Optimizing Physical Energy Functions for Protein Folding

Yoshimi Fujitsuka,[1] Shoji Takada,[1,2*] Zaida A. Luthey-Schulten,[3] and Peter G. Wolynes[4]
[1] *Graduate School of Science and Technology, Kobe University, Nada, Kobe, Japan*
[2] *PRESTO, Japan Science and Technology Corporation, Japan*
[3] *Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois*
[4] *Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California*

ABSTRACT **We optimize a physical energy function for proteins with the use of the available structural database and perform three benchmark tests of the performance: (1) recognition of native structures in the background of predefined decoy sets of Levitt, (2) de novo structure prediction using fragment assembly sampling, and (3) molecular dynamics simulations. The energy parameter optimization is based on the energy landscape theory and uses a Monte Carlo search to find a set of parameters that seeks the largest ratio $\delta E_s/\Delta E$ for all proteins in a training set simultaneously. Here, δE_s is the stability gap between the native and the average in the denatured states and ΔE is the energy fluctuation among these states. Some of the energy parameters optimized are found to show significant correlation with experimentally observed quantities: (1) In the recognition test, the optimized function assigns the lowest energy to either the native or a near-native structure among many decoy structures for all the proteins studied. (2) Structure prediction with the fragment assembly sampling gives structure models with root mean square deviation less than 6 Å in one of the top five cluster centers for five of six proteins studied. (3) Structure prediction using molecular dynamics simulation gives poorer performance, implying the importance of having a more precise description of local structures. The physical energy function solely inferred from a structural database neither utilizes sequence information from the family of the target nor the outcome of the secondary structure prediction but can produce the correct native fold for many small proteins. Proteins 2004; 54:88–103.** © 2003 Wiley-Liss, Inc.

Key words: physical energy functions; folding; Monte Carlo search; structure prediction

INTRODUCTION

An accurate yet efficiently evaluatable energy function is a key prerequisite for virtually any quantitative study on proteins. For protein structure prediction, quasienergy functions are often used that take into account both physical energies and bioinformatic terms derived from homologous sequences. Many other problems, however, require the energy functions that are entirely physics based for rational protein engineering and addressing what kinds of physical forces are most responsible for proteins being thermodynamically stable in the native state and being kinetically accessible to the native state within a biologically feasible timescale in the laboratory. Although there has been much effort to develop molecular mechanics energy functions, their accuracy and speed are not yet sufficient for many uses. Determination of potential parameters is conceptually and practically difficult. Although it would seem the most deductive and logical, determining potential parameters solely from electronic structure calculation of small molecules does not necessarily give the best performance for modeling proteins in solvent. Instead of this bottom-up approach, we might ask whether we can infer physical forces from their consequences, that is, the structures of proteins already in the structural database. Recently, structural genomics projects have started to produce thousands of 3D protein structures. If we can use this information to improve protein energy functions, this would yield energy functions that are practically powerful for many purposes and should be conceptually helpful for gaining insight into the physical principles of protein architecture.

For our present purpose, we need a theory of protein structure and folding that gives a quantitative guiding principle to infer such an energy function from structures. Here, we rely on the energy landscape theory of protein folding.[1,2] Briefly, this theory argues that proteins are thermodynamically stable below the denaturation temperature T_F, whereas protein motions are so slow that they would be trapped to one of misfolded structures below the glass transition temperature T_G. Thus, it is only in the temperature range $T_F > T > T_G$ that proteins are both thermodynamically stable and can kinetically access the native structure. Through evolution, protein sequences have acquired the property that T_F/T_G is sufficiently greater than unity so that there will be a reasonably wide temperature range that satisfies this condition. This requirement is termed the minimal frustration principle.

*Correspondence to: Shoji Takada, Graduate School of Science and Technology, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan. E-mail: stakada@kobe-u.ac.jp

Received 2 September 2002; Accepted 17 January 2003

Within a simple random energy model, T_F/T_G is shown to be proportional to the ratio of the energy gap $\delta E_s = |E_N - E_D|$ between the native state E_N and the denatured state E_D over the energy fluctuation in the denatured ensemble $\Delta E = \sqrt{\langle (E_D - \langle E_D \rangle)^2 \rangle_D}$, where the bracket stands for an ensemble average. Intuitively, a protein has a globally funnel-like energy landscape with local ruggedness on the slope where the height of the funnel corresponds to δE_s, whereas the ruggedness is of order ΔE. The large ratio T_F/T_G corresponds to the large ratio $\delta E_s/\Delta E$, which means the folding funnel is relatively smooth.

The energy landscape theory gives not only a basic perspective to think about proteins but through its mathematical formulation provides a practical framework for optimizing potential energy functions. Namely, we seek a set of energy parameters that give the largest $\delta E_s/\Delta E$ for a set of proteins. This idea was first employed and has been developed with the associative memory Hamiltonian, giving promising performance for protein structure prediction.[3–5] Scheraga and coworkers also used this optimization idea to determine the balance of several parameters in their physicochemical model.[6] The objective function by taking the ratio of δE_s to ΔE compares the native structure with an ensemble of decoys, nonnative structures. As in other pattern recognition problems, one seeks the energy function that can best discriminate the true pattern from the false patterns provided by the decoys. It is apparent from this analogy that the quality of the false patterns, decoys used in the estimate of $\delta E_s/\Delta E$, is crucial for obtaining better energy parameters. Originally, a simple threading approach was taken for producing decoys, and later studies improved the quality of decoys by performing folding simulations[4] so that they discriminate against the most likely misfolded structures to be created by the algorithm itself.

In this article, we employ an optimization scheme similar to the earlier ones to improve a physical energy function that we have been developing.[7,8] The energy function uses a coarse-grained protein-chain representation with an implicit solvent model. Its functional form is well based on physicochemical considerations so that each energetic term can be interpreted as a physical force. Because of the coarse graining, however, the large number of parameters included in the energy function cannot be easily determined with precision in a bottom-up manner. To determine these parameters, we use a pattern recognition approach based on the energy landscape theory to give an optimized energy parameter set. The present approach differs from Scheraga's in that many parameters are determined rather than just a few overall weights of different types of forces. We believe, therefore, the optimized physical potential can give us a deeper understanding of the role specific amino acid types play in protein architecture from a semimicroscopic perspective. Nevertheless, because so much physics goes into the model we believe this physical potential can be used to study even nonnatural sequences in comparison to score functions derived from bioinformatic analysis alone.[9,10]

This article is organized as follows: The next section describes the coarse-grained protein model and potential energy function in detail. We emphasize the physical motivation of individual terms; in particular the origins of solvent-induced multibody effects and cooperative effects among hydrogen bond network are explained. Then, we present algorithms for optimizing potential energy parameters using the available protein structural database. We review the energy landscape theory of proteins needed to explain how physical interactions in proteins can be inferred from structural information. We then compare the optimized physical interaction terms with the experimentally determined values, where possible. Next, we present three sets of benchmark tests of the optimized energy function. First, we perform a recognition test of the native structures against predefined set of decoys provided by Levitt and collaborators.[11] Then, the fragment assembly Monte Carlo conformation sampling developed by Baker et al. and the replica exchange molecular dynamics simulation are employed to give de novo protein structure predictions for several small proteins.[12,13] The results of these two algorithms using the same energy function are compared. Discussions and conclusions are given.

POTENTIAL ENERGY FUNCTION

Coarse-Grained Chain Representation

We describe here the coarse-grained model of proteins we use.[7,8] Each amino acid is represented by coordinates of amide N, C_α, carbonyl C in the main-chain, and by those of a centroid of side-chain atoms (GLY does not have the side-chain). The main-chain representation is basically the same as in our previous publications using a physical energy function[7,8]. Bond lengths and bond—bond angles for the main-chain atoms are fixed to their standard values. The side-chain and some of the main-chain representations are renewed. In the hydrogen bond calculation described below, the amide H atom and carbonyl O atom coordinates are derived from those of other main-chain atoms. Depending on the side-chain conformations, the side-chain centroid can take on several discrete positions relative to its main-chain placement. Each of the centroid positions corresponds to the center of mass of nonhydrogen atoms in the side-chain of a rotamer conformation given in Dunbrack's rotamer library.[14] We use the rotamers that have relatively high probabilities in the Dunbrack backbone-independent rotamer library.[14]

Potential Energy Function

The energy function we use here has evolved from our previous work.[7,8] Although many ingredients are common with those in the previous work, some new features are introduced here. In particular, (1) the hydrogen bond energy terms takes into account cooperative effects between two neighboring hydrogen bonds that turned out to play crucial roles for stabilizing β-sheets and destabilizing α-helices, (2) the hydrogen bond energy function now explicitly uses coordinates of the amide H atom and the carbonyl O atom, and (3) many of the precise functional forms are changed slightly based on survey in structural database. For the latter, we use a protein structure set used in the classic Miyazawa–Jernigan article.[15] The third point is important because energetic parameters can be readily optimized by the scheme described below, while structural parameters such as length scales are not so easily optimized and the quality of optimization in energetic parameters is highly dependent on the structural parameters used in the model potential.

