characteristics: cysteine residues (type c) can form disulfide bonds; f residues are aromatic; h residues are basic except histidine; n residues are glutamic and aspartic acids and their amide forms; s residues consist of hydroxyl (serine and threonine) and other residues; and v residues are mostly aliphatic. This reduction implies only 126 distinct quadruplet residue compositions for each of the five α types, or 630 parameters instead of 20^4. We adopt this residue classification; other classifications have been used.[14]

The four-body contact energies defined in eq. 8 are presented in Figure 2, where Q^{α}_{ijkl} is plotted versus quadruplet residue composition $(ijkl)$ for quadruplet types α = 0, 1, 2. Only these α types have Q^{α}_{ijkl} values that differ significantly from zero or the random reference state. The value of the potential for these types (which reflects that

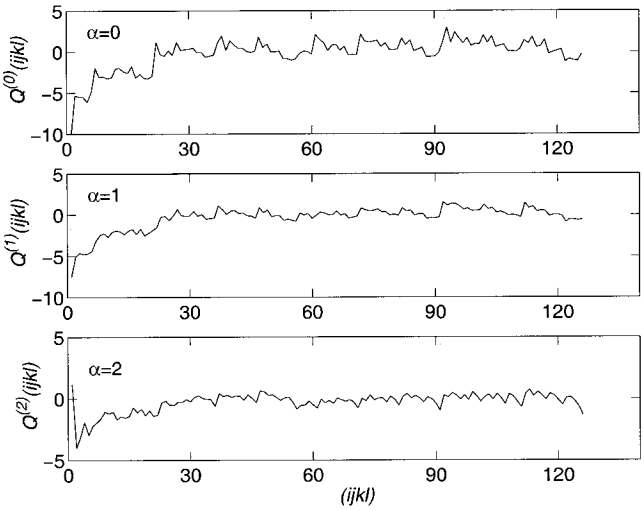


Fig. 2. Four-body potentials of types α = 0, 1, 2 are plotted as a function of residue quadruplets in the order: cccc (1), cccf (2), ccch (3), . . . , vvvv (126). Letters c, f, h, n, s, and v refer to six distinct residue classes, as described in the text. The full list of 126 quadruplets is available on our web site http://monod.biomath.nyu.edu/~hgan/delaunay.html. For α = 3, 4, the potentials are smaller in magnitude.

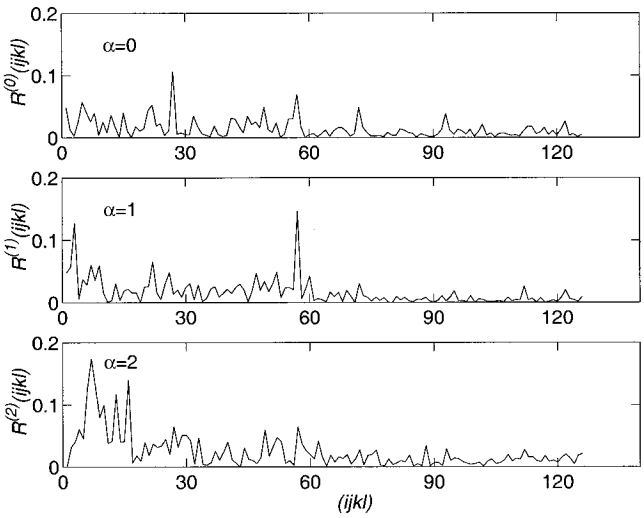


Fig. 3. Effect of protein dataset size on the four-body potential. The difference R^{α}_{ijkl} (see eq. 13) between four-body potentials evaluated with 309 and 666 representative proteins is shown as a function of the residue composition (ijkl) for three (α = 0, 1, 2) types. For α = 3, 4, the R^{α}_{ijkl} are considerably larger due to statistical fluctuations.

nonrandomness of residue contacts) decreases as α increases. Thus, nonbonded quadruplet configurations (α = 0) contribute more significantly to the potential than those types that have some bonded neighbors. This means that nonlocal four-body contacts in protein structures are largely nonrandom, unlike the contact terms with bonded neighbors. The nonrandom terms of the four-body potential should better discriminate protein-like from non-protein like conformations. The near-zero likelihoods of connected residue quadruplets is in agreement with analysis of protein sequences which suggests that native sequences are apparently random.[30] For the α = 0, 1, 2 types, the quadruplets with two or more cysteine residues (residue type *c*, the first 20 quadruplets in Fig. 2) have the largest magnitudes compared with other residue combinations (*ijkl*). This result is consistent with two-body statistical potentials where the cysteine–cysteine attractive energy is large as well.[2,31]

We also test the convergence of the four-body potential with respect to the size of the representative protein set. The Q^{α}_{ijkl} function is evaluated using two data sets: 309 proteins based on Hobohm and Sander's 1994 list versus the 666 proteins from their 1999 list; we denote the potentials generated thereby as Q^{α}_{ijkl} (309) and Q^{α}_{ijkl} (666), respectively. For this comparison, we used a cutoff R_{cut} = 11 Å, since the smaller data set requires a larger cutoff value to ensure nonzero counts for all quadruplets. Figure 3 shows the ratio of their difference:

$$R^{\alpha}_{ijkl} = |\delta Q^{\alpha}_{ijkl}/Q^{\alpha}_{ijkl}(666)| \qquad (13)$$

where $\delta Q^{\alpha}_{ijkl} = Q^{\alpha}_{ijkl}$ (666) $- Q^{\alpha}_{ijkl}$ (309). As shown in Figure 3, most quadruplets (*ijkl*) have small R^{α}_{ijkl} values (i.e., $\leq$0.05), but some values vary between 0.1 to 0.2. Clearly, larger data sets are needed to produce better convergence for the four-body potential. Unlike the four-body potential, the two-body statistical potentials are not sensitive to the size of available representative proteins.[2]

RESULTS AND DISCUSSION

The (311) cubic lattice and chain growth method for generating protein conformations are briefly outlined under Materials and Methods; detailed formulations are given elsewhere.[17] Since the use of the four-body potential for ensemble generation is prohibitively costly due to repeated tessellation of chain configurations at each step of the chain growth process, we assess it on the fully grown configurations generated by the two-body MJ potential. We typically generate a million configurations for each protein in the set of 22 small proteins examined in this discussion. Simulations are performed on a 300-MHz R12000 SGI Origin2000 computer at New York University. Sampling of a million configurations for a 30-residue protein requires about 2 CPU h.

Criteria for Selecting Nativelike Conformations

Our objective is to predict the single configuration from the ensemble that resembles the native structure as much as possible. A lowest-energy criterion assumes that the conformational entropy of native structures is small or negligible. Alternatively, since the thermodynamics hypothesis of protein folding implies that the free energy of the native state is the lowest, we can choose the states that contribute the most to the free energy expression. In the chain growth algorithm,[17] the free energy $F \sim -k_B T \ln \Sigma_{\Lambda} W(\Lambda, T)$, where W (Λ, T) is the statistical weight for configuration Λ at temperature T. The state with the highest weight W (Λ, T) corresponds to the lowest free energy. We thus compare results of selecting native configurations based on both lowest energies and highest statistical weights.

Energy/cRMSD Correlation Plots

The energy/cRMSD scatter plots for the proteins 2mhu (30 residues), 8rxna (52 residues), and 1r69 (63 residues) are shown in Figure 4. The native structure of the small protein 2mhu is disordered with no secondary structures, whereas 8rxna is a β protein and 1r69 an α protein. In the plots, only a representative 50,000 out of a million configurations from the ensembles are shown (we selected the first 50,000 (uncorrelated) configurations for plotting). The shapes of energy/cRMSD scatter plots for two-body (MJ) potential show that low-energy states are much more narrowly distributed than are the high-energy states. More important, the low-energy states tend to have lower cRMSD than do the high-energy states. This correlation is, however, not very strong, since many high-energy states have comparable low cRMSD values. Ideally, the shape of the energy/cRMSD plot should have the low-energy states shifted strongly toward low cRMSD. The four-body energy/cRMSD scatter plots for proteins 2mhu and 8rxna are similar to those for the MJ potential, but the plot for protein 1r69 shows a very different distribution: low-energy states are also broadly distributed. Thus, the

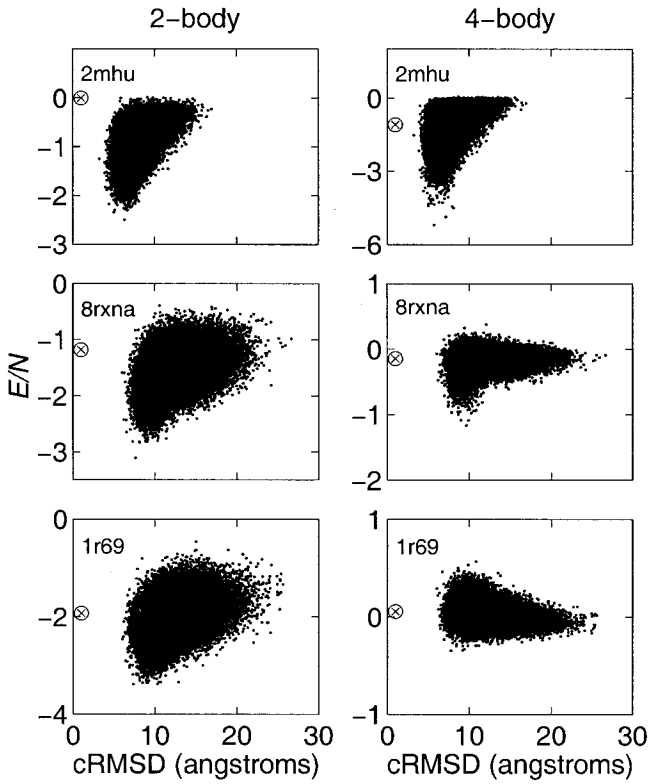


Fig. 4. Scatter plots of energy per residue versus coordinate root-mean-square deviation (cRMSD) for proteins 2mhu, 8rxna, and 1r69. The energies are expressed in units of k_BT_0, where T_0 is the room temperature (298 K). Configurations were generated with the MJ interaction matrix, but both two-body MJ and four-body energies were calculated. Left: MJ energy versus cRMSD. Right: four-body versus cRMSD. Each plot displays 50,000 uncorrelated configurations. The energy/cRMSD position of the native configuration is marked using the symbol ⊗.

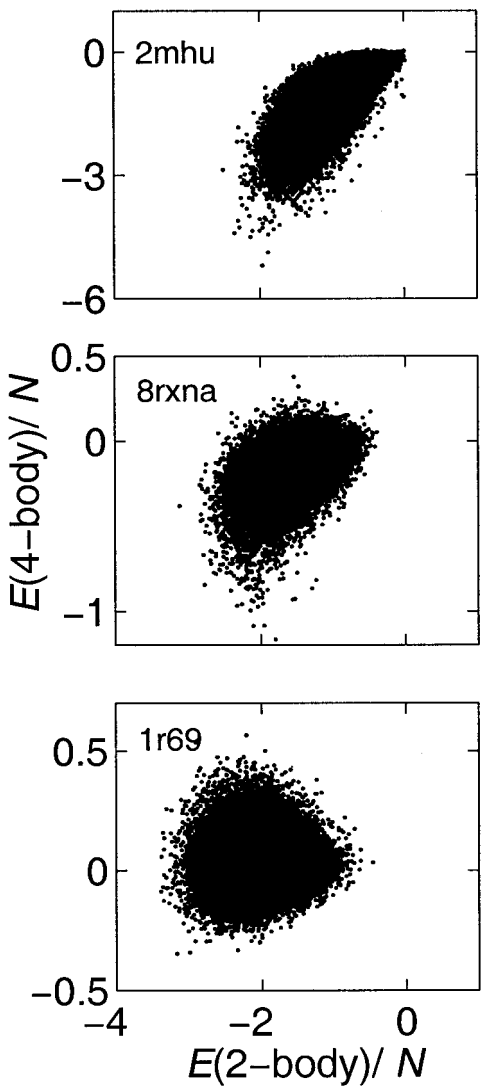


Fig. 5. Scatter plots of correlations between MJ (2-body) and four-body energies (per residue) for configurations of proteins 2mhu (top), 8rxna (middle), and 1r69 (bottom); the energies are in units of k_BT_0. Each plot displays 50,000 uncorrelated configurations.

four-body energies of the configurational ensemble are sensitive to the protein structure.