We designed the energy function in terms of torsion angles potentials (V_ω, V_ϕ, V_ψ), van der Waals interactions (V_{vdW}), hydrogen bond (V_{HB}), hydrophobicity (V_{HP}), secondary structure propensity (V_{Rama}), and pair-wise specific interactions ($V_{\text{pair-wise}}$):

$$V = V_{\varpi} + V_{\phi} + V_{\psi} + V_{\mathrm{vdW}} + V_{\mathrm{HB}} + V_{\mathrm{Rama}} + V_{\mathrm{pairwise}}. \tag{1}$$

The major sequence dependence of the energy function is in the latter three terms although some minor sequence information is included in the fourth and fifth terms as well.

The torsion angle potential V_ω keeps the torsion angle around the peptide bond close to a *trans* or *cis* conformation:

$$V_\omega = \sum_{I=1}^{Naa-1} \varepsilon_\omega (1 - \cos^2 \omega_I), \tag{2}$$

where ω_1 is the dihedral angle around the peptide bond between the Ith and I + 1th amino acid and N_{aa} is the number of residues. The strength of potential ϵ_ω is large so that no *trans–cis* transition will be realized in molecular dynamics (MD) simulations, while in the fragment assembly method such a transition can occur at the peptide bond preceding a proline residue. The other torsion angle potentials V_ϕ and V_ψ, which are the same as in our previous work, give weak hinderance of rotation for dihedral angles ϕ and ψ, respectively,

$$V_\phi = \sum_{I=2}^{Naa} \frac{1}{2} \varepsilon_\phi (1 + \cos 3\phi_I), \tag{3}$$

and

$$V_\psi = \sum_{I=1}^{Naa-1} \frac{1}{2} \varepsilon_\psi (1 + \cos 3\psi_I), \tag{4}$$

where ϵ_ϕ and ϵ_ψ are the rotation barrier height between *gauche* and *trans* or between *gauches*.

The vdW interaction V_{vdW} is separated into the 1-4 local interaction part $V_{\mathrm{vdW-lc}}$ and the nonlocal interaction part $V_{\mathrm{vdW-nonlc}}$, as in usual force fields. For the 1-4 part, its functional form strongly influences the local propensity and stiffness of the polypeptide. Although the Lennard–Jones (LJ) potential is standard in any atomic force field, it is not necessarily appropriate for the coarse-grained chain model. We first analyzed the probability distribution of the distance r_{ij} between the i and j groups in the protein structural dataset, the protein set used in the Miyazawa–Jernigan article.[15] We found that for the interactions between backbone atoms the LJ potential is suitable, while a softer functional form fits better the interaction between backbone atoms and side-chain centroids except prolines. Accordingly, we prepared two different functional forms:

$$V_{\text{vdW-lc}} = \varepsilon_{\text{vdW-lc}} \sum_{\text{1–4pairs,bb}} \left[\left(\frac{a_{\text{vdW-lc},ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{a_{\text{vdW-lc},ij}}{r_{ij}} \right)^{6} \right] + \varepsilon_{\text{vdW-lc}} \sum_{\text{1–4pairs,bs}} \left[\left(\frac{a_{\text{vdW-lc},ij}}{r_{ij}} \right)^{8} - \left(\frac{a_{\text{vdW-lc},ij}}{r_{ij}} \right)^{4} \right], \tag{5}$$

where the first sum is for pairs of backbone atoms and also the pairs between backbone atoms and the proline side-chain centroid, while the second sum is for the pairs of side-chain centroids of the nonproline residue and backbone atoms. Here, $\epsilon_{\mathrm{vdW-lc}}$ is the interaction strength and $a_{\mathrm{vdW-lc},ij}$ determines the length scale of the interaction. The latter value is taken from the probability distribution $P(r_{ij})$ of the structural data set; $a_{\mathrm{vdW-lc},ij} = 0.935\, \tilde{a}_{\mathrm{vdW-lc},ij}$, where $P(r_{ij} \le \tilde{a}_{\mathrm{vdW-lc},ij}) = 0.05$.

For the nonlocal vdW interactions, we also investigated the distance distribution in the structural data set. We see for many pairs a sharp steric repulsion at short distance and a broad attractive well at larger distance, which cannot be represented by a single (inverse) polynomial function. We thus split the attractive part from the pure repulsive part, $V_{\mathrm{vdW-nonlc}} = V_{\mathrm{vdW-att}} + V_{\mathrm{vdW-rep}}$. A generic attraction term $V_{\mathrm{vdW-att}}$ is written with the Gaussian function

$$V_{\text{vdW-att}} = \sum_{ij \in \text{nonlc}} \varepsilon_{\text{vdW-att}} \exp[-\beta_{\text{att},ij}(r_{ij} - a_{\text{att},ij})^2], \tag{6}$$

that are included for the pairs of side-chain centroids and the pairs between backbone atoms and side-chain centroids. $a_{\mathrm{vdW},ij}$ is the distance that vdW interaction are the most stabilized. The steric repulsion term $V_{\mathrm{vdW-rep}}$, is defined by

$$V_{\text{vdW-rep}} = \sum_{ij \in \text{nonlc}} \varepsilon_{\text{vdW-rep}} \times \begin{cases} 4\left\{ \left(\frac{a_{\mathrm{vdW},ij}}{r_{ij}} \right)^4 - \left(\frac{a_{\mathrm{vdW},ij}}{r_{ij}} \right)^2 \right\} + 1 & r_{ij} \le \sqrt{2} a_{\mathrm{vdW},ij} \\ 0 & r_{ij} > \sqrt{2} a_{\mathrm{vdW},ij}. \end{cases} \tag{7}$$

Note that this has a functional form that is less sharp than that for local 1-4 pairs. In the same way as 1-4 pairs, the length scale is determined from the distance distribution; $a_{\mathrm{vdW-lc},ij} = 0.90 \tilde{a}_{\mathrm{vdW-lc},ij}$, where $P(r_{ij} \le \tilde{a}_{\mathrm{vdW-lc},ij}) = 0.05$.

The hydrogen bonding terms are composed of the direct hydrogen bond interactions $V_{\mathrm{HB,\ direct}} + V_{\mathrm{HB,\ direct,4body}}$ between main-chain atoms and the Born energy $V_{\mathrm{HB,\ Born}}$,

$$V_{\mathrm{HB}} = V_{\mathrm{HB,direct}} + V_{\mathrm{HB,direct,4body}} + V_{\mathrm{HB,Born}}. \tag{8}$$

The direct interaction, similar to our previous version, is written as

$$V_{\mathrm{HB,direct}} = \varepsilon_{\mathrm{HB}} \sum_{ij(\mathrm{s.t.}I \geq J+3)} S_{\mathrm{burial},IJ} u_{\mathrm{HB}}^{(\mathrm{a,r})}(r_{ij}), \tag{9}$$

where ϵ_{HB} is the constant hydrogen bond strength and S_{burial} a measure of the buriedness of the hydrogen bond, which takes unity for completely buried bonds and is close to zero for the bond at surface of the protein. $S_{\mathrm{burial},IJ} = (S_{\mathrm{burial},I} + S_{\mathrm{burial},J})/2$, in which

$$S_{\mathrm{burial},I}(\rho_i) = \begin{cases} 0 & \rho_i < \rho_{\mathrm{burial,min}} \\ \frac{1}{2}\left(1 + \cos\left(\pi \frac{\rho_{\mathrm{burial,max}} - \rho_i}{\rho_{\mathrm{burial,max}} - \rho_{\mathrm{burial,min}}}\right)\right) & \rho_{\mathrm{burial,min}} \leq \rho_i \leq \rho_{\mathrm{burial,max}} \\ 1 & \rho_i > \rho_{\mathrm{burial,max}}, \end{cases} \tag{10}$$

where ρ_i stands for the local peptide density around the ith C_α atom, whose explicit form is defined in the next paragraph. The function u_{HB} is the actual distance-dependent part. The superscript a and r represent attraction and repulsion, respectively. As in our previous work, combining attractive and repulsive forces, we make the hydrogen bonds anisotropic.[7] One crucial improvement is that, in the attractive part, now we use the explicit coordinates of amide hydrogen and carbonyl oxygen for expressing the bond interaction. The coordinates of hydrogen and oxygen are computed from the neighboring backbone atom coordinates:

$$\mathbf{r}_{\mathrm{H},I} = \chi_1 \mathbf{r}_{\mathrm{N},I} + \chi_2 \mathbf{r}_{\mathrm{C},I-1} + \chi_3 \mathbf{r}_{\mathrm{C}_\alpha}, \tag{11}$$

and

$$\mathbf{r}_{\mathrm{O},I} = \eta_1 \mathbf{r}_{\mathrm{C}',I} + \eta_2 \mathbf{r}_{\mathrm{N},I+1} + \eta_3 \mathbf{r}_{\mathrm{C}_\alpha}, \tag{12}$$

where $\chi_1 = 2.505$, $\chi_2 = -0.825$, $\chi_3 = -0.680$, $\eta_1 = 2.439$, $\eta_2 = -0.683$, and $\eta_3 = -0.756$. Using these coordinates, we define $u_{\mathrm{HB}}{}^{(\mathrm{a})}$ as

$$u_{\mathrm{HB}}^{(\mathrm{a})} = \varepsilon_{\mathrm{HB}}\left[\left(\frac{\sigma_{\mathrm{OH}}}{r_{\mathrm{OH}}}\right)^{12} - 2\left(\frac{\sigma_{\mathrm{OH}}}{r_{\mathrm{OH}}}\right)^{6}\right], \tag{13}$$

where σ_{OH} is the potential minimum. Repulsive terms are introduced between N–N, N–Cα, Cα–C, and C–C pairs with the functional form

$$u_{\mathrm{HB}}^{(r)} = \varepsilon_{\mathrm{HB}} \begin{cases} 4\left\{\left(\frac{\sigma_{ij}}{r_{ij}}\right)^4 - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^2\right\} + 1 & r_{ij} \leq \sqrt{2}\sigma_{ij} \\ 0 & r_{ij} > \sqrt{2}\sigma_{ij}. \end{cases} \tag{14}$$

The four-body interaction takes into account cooperativity of neighboring two hydrogen bonds and is written as

$$V_{\mathrm{HB,direct,4body}} = -\varepsilon_{\mathrm{HB}} \sum_{ij(\mathrm{s.t.}I \geq J+3),kl(\mathrm{s.t.}I \geq J+3),x} \times c_{\mathrm{HB,4body},x} \min(0, u_{\mathrm{HB}}^{(\mathrm{a})}(r_{ij})) \min(0, u_{\mathrm{HB}}^{(\mathrm{a})}(r_{kl})). \tag{15}$$

We consider three different types of cooperativity related to three types of secondary structures: α-helix, antiparallel β-sheet, and parallel β-sheet. For the α-helix, we take into account couplings between double donor and double acceptor bonds, which should have negative cooperativity; $x = \alpha$ helix corresponds to the quartets; $\{i,j;k,l\} = \{\mathrm{CO}(I), \mathrm{NH}(I+3); \mathrm{CO}(I), \mathrm{NH}(I+4)\}$, $\{\mathrm{CO}(I), \mathrm{NH}(I+4); \mathrm{CO}(I), \mathrm{NH}(I+5)\}$, $\{\mathrm{CO}(I), \mathrm{NH}(I+4); \mathrm{CO}(I+1), \mathrm{NH}(I+4)\}$, and $\{\mathrm{CO}(I), \mathrm{NH}(I+5); \mathrm{CO}(I+1), \mathrm{NH}(I+5)\}$. For hydrogen bonds that contribute to β-sheets, we introduce a positive cooperativity that is induced by desolvation effects.[16] In particular, x = antiparallel β-sheet corresponds to $\{i,j;k,l\} = \{\mathrm{NH}(I), \mathrm{CO}(J); \mathrm{CO}(I), \mathrm{NH}(J)\}$ and x = parallel β-sheet corresponds to $\{i,j;k,l\} = \{\mathrm{NH}(I), \mathrm{CO}(J); \mathrm{CO(I)}, \mathrm{NH}(J+2)\}$. Finally, the Born self-energy term represents the free energy cost upon burial of charged groups, amide group, and carbonyl group in backbone, and is expressed as before:

$$V_{\mathrm{HB,Born}} = \varepsilon_{\mathrm{HB}} c_{\mathrm{HB,c}} \sum_{I} S_{\mathrm{Born},I}, \tag{16}$$

where $S_{\mathrm{Born},\ I}$ has the same functional form as $S_{\mathrm{burial},\ I}$ with different parameter values.

We describe the amino acid side-chain (sequence)-dependent terms below. The hydrophobic interaction is written in the form

$$V_{\mathrm{HP}} = -\varepsilon_{\mathrm{HP}}^{(\alpha)} \sum_{i \in C_\alpha} S_{\mathrm{HP}}(\rho_i) - \sum_{i_s \in C_\beta} \varepsilon_{\mathrm{HP},A(i_s)}^{(\beta)} S_{\mathrm{HP}}(\rho_{i_s}), \tag{17}$$

where $i \in C_\alpha$ denotes α carbon atoms and $i_s \in C_\beta$ denotes the side-chain beads. $\epsilon_{\mathrm{HP}}^{(\alpha)}$ and $\epsilon^{(\beta)}{}_{\mathrm{HP},\ A(i_s)}$P are hydrophobic energy parameters for the C_α and the centroid of side-chain, respectively. The function $S_{\mathrm{HP}}(\rho_\mathrm{i})$ represents the buriedness of the atom group i and is defined in terms of the local density of peptide atoms ρ_i as

$$S_{\mathrm{HP}}(\rho) = \begin{cases} 1 & 1 < \rho \\ c_{\mathrm{linear}}\rho + \frac{1}{2}(1 - c_{\mathrm{linear}})\left[1 + \cos\frac{\pi(1-\rho)}{1-\rho_{\mathrm{min}}}\right] & \rho_{\mathrm{min}} < \rho \leq 1 \\ c_{\mathrm{linear}}\rho & \rho < \rho_{\mathrm{min}}, \end{cases} \tag{18}$$

where ρ_{min} and c_{linear} are used for combination of two function forms. The pair-wise contact is considered when ρ is small, while nonadditive hydrophobic interaction is considered for large ρ. The local density of peptide atoms ρ_i at the ith group is defined as

$$\rho_i = \frac{\sum_{j \in C\alpha C\beta} n_j u_{\mathrm{HP}}(r_{ij}, r_{\mathrm{min},A(i),A(j)}, r_{\mathrm{max},A(i),A(j)})}{n_{\mathrm{max},A(i)}}, \tag{19}$$

where $A(i)$ means the type of ith united atoms. The constant n_i corresponds to the number of nonhydrogen atoms in the all-atom representation of the group, which coarse-grained atom i represents, for example $n_i = 4$ for C_α and $n_i = 1$ for the side-chain bead of alanine. The constant $n_{\mathrm{max},\ A(i)}$ is the knowledge-based maximal coordination number for atom type i and correlates with its size. $r_{\mathrm{min},\ A(i),A(j)}$ and $r_{\mathrm{max},\ A(i),A(j)}$ are distance parameters. The function u_{HP} represents the degree of the contact between atom groups i and j and is defined as having a sigmoidal shape:

$$u_{\mathrm{HP}}(r, \sigma_{\mathrm{HP1}}, \sigma_{\mathrm{HP2}}) = \begin{cases} 1 & r < \sigma_{\mathrm{HP1}} \\ \dfrac{1}{2}\left(1 + \cos \pi \dfrac{r - \sigma_{\mathrm{HP1}}}{\sigma_{\mathrm{HP2}} - \sigma_{\mathrm{HP1}}}\right) & \sigma_{\mathrm{HP1}} < r < \sigma_{\mathrm{HP2}} \\ 0 & r > \sigma_{\mathrm{HP2}}. \end{cases} \qquad (20)$$