Levitt's group[32,33] has examined factors affecting the quality of energy functions through the energy/cRMSD plots. They found poor energy/cRMSD correlations for Hinds-Levitt[31] and MJ statistical potentials. Heuristic statistical potentials retain much fewer details of interaction between residues, and they are expected to show weaker energy/cRMSD correlations. This has led to the development of energy discriminatory functions with a detailed representation of atomic sites.[34] Moreover, recent folding simulation studies with all-atom potentials plus solvation demonstrate that these potentials can discriminate native states quite successfully.[35,36] Although detailed atomic potentials offer good discriminatory functions, they are much more costly to use in this context. A procedure that combines the strengths of different energy functionals could be a fruitful strategy for sampling and evaluating configurations.

For the three proteins 2mhu, 8rxna, and 1r69 shown in Figure 4, the cRMSD values based on the lowest two-body MJ energy are 5.10, 7.77, and 7.53 Å, respectively. The corresponding cRMSD values based on the lowest four-body energies are 4.37, 8.61, and 8.88 Å. Thus, in these cases, the four-body potential does not always yield cRMSD values lower than those predicted by the two-body MJ potential. Although the cRMSD values obtained in this case are similar to others reported in the literature, the predicted accuracy of the structures is quite low.

Correlation and Comparison Between Two- and Four-Body Energies

The above differences and similarities between the energy/cRMSD plots of two- and four-body statistical potentials can be quantified by plotting the correlations between the two- and four-body energies of configurational ensembles for proteins 2mhu, 8rxna, and 1r69, as shown in Figure 5. For proteins 2mhu and 8rxna, an overall correlation exists: configurations with low two-body energies also have low four-body energies, although there is a consider-

TABLE I. Modified Miyazawa-Jernigan (MJ) and Four-Body Energies (Per Residue) of Native and Lowest-Energy Structures Compared for a Set of 22 Proteins[†]

Proteins	Size	Native MJ	Native four-body	Lowest MJ	Lowest four-body
Disordered peptide					
2mhu	30	0.01	−1.07	−2.93	−5.29
α-class proteins					
sini	31	−1.59	−0.11	−5.05	−0.55
1ppt	36	−1.85	−0.14	−3.67	−0.45
1r69	63	−1.92	0.06	−3.85	−0.41
2cro	65	−2.11	0.01	−4.71	−0.48
4icb	76	−1.53	0.00	−4.08	−0.33
1bg8	76	−1.41	−0.06	−3.56	−0.46
β-class proteins					
1apo	42	−0.96	−0.99	−2.84	−2.28
2bds	43	−1.51	−0.91	−4.39	−2.38
1atx	46	−1.95	−1.75	−3.88	−1.91
2ech	49	−1.12	−1.87	−2.61	−2.69
1tpm	50	−1.81	−0.35	−4.09	−1.17
8rxna	52	−1.17	−0.14	−3.19	−1.27
1hcc	59	−1.31	−0.59	−3.57	−0.89
Mixed α and β					
5znf	30	−0.64	0.07	−1.95	−0.73
1pnh	31	−0.87	−1.71	−5.38	−3.25
1sis	35	−2.27	−2.66	−4.19	−3.81
1shp	55	−2.07	−0.53	−4.02	−1.42
7pti	58	−1.99	−0.28	−3.93	−1.04
2sn3	65	−1.70	−1.09	−3.44	−1.67
1ctf	68	−2.00	−0.14	−3.69	−0.58
1ubq	76	−1.73	−0.14	−4.16	−0.38

[†]Native MJ and four-body energies are in columns 3 and 4, respectively, and their lowest energies in generated ensembles of one million configurations are in the last two columns. Energies are in units of k_BT_0, where T_0 is room temperature.

able spread in the low-energy end of the plot. By contrast, the 1r69 example shows almost no correlation, as evident from the energy/cRMSD plots in Figure 4. Thus, the correlation between two- and four-body potentials is protein dependent. Furthermore, the magnitude of the four-body energy, unlike its two-body counterpart, varies greatly from protein to protein. The origin of these observed differences is discussed below.

For this assessment, Table I summarizes the two and four-body energies per residue (in k_BT units at room T_0) associated with the native (middle two columns) and lowest-energy structures in corresponding ensembles for all 22 proteins tested here. Noting that the reference state for the two potentials differs, the four-body energies of native proteins are higher than their two-body energies. Table I shows that the four-body energies of α proteins are significantly higher than those for β and mixed α/β proteins. This may reflect larger contributions from nonlocal contacts (i.e., type 0 potential) for β and mixed α/β proteins over the local contacts (i.e., type 4) for α proteins. Native four-body energies (column 4) of β and mixed α/β proteins are comparable in magnitude, although there are considerable fluctuations within the protein sets, and we find no notable dependence on protein size. In marked contrast, the native two-body energies (per residue) (MJ column 3 in Table I) for the proteins are fairly uniform except for the smallest proteins. Similar behavior is seen in the lowest ensemble energies (columns 5 and 6, respectively, in Table I) for both potentials.

Table I indicates that the two-body potential is the dominant term and that the four-body potential is sensitive to specific proteins types; this is expected, given that multibody interactions depend more strongly on protein conformations. Traditionally, the energy function of a system is written as a sum of the two, three, four-body, and so on, energy contributions. Thus an additive combination of two and four-body energies appears reasonable. However, we found that the energy/cRMSD plots do not improve significantly with such a combined energy function. Perhaps the quality of the configurational ensembles generated with the MJ energy needs to be improved to make the new energy discriminating functions fruitful.