The secondary structure propensity is related to the side-chain entropy loss upon forming the secondary structure and depends on the amino acid sequence.

$$V_{\mathrm{Rama}} = -\sum_{I} \varepsilon_{\mathrm{Rama},\alpha,A(i_s)} \exp\left[-\frac{(\phi_I - \phi_{\alpha 0})^2}{2\sigma^2_{\alpha\phi}} - \frac{(\psi_I - \psi_{\alpha 0})^2}{2\sigma^2_{\alpha\psi}}\right] - \sum_{I} \varepsilon_{\mathrm{Rama},\beta,A(i_s)} \exp\left[-\frac{(\phi_I - \phi_{\beta 0})^2}{2\sigma^2_{\beta\phi}} - \frac{(\psi_I - \psi_{\beta 0})^2}{2\sigma^2_{\beta\psi}}\right], \qquad (21)$$

where $A(i_s)$ represents the type of the side-chain of the Ith amino acid. This potential makes two wells, the α-helix region and the β-strand region, in the Ramachandran plot. ϵ and σ mean the depth and width of wells, respectively. $(\phi_{\alpha 0},\psi_{\alpha 0})$ and $(\phi_{\beta 0},\psi_{\beta 0})$ are the centers of the wells.

The pair-wise contact energies represent amino acid-specific interactions between side-chains. This potential represents additional hydrophobic effects, aromatic interactions, and charge–charge interactions. We consider only attractive interactions in the present work. Our derivation of the potential function form is as follows. First, we analyzed the distance distribution between side-chains from the structural database. We then tried to fit two different functional forms, Gaussian and LJ forms, to the logarithm of the frequency in the distribution. We find that the Gauss function fits better the distribution of most side-chains except for that between alanine and alanine. We do not consider amino acid pairs of which distributions do not show any significant peaks. So, we introduced the pair-wise potential,

$$V_{\mathrm{pairwise}} = V_{\mathrm{pairwise.Ala,Ala}} + V_{\mathrm{pairwise,others}}, \qquad (22)$$

where

$$V_{\mathrm{pairwise,Ala,Ala}} = -\varepsilon_{\mathrm{pairwise,Ala,Ala}} \times \sum_{i<j}\left(\left(\frac{a_{\mathrm{AlaAla}}}{r_{ij}}\right)^{12} - 2\left(\frac{a_{\mathrm{AlaAla}}}{r_{ij}}\right)^{6}\right), \qquad (23)$$

and

$$V_{\mathrm{pairwise,others}} = -\sum_{i<j} \varepsilon_{\mathrm{pairwise},A(i),A(j)} \exp\lfloor -\beta_{A(i)A(j)}(r_{ij} - r^0_{A(i),A(j)})^2 \rfloor. \qquad (24)$$

Here, $r^0_{A(i),A(j)}$ is the interaction distance at which the two amino acids are located most frequently in the structural database and $\beta_{A(i)A(j)}$ is the width of interaction well. These parameters are determined so that the Gauss potential is best fit to the histogram from the structural database.

So far, we described the potential that is used for main-chain movements. Now, we describe the rotamer potential V_{rot} that is used for the side-chain moves. We did not write this potential in eq. 1 because it cannot be differentiated in the simulation. In the simulation, we calculate main-chain and side-chain moves in different ways. First, we calculate the main-chain move using molecular dynamics or Metropolis Monte Carlo (MC). After certain steps of main-chain move, we try to move side-chain conformations using the MC method. This procedure is iterated. The rotamer potential is added to the potential and is used for MC of side-chain move. The rotamer's intrinsic energy V_{rot} is derived from the probabilities of rotamer conformations, that is,

$$V_{\mathrm{rot}} = -\sum_{I} \varepsilon_{\mathrm{rot}} \ln p(r_k(A(I)), A(I)), \qquad (25)$$

where $r_k(A(I))$ represents the rotamer conformation of amino acid A and ϵ_{rot} is a scaling factor for changing probability to energy.

All scales and parameters used above may be downloaded from our Website (http://theory.chem.sci.kobe-u.ac.jp/).

OPTIMIZATION OF THE POTENTIAL ENERGY FUNCTION

Method: *Z* Score Approach

As explained in the Introduction, the energy landscape theory states that naturally evolved proteins should have a sufficiently large ratio $T_{\mathrm{F}}/T_{\mathrm{G}}$, which then leads to the large ratio $\delta E_s/\Delta E$ within a simple random energy model. This can be utilized as a guiding principle for deriving an energy function; the energy function should give a large ratio $\delta E_s/\Delta E$ for natural proteins. In other words, we can optimize an energy function so that we maximize this ratio $\delta E_s/\Delta E$ or its modified version.

For this purpose, we rewrite the potential energy function in the form

$$V = V_0 + \sum_i \varepsilon_i u_i,$$

where V_0 represents the prefixed terms that are not subject to optimization, which include van der Waals energy between adjacent amino acids and hydrophobic energy of the main-chain. The second term is the major sequence-dependent interaction terms that are optimized. Here, note that i stands for the types of interaction parameters: They include 19 parameters for hydrophobicity of each amino acid side-chain (except GLY), $\epsilon^{(\beta)}{}_{\mathrm{HP},A}$, 40 parameters for propensity of secondary structure (α-helix or β-strand) of each amino acid, $\epsilon_{\mathrm{Rama},SS,A}$, and 33 parameters for specific pair-wise contact energies that depend on pair of amino acid types, $\epsilon_{\mathrm{pairwise},A,B}$. For the latter, it is possible to incorporate other interactions between side-chains into the potential, in principle. In the survey of structural data set, however, the pair distance probability distributions of many pairs do not exhibit clear peaks, implying no specific attractions. Some others are found too rarely in our training set; if such poorly sampled situations are subject to optimization, they may cause overfitting in the optimization procedure. We therefore restrict the pair-wise interactions to those that appear more than nine times in the training set.

For a given energetic parameter set ϵ (a vector containing 92 rows), we define the Z score, which is the negative of $\delta E_s/\Delta E$,

$$Z(\varepsilon) = \frac{V(r^N, \varepsilon) - \langle V(r, \varepsilon)\rangle_D}{\Delta V(r, \varepsilon)_D},$$

where

$$\Delta V_D = \sqrt{\langle V(r)^2\rangle_D - \langle V(r)\rangle_D^2}.$$

Here, r^N stands for the native structure and the subscript D denotes ensemble of decoy structures r that are generated by the method described in the next subsection and <> means ensemble average over the D ensemble. Because the structures in the D ensemble have a broad distribution in energy, a simple average over the D ensemble gives an energy too high relative to that at the native structure. Therefore, the canonical average is used:

$$\langle \Omega\rangle_D = \frac{\Sigma_{\text{ensemble}}\ \Omega \exp(-\beta V)}{\Sigma_{\text{ensemble}} \exp(-\beta V)},$$

where Ω is a physical quantity and β^{-1} is a selection temperature not necessarily corresponding to the room temperature, and is chosen so that the effective number of configurations that contribute to the average is sufficiently large.

The Z score above is defined for a single protein, while we seek the potential energy function that makes Z scores of "all" proteins simultaneously low enough, that is, negative and large in the absolute value. Therefore, we need a single objective function that reflects the Z scores of the many proteins in the training set. If we minimize the simple algebraic average of Z scores over proteins, we obtain an energy function that gives small Zs for many proteins but large Zs for a few proteins. This is not desired because the Z scores of "all" proteins need to be small. This problem arises partly because proteins in the training set have different quality decoys and partly because the current energy function allows some proteins to be more easily recognized than others. Another extreme approach may be to minimize the maximum of Z scores of proteins in the training set as in our previous work.[7] With this extreme, however, the optimized parameters are largely determined by few proteins, which may be exceptional ones. In particular worrisome is that these could be structures with some errors. This tends to lead to unphysical energy parameters. Some intermediate approaches have been proposed by Koretke et al.[4] as well as by Mirny and Shakhnovich,[17] where average is taken so that the larger (i.e., worse) Z score has more weight in average. In this work, we chose a weighted average somewhat similar to Koretke et al.[4] The objective function Z_{ave} to be minimized is the Boltzmann-like weighted average of $Z^{(m)}$ in the training set:

$$Z_{\text{ave}} = \frac{\Sigma_m^M Z^{(m)} \exp(\beta_p Z^{(m)})}{\Sigma_m^M \exp(\beta_p Z^{(m)})},$$

where β_p is a constant value for weighting and m indicates the protein in the training set.

We apply two constraints during optimization. First, the sum of the magnitude of the interaction terms $\Sigma|\epsilon_i|$ is restricted to be a constant; otherwise, all the energetic parameters will tend to become large relative to the prefixed part of the energy function V_0. Second, the potential parameters for secondary structure propensity are restricted to be positive.

For the minimization of Z_{ave}, the Monte Carlo procedure is performed. Starting from random interactions, at each step, one of energetic parameters is chosen at random and changed. A change in Z score (ΔZ) is computed. The change is accepted with the probability, min(1,exp($-\beta_z \Delta Z$)). Here, $\beta^{-1}{}_z$ is a selection "temperature" for Z scores. After the optimization of Z score, we check the potential parameters by running MD simulations. We find the radius of gyration of structures obtained in MD is, in general, smaller than those of native structures. We find that pair-wise contact terms obtained from the Z score approach tend to be too large and thus we decided to reduce the pair-wise parameters by a factor of 0.20 for giving a radius of gyration similar to nature.

Choice of Training Set Proteins

A training set that includes 40 polypeptide chains in the Brookhaven Protein Data Bank (PDB) is used for the optimization of the energetic parameters. We first chose a protein set in which any two proteins have their sequence homology less than 25%. All the proteins sequences are longer than 20 residues and shorter than 200. Bastolla et al. presented evidence that failures in optimization can often be attributed to the neglect of interactions between different chains in oligomeric proteins or with cofactors.[18] They also reported NMR structures are more difficult to guarantee optimal stability. Thus, we restricted our training set to protein structures determined by X-ray crystallography. We excluded metal-binding proteins, membrane proteins, oligomeric proteins, and proteins that have obvious chain breaks in middle and extremely extended structures. The training set thus obtained are listed below by their PDB codes (chain length): 153l(185), 1agy(197), 1aho(64), 1amm(174), 1amx(150), 1bd8(156), 1bea(116), 1bfg(126), 1bgf(124), 1bj7(150), 1bm8(99), 1bv1(159), 1fna(91), 1fus(105), 1gpr(158), 1ctf(68), 1hoe(74), 1hyp(75), 1ifc(131), 1koe(172), 1lcl(141), 1lki(172), 1msi(66), 1npk(150), 1orc(64), 1pdv(195), 1pdo(129), 1pne(139), 1sfp(111), 1vie(60), 1whi(122), 1who(94), 1xnb(185), 2cpl(164), 2end(137), 2i1b(153), 2igd(61), 2pth(193), 2rn2(155), and 3tss(190).

Generation of Decoys

Decoys (nonnative structures) are generated with a simple gapless threading method. A set of proteins used for threading template is defined in the same way as the training set except that the chain length is restricted to be

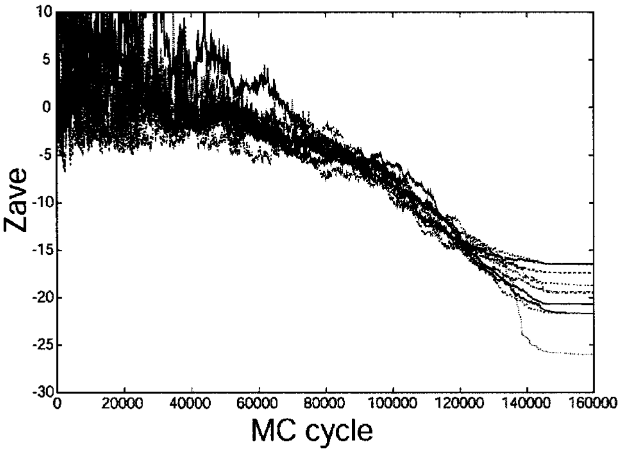

Fig. 1. *Z* score change along the MC cycle in the *Z* score optimization procedure. The results of 10 independent simulations are overwritten. Plotted here are the weighted average Z values over the training set proteins. “Temperature” for the MC run is decreased gradually during the simulation and is set to zero in about the last 16,000 cycles.

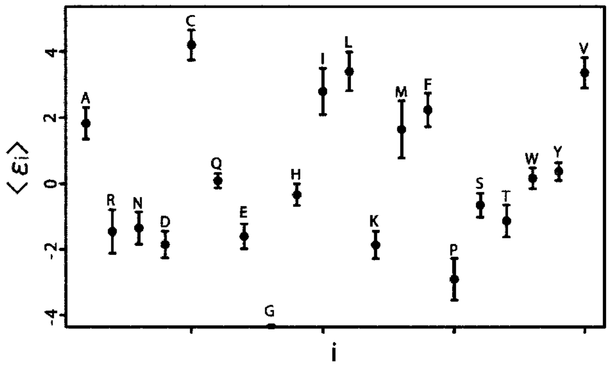

Fig. 2. Optimized energy parameters for the hydrophobicity.

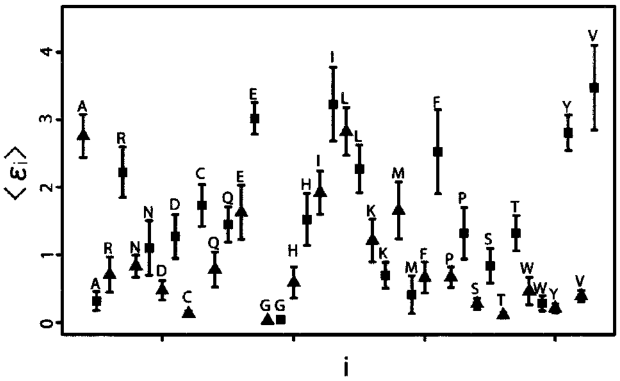

Fig. 3. Optimized energy parameters for the secondary structure propensity. The triangles and circles represent α-helical propensity and β-strand propensity, respectively.

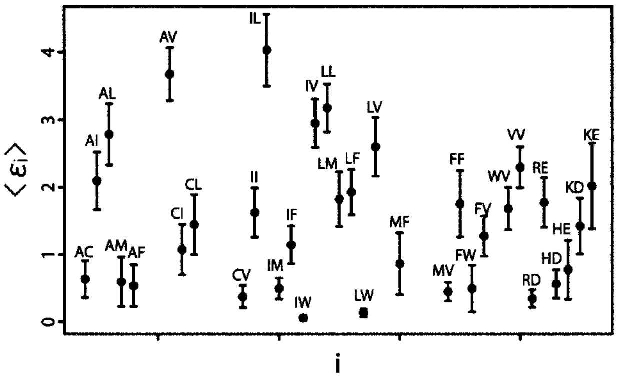

Fig. 4. Optimized energy parameters for the pair-wise contact interaction.

larger than 100 and smaller than 1000. The total number of proteins in the threading template set is 219. Decoys are constructed by threading each protein sequence in the training set onto protein structures for which sequences are longer than that of the trained protein. On average, 600 decoys are generated for each protein in the training set.

Optimization of Potential Energy Function

To obtain the potential that minimizes Z scores for proteins in the training set, we used Monte Carlo search with simulated annealing in the space of energetic parameters. The weight factor for averaging physical quantity for a protein, β, and the weight factor for averaging Z scores over proteins, β_p, are empirically set to (1000 K_B kcal/mol)$^{-1}$ and 1.0, respectively. We performed 16,000 MC cycles. The parameter temperature $\beta^{-1}{}_z$ in the Monte Carlo search is exponentially decreased from 10. After the temperature reaches 0.01, it is reset to 0. Starting from random potentials, the Monte Carlo search converges fast (Fig. 1). Among 10 independent simulations performed, the average of correlation between the energy parameters from 2 simulations is 0.56. This relatively high correlation indicates that optimization has succeeded. Hereafter, we use the averages of derived potential parameters that are shown in Figures 2, 3, and 4.

PHYSICAL INTERACTIONS INFERRED FROM PROTEIN STRUCTURAL DATA ALONE

Comparison of Optimized Energy Parameters with Experimental Data

We now compare the physical interaction parameters inferred from the PDB with interaction parameters derived from thermodynamic experiments. Figure 5 compares hydrophobic interaction parameters of each amino acid with corresponding experimental data.[19] The latter are the transfer energy of *N*-acetyl amino acid amide from water solvent to organic solvent. The correlation coefficient $r = 0.73$ shows that the energy function form properly reflects this property of amino acids and gives confidence in the optimization procedure.