Protein Classes and Predicted Structures

As in Table I, results are organized in Table II according to protein classes (all-α, all-β, and mixed α/β classes, and a disordered protein) to assess predictions based on lowest MJ and four-body energies, highest statistical weight (SW), and the lowest RMSD values in the conformational ensembles. The average cRMSD values of the proteins

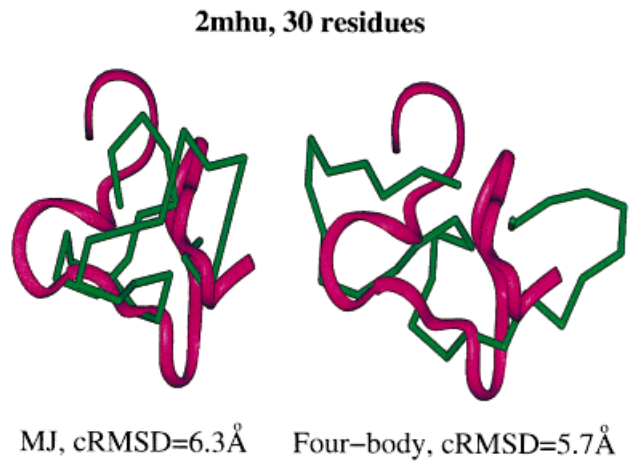
TABLE II. cRMSD of Predicted Structures and Ensemble-Averaged dRMSD ($\sqrt{\langle D^2_{drms}\rangle}$) for 22 Proteins†

Proteins	Size	MJ	Four-body	SW	Lowest cRMSD	$\sqrt{\langle D^2_{drms}\rangle}$
Disordered peptide						
2mhu	30	5.10	4.37	5.10	3.27	3.71
α-class proteins						
sini	31	7.62	6.71	7.32	4.18	6.58
1ppt	36	6.31	7.66	8.73	4.93	7.31
1r69	63	7.53	8.88	7.53	5.96	4.81
2cro	65	8.68	9.14	8.68	6.25	5.61
4icb	76	8.08	11.98	10.75	5.92	6.02
1bg8	76	10.91	10.93	11.79	6.36	7.64
Average:	57.8	8.19	9.21	9.13	5.60	6.32
β-class proteins						
1apo	42	10.03	8.99	10.03	5.49	6.56
2bds	43	8.98	8.14	8.98	5.34	5.40
1atx	46	8.14	8.89	8.14	5.79	5.29
2ech	49	8.68	8.27	10.58	5.90	6.77
1ptm	50	8.46	12.00	10.38	7.70	7.56
8rxna	52	7.77	8.61	8.39	5.11	5.09
1hcc	59	10.81	10.31	11.60	6.47	8.03
Average:	48.7	8.97	9.36	9.73	5.97	6.39
Mixed α and β						
5znf	30	5.67	6.03	7.20	4.83	4.78
1pnh	31	7.36	7.71	7.36	4.21	5.46
1sis	35	8.37	7.58	7.59	4.17	4.55
1shp	55	10.39	11.20	10.04	5.17	6.30
7pti	58	9.72	10.44	10.23	5.90	6.84
2sn3	65	12.22	11.25	12.22	7.68	6.96
1ctf	68	9.76	12.41	9.36	6.46	6.05
1ubq	76	11.81	14.64	8.97	6.61	6.05
Average:	52.3	9.41	10.16	9.12	5.63	5.87

cRMSD, coordinate root-mean-square deviation; dRMSD, distance root-mean-square deviation; MJ, Miyazawa–Jernigan.
†The cRMSD values for structures with lowest MJ (column 3) and four-body (column 4) energies and highest statistical weight (SW, column 5) are compared; statistical weight is defined in eq. 19. Also shown are the lowest cRMSD values (column 6) and ensemble-averaged dRMSD (column 7) over ensembles of one million configurations. All RMSD values are reported in Å.

vary significantly, depending on the protein size and the class to which they belong, and are more meaningful measures of the overall performance of different potentials. For the MJ potential, the α class is predicted to have an average cRMSD of ~8 Å compared with about 9 Å for the four-body potential. The all-β class proteins yield an average cRMSD of about 9 Å for both potentials. The average number of residues in this set is only 49 compared with 58 for the all α protein set. Thus, there is some deterioration in the MJ and four-body results for all β proteins. The predicted cRMSD values for the mixed α/β class are also poorer than those for the all-α class. Here, the results of the four-body potential show marked deterioration with average cRMSD value of more than 10 Å for an average of 52 residues. On the whole, α proteins are predicted with much lower cRMSD values than the β and mixed classes, which agrees with other studies. For comparison, random conformations for the protein sizes computed have expected cRMSD of about 11 to 12 Å.[37] Therefore, the predicted cRMSD values for all α and β proteins are only about 3 Å better than random, which still is poor, although some predicted proteins have better cRMSD values.

Figures 6 and 7 show eight examples of predicted structures in comparison with their native folds. Figure 6 compares predicted structures for proteins 2mhu and 8rxna with lowest MJ and four-body energies; they are superimposed with their native structures. The four-body potential's prediction for 2mhu compares more favorably with the native structure than the MJ potential, but the reverse is true for protein 8rxna. Predicted structures from both MJ and four-body show rough agreement with the native folds, but significant distortions from the native structures are evident. Figure 7 shows four additional predicted structures (superimposed with native structures) with two each from MJ and four-body potentials. The MJ structures for 1r69 and 4icb have cRMSD values of about 8 Å; the folds show rough overall agreement with the native proteins. However, the helical elements of these α protein are not evident. Structure-derived potentials are not capable of reproducing the secondary structural elements without additional secondary energy biases.[8] The



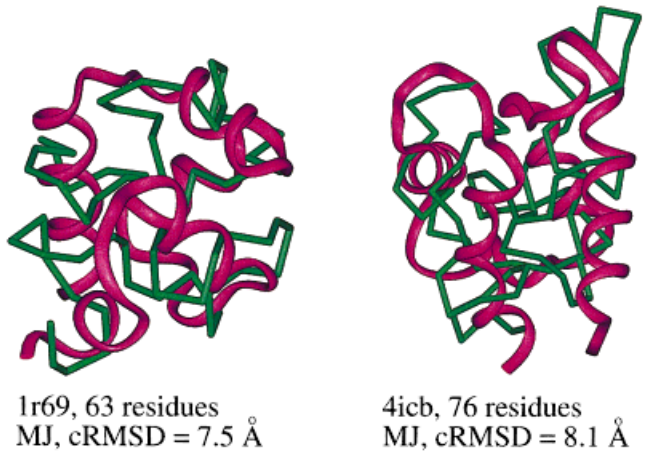
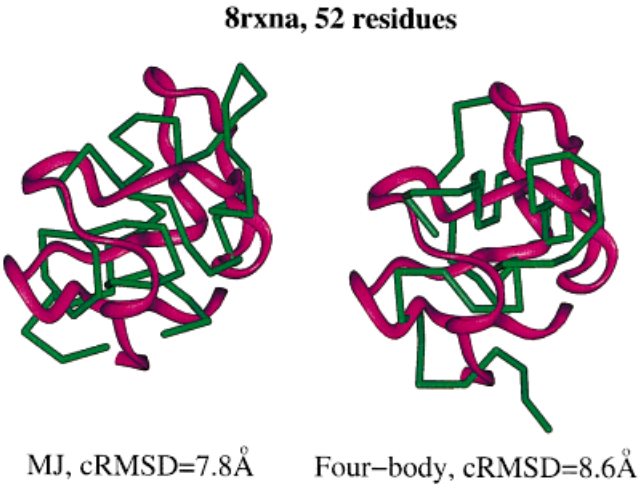