There are two outliers: proline and tryptophan. The proline is intrinsically hydrophobic but the energy parameter shows the most hydrophilic propensity of all amino acids. Most prolines are located in loops in protein structures, especially in small proteins. The prolines exposed to the solvent are inferred as hydrophilic ones in the optimization procedure. If some of the training proteins contain prolines in buried positions, the propensity of proline may

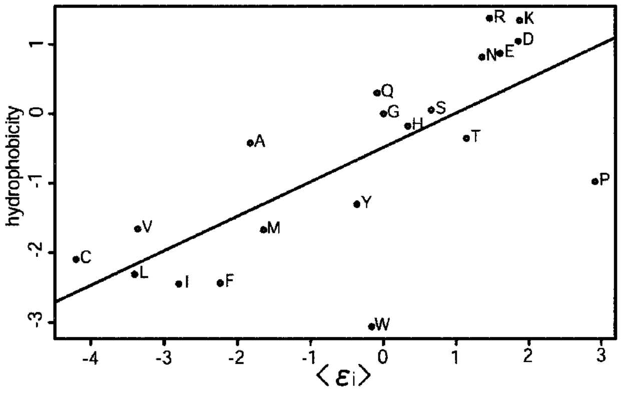

Fig. 5. Derived hydrophobic energy parameters versus experimentally measured hydrophobicity scales.[19]

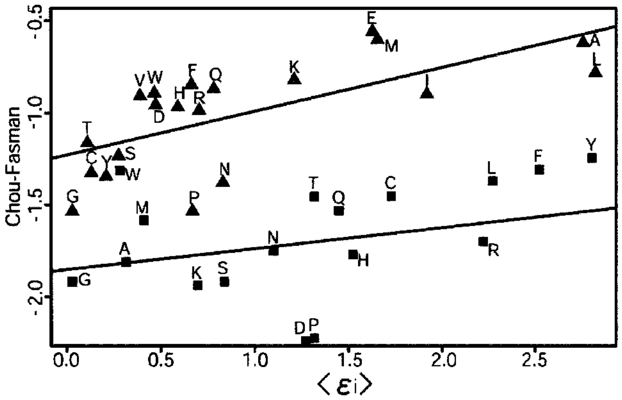

Fig. 6. Derived secondary structure propensity versus statistical data. The triangles and circles represent α-helical propensity and β-strand propensity, respectively. Outliers of parameters about the β-sheet are omitted.

be correctly recognized as hydrophobic. Tryptophan is rare in proteins. The hydrophobic potential parameter of tryptophan shows it to be neither hydrophobic nor hydrophilic. We note that pair-wise contact parameters of tryptophan indicate strong interaction with other hydrophobic amino acids. In the current energy function, hydrophobic interaction and some pair-wise contact terms are inherently correlated. Thus, the hydrophobic property of tryptophan is largely represented by the pair-wise terms.

Comparison of Optimized Energy Parameters with Statistical Data

We also investigated the correlation between the calculated potential for the secondary structure propensity and the corresponding statistical data often used (Fig. 6). For the specific comparison, we use Chou Fasman statistics,[20] $\log(f_{A,SS})$, which is the logarithm of frequency, $f_{A,SS}$, that an amino acid A forms a secondary structure SS (α or β). The correlation coefficients are 0.67 and 0.30 for the α-helix and β-strand Ramachandran bias parameters, respectively. We note that formations of secondary structure are dependent on not only potential parameters $\epsilon_{\mathrm{Rama,SS,A}}$ but also van der Waals interactions with their neighboring amino acids. Therefore, a simple one-to-one comparison of secondary structure propensity may not be appropriate. Relatively high correlation for α-helix and low correlation for β-sheet may correspond to that helical propensity is largely determined by intrinsic amino acids, while β-sheet propensity is more dependent on the tertiary context and thus is more difficult to be inferred from limited size of the training set.

BENCHMARK TESTS OF THE OPTIMIZED ENERGY FUNCTION

We performed three different benchmark tests of the energy function optimized above. The first uses the predefined protein decoy set and checks if the energy function correctly recognizes the native structure from wrong ones. Then, we move to a much more challenging structure prediction test where we combine our energy function with the so-called fragment assembly sampling method that has been successfully used by Baker's group and others.[21,22] This simple sampling method limits the sampling space severely, which makes constructing prediction models quick. Finally, we also use the replica exchange MD simulations to sample protein conformations more widely. We compare prediction performance of the fragment assembly method with those of the replica exchange method.

Discrimination on Decoys'R'Us

As a first test of the optimized energy function, we use a predefined set of protein nonnative structures, decoys, and see if the current energy function gives a lower energy at the native structure than for the decoys. For the decoys set, Decoys'R'Us of Levitt and his collaborators, which is downloaded from their Website, was used.[11] We tested our potential on two types of decoy sets: 4-state-reduced (Park and Levitt) and lmds (Kesar and Levitt).[23,24] The former were prepared from 7 proteins and include, on average, 665 decoys per protein. The latter were produced by Kesar and Levitt by using minimization with a complex potential that contains a significant pair-wise component as well as cooperative hydrogen bonds.[24] Prior to energy comparison, all decoys and native structures are locally optimized by the steepest descent method with the current energy function.

Figure 7 plots energies and the root mean square deviation (RMSD) from the native structure for the decoy and native structures in 4-state-reduced sets. Because the steepest decent minimization results in a small RMSD change from its initial value, the initial RMSD value is used in the figure. We see that the native structures give the lowest energies for five of seven proteins and that the structures with RMSD smaller than 3.0Å give the lowest energies for all seven proteins. The energy gaps between the native structures and lowest-energy decoys are variable from protein to protein, from −5 kcal/mol of 3icb to 50 kcal/mol of 1ctf. More importantly, we see that as the RMSD increases energy becomes larger and larger. This is a desirable property. We also note that 4rxn and 3icb, for which the native structures do not have the lowest energy, possess metal ligands (iron for 4rxn and calcium for 3icb). For these proteins, stability originates from interactions

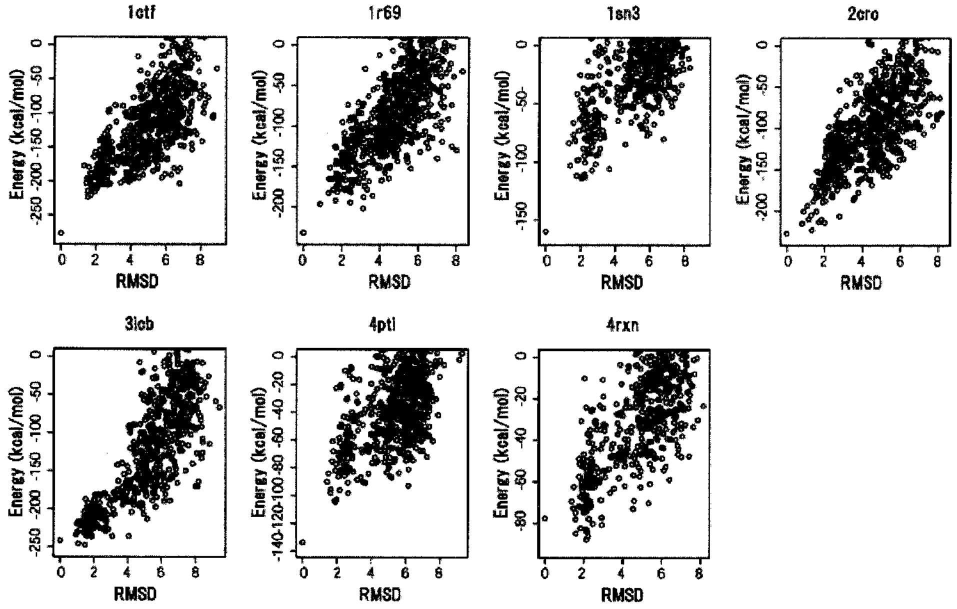

Fig. 7. Optimized energies for the native and all decoy structures versus RMSD from the native structure.