Fig. 6. Three-dimensional structures of proteins 2mhu and 8rxna with the lowest MJ and four-body energies in generated ensembles (green) are superimposed with their native structures (magenta); superimpositions were done with InsightII molecular graphics program. Ensembles of one million configurations were generated to obtain these results.



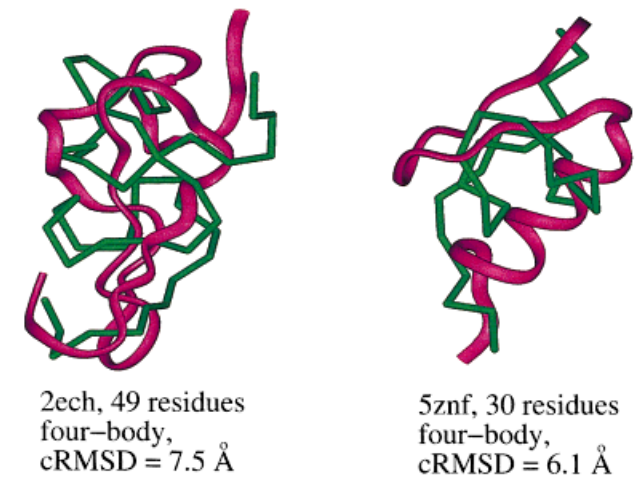


Fig. 7. Three-dimensional structures (in green) of proteins 1r69 and 4icb with the lowest MJ energies and of proteins 2ech and 5znf with the lowest four-body energies; they are superimposed with their native structures (magenta). Ensembles of one million configurations were generated to obtain these results.

four-body structures in Figure 7 show a similar degree of resemblance to their native folds as the MJ structures, although the sequence lengths for the proteins are smaller.

Selection of nativelike structures based on highest statistical weights (SW) in the ensembles yields cRMSD values of just over 9 Å for all three protein classes. In the case of mixed α/β class, the average cRMSD for SW is 9.12 Å, compared with 9.41 Å and 10.16 Å for the structures with lowest MJ and four-body potentials, respectively. In a few cases, the cRMSD values of SW are identical to the MJ results. This means that for certain proteins the states with the lowest MJ energies also contribute the most to the thermodynamic free energy. Inspection of the distribution of statistical weights shows that, for some protein cases, the free energy is dominated by a few low-energy configurations. This is especially true when the temperature is lowered below T_0. The fact that some, or most, configurations may be redundant from the viewpoint of thermodynamics can be a basis for designing more efficient chain growth algorithms.[38]

Since many early as well as current protein structure prediction studies have presented only a few test proteins, it is interesting to determine by how much our results might improve if we select only the five most favorable predictions from the set of 22 proteins for each method. For the MJ results, we could select the protein set {2mhu, 1r69, 4icb, 8rxna, 5znf} (see structures of 1r69 and 4icb in Fig. 7) to yield an average cRMSD of only 6.83 Å (for 50 residues/protein). This result is better by 2 Å than the average cRMSD values for MJ potential in Table II. From the four-body results in Table II, if we choose the set {2mhu, 2cro, 2ech, 8rxna, 5znf} (45 residues on average) we obtain an average cRMSD of 7.28 Å. Again an overall improvement of about 2 Å (see structures of 2ech and 5znf in Fig. 7). Finally, the favorable protein set for the SW results is {2mhu, 1r69, 8rxna, 1ubq, 1sis}, with an average cRMSD of 7.52 Å for 51 residues. Thus, employing a small protein set

can bias the reported cRMSD values, making it difficult to compare the performance of different models and algorithms.

Proteins in solution exhibit some flexibility as revealed by nuclear magnetic resonance (NMR)-determined structures. The accessible conformations have different thermodynamic weights, but the thermodynamic average defines the equilibrium state of a native protein in solution. A measure of the ensemble deviation of the predicted structures from the native state is the thermal-averaged distance RMSD (dRMSD)[17]

$$\langle D^2_{\mathrm{drms}}\rangle_\beta = \sum_\Lambda D^2_{\mathrm{drms}}(\Lambda) W_\Lambda(\beta) / \sum_\Lambda W_\Lambda(\beta) \qquad (14)$$

where the distance RMSD, D_{drms}, is defined as follows:

$$D_{\mathrm{drms}} = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(D_{ij} - D_{ij}^{\mathrm{nat}})^2} \qquad (15)$$

Here, D_{ij} and D_{ij}^{nat} are distances between residues i, j of a generated configuration and the native protein, respectively. All allowed configurations contribute to the average dRMSD, but some configurations contribute more than the others, and this is determined by the statistical weight $W_\Lambda(\beta)$ of the configuration. Unlike the cRMSD values in Table II for specific configurations, this average dRMSD reflects the property of the configurational ensemble or equilibrium configurations. An advantage of the thermal-averaged dRMSD value is that it is not sensitive to statistical fluctuations from one run to another. In other words, the dRMSD values reported in this way are reproducible. A folding algorithm that rarely folds a protein to a low RMSD value does not imply a significant thermodynamic result. The average dRMSD from the MJ potential shown in Table II for the protein classes vary between 5.9–6.4 Å. The cRMSD values (in Table II) are about 30% to 60% larger than the average dRMSD values. Covell[9,10] found smaller dRMSD values in his simulation studies on a simple cubic lattice with an MJ-type potential, but the average cRMSD for a set of eight proteins is about 8.4 Å, which is rather similar to our results.