TABLE I. Performance of Scoring Functions on the Decoys'R'Us Collection of Decoys[†]

	Total			Hydrophobicity		Rama		Pair-wise		MJ	
Protein	Rank	*Z*	Low	Rank	*Z*	Rank	*Z*	Rank	*Z*	Rank	*Z*
Park and Levitt											
1ctf	1	−2.5	0.0	8	−2.2	1	−3.1	3	−3.1	1	−3.7
1r69	1	−2.5	0.0	13	−2.1	1	−3.8	1	−3.8	1	−4.1
1sn3	1	−3.2	0.0	5	−2.1	1	−4.0	122	−0.9	2	−3.2
2cro	1	−2.3	0.0	9	−2.3	1	−3.4	1	−3.3	1	−4.3
3icb	3	−1.6	1.5	10	−1.8	1	−3.2	144	−0.7	—	—
4pti	1	−2.7	0.0	33	−1.7	1	−3.2	144	−0.8	3	−3.2
4rxn	12	−1.7	2.1	21	−1.8	9	−2.2	53	−1.5	1	−3.1
lmds											
1ctf	1	−4.7	0.0	7	−2.0	38	−1.4	14	−2.1	1	−3.9
1dtk	2	−2.3	9.5	2	−2.5	84	−0.3	176	0.9	13	−1.7
1igd	1	−6.2	0.0	3	−2.8	2	−2.7	4	−3.4	—	—
2cro	1	−4.0	0.0	1	−5.4	488	1.9	1	−4.5	1	−5.1
2ovo	1	−4.1	0.0	4	−2.8	132	−0.4	18	−1.8	1	−5.1
4pti	17	−1.8	9.2	13	−1.9	176	0.1	63	−0.9	—	—

[†] *Z* score, defined in Results; total, hydrophobicity, Rama, pair-wise, results for our potential; MJ, potential derived by Miyazawa and Jernigan; low, RMSD of the lowest energy.

with the ligand, while the current optimized energy function does not take into account these ligands. The fact that the native structures are not the lowest state only for these two proteins may be reasonable.

Table I shows native energy rank and *z* score of total energy and energy components. To compare our results with other potential, the results of MJ, a contact potential derived by Miyazawa and Jernigan, are also shown.[25] For comparison, we define *z* score:

$$z \equiv \frac{V_N - V_D}{\sigma},$$

in which V_N and V_D are native energy and algebraic average of decoys, respectively. σ is the standard deviation of energies including native structure. It is noticed that *z* score in this discrimination test is different from the one used in the optimization procedure. Although the native *z* score of our potential is slightly lower than that of MJ, native ranks of our potential were lower than MJ except for 4rxn. We note that MJ includes 210 energetic parameters derived from the PDB, while the current one has 92 parameters to be optimized. Hydrophobicity is the dominant contribution to discrimination both in 4-state-reduced and lmds sets. But, the hydrophobic energies stabilize both native structure and some of decoys and thus hydrophobicity alone cannot discriminate the native structure from decoys. The secondary structure propensity energies contribute significantly to the recognition of the native structure on the 4-state-reduced set. The exception is 4rxn, of which the native structure has few secondary structures. Most of the torsion angles of decoy in 4-state-reduced sets are outside of the effective region of our potential. Pair-wise energies contribute to discrimination between native structure and decoys to some extent. Some decoys have lower energy than the native structure but the energy differences never exceed 2 kcal/mol. It may be important to emphasize that no single energy term is sufficient to recognize the native structure from decoys and, in this sense, the native structure is apparently designed with the use of many physical interactions.

Protein Structure Predictions with the Fragment Assembly Method

Next, with the optimized energy function, we performed a benchmark test of structure prediction using the fragment assembly method developed by Baker and others.[12] The fragment assembly protocol consists of two parts: preparation of fragment structural candidates and generation of protein tertiary structures by assembling fragments.

First, we prepare the fragment candidates for a target. Following Baker and coworker's procedure, we use nine- and three-residue fragments: They showed that there is stronger correlation between local sequence and local structure for the nine-length fragments than other length ones. Three length fragments are also prepared to provide the smooth moves in the following Monte Carlo procedure. For each segment of 9 residue sequences, about 20 fragment sequences are chosen by a simple sequence–sequence comparison to nonredundant structural database, in which the BLOSUM62 substitution matrix is utilized for scoring.[26] For the three-residue segments, we pick up both fragments that have exactly the same sequence as the target segment and central three-residue fragments of the nine-length fragments library prepared above. All the chosen fragment candidates are structurally optimized, prior to construction of the whole protein model, with respect to the bond lengths and angles. All fragments that are selected from homologs of the target proteins are removed.

Next, complete tertiary structures were generated. All simulations start from different random conformations. At a randomly chosen nine or three segments, the torsion angles are replaced with those of a randomly chosen fragment that was prepared for this segment in advance. Moves are then evaluated according to the standard Metropolis criterion. This procedure is repeated with decreasing temperature: the simulated annealing. As is in general known, most of the torsion angle replacements are likely to be rejected because of atomic overlaps in a trial structure. Further, because the fragment assembly method gives discrete conformational space it is difficult to attain accurate nonlocal hydrogen bonds in the β-sheets. So, we introduce some modification of the potential. Namely, we set the upper limits for the repulsive parts of hydrogen bond and van der Waals potentials, enabling protein atoms to overlap to some extent. This modification increases the acceptance ratio of Monte Carlo trial moves and enlarges the sampled conformation space.

Because we find that conformations sampled with the physical interactions alone tend to be extended, we also introduce an additional potential that induces compactness in the sampled conformations:

$$V_{\mathrm{Rg}} = \frac{1}{2} k_{\mathrm{Rg}}(R_g - R_g^0)^2,$$

where

$$R_g^0 = 2.96N_{aa}^{1/3} - 0.84$$

Here, k_{Rg} is strength of bias. R_g and $R^0{}_g$ are the radius of gyration of simulated chain and its statistical value, respectively. The latter $R^0{}_g$ is estimated from the classic Miyazawa–Jernigan protein set.[15] This propensity to compactness is effective but not indispensable. This additional potential just eases the search for compact structures.

The test proteins were chosen from main classes in the Cath database: mainly α [albumin-binding domain (1prb), 434 repressor (1r69)], mainly β [major cold-shock protein (1nmg)], and α/β protein [protein G (2gb1), protein L (2ptl), ubiquitin (1ubi)]. For each target protein, we repeated MC annealing simulations 400 times. Because structures constructed from fragments lack structural plasticity, nonlocal hydrogen bond distances and vdW packing are usually poorly optimized. Therefore, after the fragment assembly simulation we performed two structural optimization steps: First, we minimized the structure, eliminating the steric hindrance caused by the modified hydrogen bond potential. Next, we performed a 1-ns MD simulation using the current energy function and quenching to attain better hydrogen bond formation. The hydrogen bond parameters in these optimization procedure, such as ϵ_{HB}, are those for MD, which are slightly different from those in MC procedures. The details of MD simulation are described in the next subsection. We then performed clustering of 400 obtained structures using the pair-wise C_α RMSD as a measure.[27,28] The cluster cutoff was tuned so that the size of the largest cluster is about 8–10% of the total number of structures. We finally chose the centers of the five largest clusters as the prediction models (Table II). As in Refs. 27 and 28, we assume that the minimum of the native topology is broader than any other minimum in the energy landscape. The cluster size is regarded as the broadness of basin in the energy landscape. Thus, we examine whether one of the five biggest clusters contains native topology as the measure of the success.

Figure 8 plots energies of 400 obtained structures and their RMSD from the native structure. We find weak correlation between them: Structures with lower RMSD tend to have lower energies. This weak correlation may be explained in the following way: Available torsion angles are discrete in the fragment assembly method, while the hydrogen bond potential shape is sharp, having a narrow well. Some structures, irrespective of their global fold, happen to form optimal hydrogen bonds while others include poor hydrogen bonds. The former may have lower

TABLE II. Summary of Predictions with the Fragment Assembly Method†

		Best	Low *E*	Cluster			
Name	Nres	RMSD (Å)	RMSD (A)	RMSD (Å)	LCS 5 Å	GDT 6 Å	CE *Z*
1prb	47	1.8	1.8	3.0	47	47	3.5
1r69	63	5.0	10.8	5.8	59	50	3.7
2gbl	56	4.5	4.9	5.4	53	42	3.5
2ptl	60	5.3	10.4	5.8	49	48	3.5
1nmg	67	5.1	9.4	5.1	63	51	3.7
1ubi	76	6.4	10.4	10.0	27	32	2.0

†Nres, length of protein chain; best, RMSD of the best structure in the 400 MC simulations; low *E*, RMSDs of the lowest-energy structure; RMSD, LCS, GDT, and CE *Z*, results of the best cluster center in the top five largest clusters; LCS and GDT, results of LGA server analysis; CE *Z*, results of CE server analysis.