A Conformational Ensemble from Four-Body Potential

The four-body potential has been used solely as a discriminating function for configurations generated with the two-body potential. As a test case, we generated a configurational ensemble with the four-body potential for the small protein/peptide 1fct (32 residues). For this protein, the CPU time to produce 10^5 configurations is about 2 weeks compared with about 2 h for a million configurations with the MJ potential. The best configuration has a dRMSD of 2.8 Å and the lowest MJ and four-body energy configurations yield comparable dRMSD values of 4.7 Å and 4.5 Å, respectively. Although these predicted configurations have dRMSD values similar to the average dRMSD of the small proteins (30–35 residues) in Table II, the results here are probably not yet converged, since we have only generated 10^5 configurations. Our previous study using the chain growth algorithm showed that protein properties such as thermodynamic functions require millions of configurations to obtain accurate results.[17] Significant computational improvements are needed to produce convergent results with the four-body potential.

SUMMARY AND CONCLUSIONS

Multibody potentials are natural extensions of the commonly used two-body interactions. Our theoretical formulation of multibody (mean) potentials shows that two, three, four-body potentials are related. Since multibody mean potentials are interpreted as the probability of observing clusters of residues in proximity, they yield more information than do two-body potentials regarding correlations in folded protein structures. We showed how a four-body statistical potential is evaluated using the methodology of statistical geometry and its incorporation in a chain growth sampling algorithm for protein structure prediction.

The four-body potential is similar to the two-body counterpart in that cysteine-rich quadruplets are the dominant term in the potential, although the MJ versus four-body energy plots show that they are not necessarily correlated. Our analysis indicates that some terms of the four-body potential may not be sufficiently convergent, given the size of current representative protein structures. We examined the quality of the four-body potential for structure prediction, in comparison with MJ potential, through energy/cRMSD plots and calculated cRMSD values. Our results show that the four-body potential as implemented (to assess the MJ ensemble, rather than to generate one de novo) yields cRMSD values that are about the same as, or slightly poorer than, those from the two-body MJ matrix on a representative set of 22 proteins. Finally, both the two- and four-body statistical potentials employed are not sufficiently discriminating since the native structures have less favorable energies than the non-native states (Fig. 4). Thus, not only must the quality of the four-body potential be improved, statistical potentials beyond the C_α interaction site models are likely required to successfully discriminate native from nonnative structures generated on a lattice.

This failure of the four-body potential can be attributed to several factors. First, its construction requires a much larger set of representative native proteins than two-body potentials. Our use of only six residue types makes implementation tractable but not necessarily accurate. As the number of solved protein structures increases, this derivation difficulty may be alleviated. Second, the generation of configurational ensembles with the four-body potential is prohibitively costly due to the necessity to perform Delaunay tessellation at each step of the chain growth algorithm. Our use of the four-body potential as a postprocessing tool—to screen the configurations generated using the two-body MJ potential—may therefore not judge the four-body potential most favorably. With the MJ potential, the lowest cRMSD value attained in the generated ensembles

is still not accurate, about 5–6 Å (Table II). More algorithmic work is needed to make direct ensemble generation via the four-body potential feasible. For example, by allowing local tessellation as the chain grows, we should reduce the computational time substantially.

The four-body potential can also be improved by using smaller (closer to physical range for) cutoff length R_{cut}, relaxing the assumption about the equivalence of four-body terms $Q^{\alpha}_{ijkl} = \ldots = Q^{\alpha}_{ikji}$, and incorporating geometric features of four-body clusters in native proteins. In addition, the role of potentials with longer-range than our contact potentials should be examined. These improvements, however, rely on availability of a larger representative protein database. Our approximate implementation of the four-body potential, as discussed here and in the preceding paragraph, biased our conclusions to some extent about the relative performance of the two- and four-body potentials. By lifting some of the approximations in future studies, we will be able to make a sharper distinction between them.

Our simplified C_α model of proteins ignores vital structural details of proteins. Certainly, explicit modeling of side chains is required to describe the regularity in residue–residue packing configurations of helices and β-sheets. Such a refinement can be implemented with lattice protein models.[8] Thus, our current C_α model is inherently of low resolution as evident from our best structures (5–6 Å cRMSD). The similarity between the two- and four-body results partly reflects this limited resolution of our C_α protein model. Future improvements will undoubtedly be realizable by modeling side chain interaction centers, as well as reconstructing all-atom models from low-resolution protein configurations.[39]

The importance of multibody potentials in structure prediction has been recognized by computational structural biologists.[4,8,11,21] Further work in this area should improve our understanding of the role of multibody potentials in folded protein structures. Recent developments in nonhomology-based methods in protein structure prediction have incorporated knowledge of secondary structure[40] and tertiary restraints.[41] These methods are quite successful in producing nativelike conformations with cRMSD of about ≤6 Å for a number of small proteins. It is far more difficult to predict folds well de novo based on physical principles, such as the method employed in this work.

Analysis of the results reported at the Third Critical Assessment of Techniques for Structure Prediction (CASP3) showed that most ab initio predictions gave cRMSD values of ~10 Å for a set CASP3 proteins,[42] although a few significantly better predictions were reported.[43] Despite this low accuracy attainable to date, the rapid increase in the number of protein structures should improve the prospect of the ab initio approach to structure prediction in the near future. Results of the CASP3 meeting (http://predictioncenter.llnl.gov/casp4/Casp4.html) indicated considerable progress in fully automated approaches for structure prediction. Significant improvements in the prediction of small proteins with no known structures were reported. Thus, if modeling improvements continue steadily, and targets for modeling are selected carefully, computer modeling might become an important resource for protein structure prediction. In the genomic era, the need for large-scale structural assignment of protein sequences ensures continued importance of computer modeling.

MATERIALS AND METHODS

Chain Geometry and Simulation Lattice

We now define the protein model, simulation lattice and the chain growth move directions on the lattice. Since our model has been presented recently,[17] we only summarize the procedure here. We use a C_α representation for protein chains with no side-chains. Geometric characteristics of the protein backbones are reproduced by limiting the C_α pseudo-bond angles to the range of 63–143° and associating each C_α vertex with a set of excluded-volume sites (defined below). Protein configurations generated on the lattice must satisfy these geometric constraints.

Our (311) lattice model is similar to the family of refined cubic lattices investigated by Kolinski and Skolnick.[8] On this lattice, possible growth (move) directions are given by the vectors $\mathbf{v} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, where $(x, y, z) \in \{(\pm 1, \pm 3, \pm 1), (\pm 3, \pm 1, \pm 1), (\pm 1, \pm 1, \pm 3)\}$. In the chain growth process, many of these move directions are prohibited by the pseudo-bond angle restrictions and the excluded-volume requirements. The 26 excluded volume sites associated with each C_α vertex are separated from it by the vectors $\{(\pm 1, \pm 1, \pm 1), (\pm 1, \pm 1, 0), (\pm 1, 0, \pm 1), (0, \pm 1, \pm 1), (\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)\}$. Here, the lattice spacings are measured in units of $L = 3.8/\sqrt{x^2 + y^2 + z^2}$ Å = 1.146 Å, where the C_α pseudo-bond is 3.8 Å. As found empirically, this lattice leads to lattice-mapped structures of native proteins with cRMSD of about 1.5 Å, which is sufficient to reproduce elements of the secondary structures and overall protein folds.

Structure Generation by Chain Growth Algorithm

Overview

The chain growth algorithm is conceptually distinct from current approaches to protein folding based on Metropolis algorithms[44] and the multicanonical[45] or entropy[46] sampling. We recently applied this algorithm to lattice proteins and showed that it is an effective approach to thermodynamics and generation of lattice protein structures. Here, we summarize the methodology detailed earlier.[17]

Essentially, the chain growth algorithm generates chain configurations by sequential addition of links until the full length of the chain is reached. Since each configuration is generated de novo, configurations are statistically independent. This approach differs markedly from the standard Metropolis algorithm, in which the successive configurations are linked by a prescribed set of perturbation moves. The chain growth process is guided by a temperature-dependent transition probability which is a normalized Boltzmann factor.[47–49] The Boltzmann-weighted transition probability tends toward growth directions with favorable contacts, thereby allowing generation of both open and compact chain configurations depending on the tem-

perature. This is an extension of the original algorithm for self-avoiding (athermal) chains by Rosenbluth and Rosenbluth.[50] The chain growth algorithm has been applied successfully to simple polymer and peptide systems.[47–49] A large number of chain growth configurations is generated to estimate thermal averages which are calculated using an importance sampling procedure whereby each contributing configuration is appropriately weighted.[48]

Chain generation, ensemble averaging, and convergence

We generate chains on (311) cubic lattice guided by the following temperature-dependent transition probability[17]

$$P_i(\mathbf{R}_i + \mathbf{v}_{k_i}|\mathbf{R}_1, \ldots, \mathbf{R}_i; \beta) = \exp[-\beta u_i(\mathbf{R}_i + \mathbf{v}_{k_i})] / \sum_{k_i=1}^{C_i} \exp[-\beta u_i(\mathbf{R}_i + \mathbf{v}_{k_i})] \quad (16)$$

where the incremental, nonbonded potential energy is

$$u_i(\mathbf{R}_i + \mathbf{v}_{k_i}) = \sum_{j=1}^{i-1} u_{ij}(R_{ij}) \quad (17)$$

Other symbols are defined as follows: $\mathbf{R}_1, \ldots, \mathbf{R}_i$ are position vectors of (interaction) sites $1, 2, \ldots, i$; $\mathbf{v}_{k_i}$ is the lattice vector for the chosen direction k_i; and C_i is the number of vacant sites at step i. The first link can be placed in any direction, and the move directions for subsequent links are selected according to the transition probability by a Monte Carlo procedure. At each step, the potential energies, $u_i(\mathbf{R}_i + \mathbf{v}_{k_i})$, for all allowed lattice growth directions are determined and their transition probabilities evaluated. The growth directions must satisfy the prescribed range of pseudo-bond angles and excluded volume requirements. For compact configurations, the number of allowed moves is much less than the maximum of 26. This growth process is continued until the entire chain length (N links) is exhausted. When a dead-end configuration, i.e., $C_i = 0$, is encountered before the chain is fully grown, the growth process is terminated and the chain discarded, the next chain is then regrown from scratch. For our simple C_α model on (311) lattice, more than 90% of configurations are successfully grown. We generate about a million configurations to obtain accurate estimates of thermal averages. More details on chain generation procedure are described by Gan et al.[17]

From the chain growth configurations, the average of a property A in canonical ensemble is given by

$$\langle A \rangle_\beta = \sum_{\{\Lambda\}} A_\Lambda W_\Lambda(\beta) / \sum_{\{\Lambda\}} W_\Lambda(\beta) \quad (18)$$

where A_Λ is the value of property A for configuration Λ and the statistical weight

$$W_\Lambda(\beta) = \prod_{i=1}^{N} \{ \sum_{k_i=1}^{C_i} \exp[-\beta u_i(\mathbf{R}_i + \mathbf{v}_{k_i})] \} \quad (19)$$

We have used an importance sampling procedure to obtain the above statistical average, which ensures that the (biased) chain growth configurations are assigned appropriate weights.[48] All successfully grown configurations are counted, each with a statistical weight W_Λ. To ensure accuracy and convergence of the average $\langle A \rangle_\beta$, the size of the configurational sample must be sufficiently large that equilibrium configurations belong to the sample.

The convergence of the average property $\langle A \rangle_\beta$ depends on the size of the configurational ensemble $\mathcal{N}$, the number of links b placed at each step of the growth process, and the lattice coordination number n_c. Since the chain growth algorithm explores all available growth directions at each step, the number of energy evaluations per step is proportional to $(n_c)^b$. For our (311) lattice with $n_c = 24$, $(n_c)^b$ is large even for small values of b. We thus choose $b = 1$, but compensate by generating a large sample size on the order of 1 million. Larger b values would mean reducing the sample size $\mathcal{N}$. This could in turn render the importance sampling approach to computing thermal averages ineffective.

In this and our previous work,[17] we have demonstrated convergence of thermodynamic and configurational properties in our implementation of the chain growth algorithm. Specifically, we found that more than 90% of the chain configurations are successfully grown. Therefore, the algorithm is considered adequate for the simple model used. Greater efficiency in the chain growth procedure can be achieved by optimizing the value of b (number of links grown at each step), and the sample size. Moreover, the possibility of using more efficient chain growth algorithms, as developed recently, might be considered.[38] Other efficient Monte Carlo algorithms, such as entropy sampling[46] and multicanonical[45] algorithms may be viable alternatives for more complex protein models.

ACKNOWLEDGMENTS

We thank Stephen Cammer for help with Figure 1. This work was supported by NIH grant GM55164, NSF grants BIR-94-23827EQ and ASC-9704681, and a John Simon Guggenheim fellowship (to T.S.). A.T. acknowledges receiving support from NIH Research Resource grant RR08102. T.S. is an investigator of the Howard Hughes Medical Institute.

REFERENCES

1. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
2. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256: 623–644.
3. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213: 859–883.
4. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem 1997;18:849–872.
5. Bryant SH, Altschul SF. Statistics of sequence-structure threading. Curr Opin Struct Biol 1995;5:236–244.

6. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
7. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 1998; 95:13597–13602.
8. Kolinski A, Skolnick J. Lattice models of protein folding, dynamics, and thermodynamics. Austin, TX: Landes; 1996.
9. Covell DG. Folding protein α-carbon chains into compact forms by Monte Carlo methods. Proteins 1992;14:409–420.
10. Covell DG. Lattice model simulations of polypeptide chain folding. J Mol Biol 1994;235:1032–1043.
11. Liwo A, Kazmierkiewicz R, Czaplewski C, et al. United-residue force field for off-lattice protein-structure simulations. III. Origin of backbone hydrogen-bonding cooperativity in united-residue potentials. J Comput Chem 1998;19:259–276.
12. Hill T. Statistical mechanics: principles and selected applications. New York: McGraw-Hill; 1956.
13. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
14. Tropsha A, Singh RK, Vaisman II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. In: Hunter L, Klein TE, editors. Pacific Symposium Biocomputing. Singapore: World Scientific; 1996. p 614–623.
15. Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. In: Hunter L, Klein TE, editors. Pacific Symposium Biocomputing. Singapore: World Scientific; 1997. p 486–497.
16. Munson PJ, Singh RJ. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci 1997;6: 1467–1481.
17. Gan HH, Tropsha A, Schlick T. Generating folded protein structures with a lattice chain growth algorithm. J Chem Phys 2000;113:5511–5524.
18. Thomas PD, Dill K. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
19. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
20. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
21. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;9:361–369.
22. Mulmuley K. Computational geometry: an introduction through randomized algorithm. New York: Prentice-Hall; 1994.
23. Bernal JD. A geometric approach to the structure of liquids. Nature 1959;183:141–147.
24. Richards FM. Areas, volumes, packing, and protein structures. Annu Rev Biophys Bioeng 1977;6:151–176.
25. Voronoi and Delaunay tessellation programs are available at http://www.geom.umn.edu/software/qhull which is located at the Geometry Center, University of Minnesota.
26. Seidel R. Exact upper bounds for the number of faces in d-dimensional Voronoi diagrams. In: Gritzmann P, Sturmfels B, editors. Applied geometry and discreet mathematics—the Victor Klee Festschrift, DIMACS Series in Discreet Mathematical and Theoretical Computer Science Science. American Mathematics Society, Providence, RI; 1991. p 517–529; see also K. Fukuda, Polyhedral computation, http://www.ifor.math.ethz.ch/~fukuda/index.html.
27. Representative protein lists are available at http://www.sander-.embl-heidelberg.de/pdbsel/ and the algorithm for generating the lists is described in Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
28. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
29. George DG, Hunt LT, Barker WC. Computer methods in sequence comparison and analysis. In: Schlesinger DH, editor. Macromolecular sequencing and synthesis: selected methods and applications. New York: Liss; 1988. p 127–149.
30. White SH, Jacobs RE. Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. Biophys J 1990;57:911–921.
31. Hinds D, Levitt M. A lattice model for protein structure prediction at low resolution. Proc Natl Acad Sci USA 1992;89:2536–2540.
32. Park BH, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–392.
33. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
34. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
35. Daura X, Juan B, Seebach D, van Gunsteren WF, Mark A. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280:925–932.
36. Lazaridis T, Karplus M. Discrimination of the native from the misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
37. Cohen FE, Sternberg MJE. On the prediction of protein structure: the significance of the root-mean-square deviation. J Mol Biol 1980;138:321–333.
38. Grassberger P. Pruned-enriched Rosenbluth method: simulation of θ polymers of chain length up to 1000000. Phys Rev E 1997;56:3682–3693.
39. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CI III. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.
40. Eyrich VA, Standley DM, Felts AK, Friesner RA. Protein tertiary structure prediction using a branch-and-bound algorithm. Proteins 1999;35:41–57.
41. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.
42. The CASP3 target proteins and analysis of all results are available at http://predictioncenter.llnl.gov/casp3/Casp3.html
43. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;3(suppl):149–170.
44. Skolnick J, Kolinski A. Monte Carlo approaches to the protein folding problem. Adv Chem Phys 1999;105:203–242.
45. Hansmann U, Okamoto Y. Monte Carlo algorithms for protein folding. Curr Opin Struct Biol 1999;9:177–183.
46. Hao MH, Scheraga HA. Monte Carlo simulation of a first-order transition for protein folding. J Phys Chem 1994;98:4940–4948.
47. Meirovitch H. Improved computer simulation method for estimating the entropy of macromolecules with hard-core potential. Macromolecules 1983;16:1628–1631.
48. Meirovitch H, Vasquez M, Scheraga HA. Stability of polypeptide conformational states. II. Folding of a polypeptide chain by the scanning simulation method, and calculation of the free energy of the statistical coil. Biopolymers 1988;27:1189–1204.
49. Meirovitch H. Statistical properties of the scanning simulation method for polymer chains. J Chem Phys 1988;89:2514–2522.
50. Rosenbluth MN, Rosenbluth AV. Monte Carlo calculation of the average extension of molecular chains. J Chem Phys 1995;23:356–359.