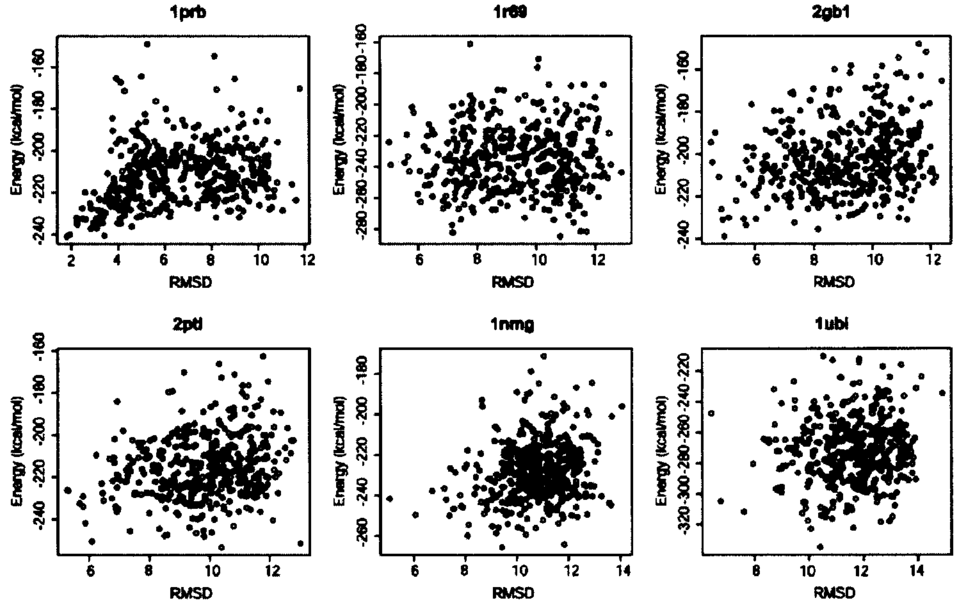

Fig. 8. Total energies and RMSDs from the native for structures sampled by the fragment assembly method. Each circle corresponds to a structure quenched from the final structure of the fragment assembly MC simulation. Proteins simulated are the albumin-binding domain of protein A (1prb), the 434 repressor (1r69), the protein G (2gb1), the protein L (2ptl), the cold-shock protein B (1nmg), and ubiquitin (1ubi).

energy than the latter simply because of the better hydrogen bonding attained. This noise of hydrogen bond potential is significantly large. Thus, the cluster size may be more informative than the energy itself for selecting the best prediction model. In Figure 8, the lowest RMSD is far from 0 (native), in part due to discrete torsion angles.

The summary of the resulting predictions for six small proteins is given in Table II. For 5 of 6 proteins, the best RMSD out of 400 structures is less than 6.0Å. The cluster cutoff that leads to the largest cluster being about 10% of all structures correlates with the degree of prediction success that can be monitored by the best RMSD of five models. Cartoon diagrams illustrating some of the models are presented in Figure 9. The model structures drawn as red lines are the cluster centers that have the minimal RMSD of the five largest clusters. For the simplest topological protein that has the three-helix-bundle topology, 1prb, the cluster center of the largest cluster is close to the native structure (RMSD of 3.0 Å). For 434 repressor, 1r69, although the global topologies of the drawn cartoons differ from the native one the central three helices of the predicted structure are similarly arranged to those in the native structure. For protein G and protein L, which have 1 α-helix and 4 β-strands, 2gb1 and 2ptl, respectively, the third and fifth largest cluster centers are closest to the native structure, respectively. For the major cold-shock protein, 1nmg, the additional helix is formed but global topology is correct. Unless predicted structures have the same overall fold as the native, evaluating these structures with the global RMSD is not necessarily appropriate for accessing performance in partial regions of the proteins. Thus, we also evaluate structures with two other methods, which employ more local criteria of assessment. One is the evaluation of the similarity between target and predicted structures using the LGA server, which has been used in CASPs automatic accessment.[29] We submit the best RMSD structures of the five largest clusters to the LGA server. The predictions are analyzed with two methods. LCS is the longest continuous segments under specified C_α RMSD cutoff (5 Å). GDT is the global distance test,

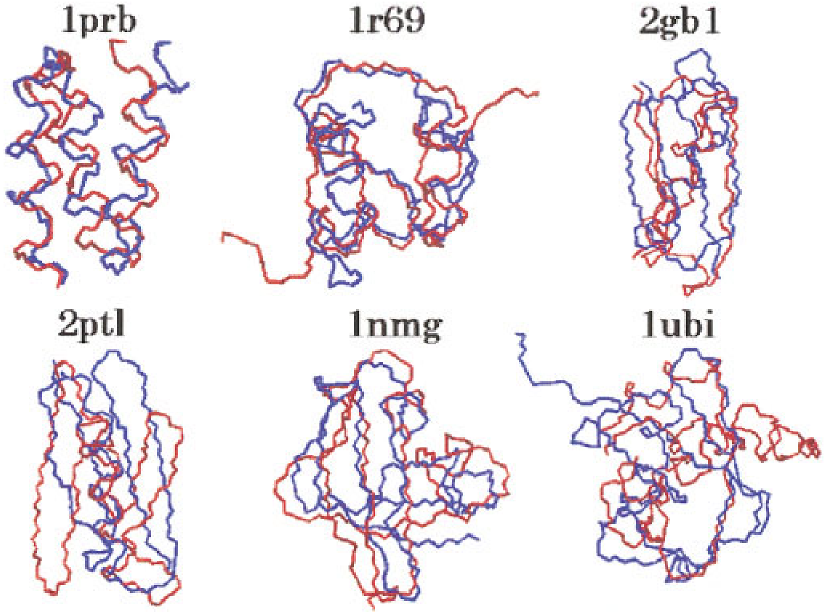

Fig. 9. Superimposition of experimentally observed structures and those predicted by the fragment assembly sampling for six small proteins. Blue and red depict the native structure and the best cluster center of the top five clusters as identified in Table II, respectively. The proteins studied are the same as in Fig. 8.

which identifies the sets of residues that do not deviate from the target structure by more than a specified C_α distance cutoff (6 Å). The other is the combinatorial extension (CE), which builds an alignment between two protein structures.[30] The longest alignment path is evaluated for statistical significance represented as a CE Z score. A CE Z score greater than 3.5 is considered to be a probably correct assignment of two structures to the same protein family. Thus, we see that fragment assembly with this energy function gives structures in the correct class. Ubiquitin is the exception to this rule.

Protein Structure Predictions with the Replica Exchange Molecular Dynamics

First, we present MD simulation protocols. To be precise, MD in this article means the overdamped Langevin dynamics that we used in our previous studies.[7] A timestep of 30 fs is used and the LINCS algorithm is employed to constrain bond lengths and angles.[31] For investigating appropriate temperature range to be incorporated, we perform three 420-ns folding runs from different random conformations for each protein and one 210-ns run from the native structure.

We then employ the replica exchange MD to accelerate conformational search.[13] The replica exchange method is more efficient than standard simulated annealing. A total system in the replica exchange method consists of noninteracting copies of proteins, each of which is in the canonical ensemble at a certain temperature. The temperatures are chosen to cover the folding transition region so that the protein changes its structure from extended at the highest temperature to the compact ones at the lowest temperature. We prepare 8 copies for a small protein and 16 copies for midsized proteins. Pairs of replicas that have the neighboring temperatures are tried to exchange at some intervals. Thus, a random walk in "temperature space" is realized for each copy. In Table III we summarize the temperature parameters; the temperature is distributed algebraically. Structures are saved every 1 ns from the replica with the lowest temperature for analysis.

TABLE III. Summary of Temperatures in the Replica Exchange Simulations

Protein	No. of replicas	Temperature range (K)
1prb	8	250–350
1r69, 2gbl, 2ptl, 1nmg	16	250–350
1ubi	16	250–500

We performed MD simulations based on the replica exchange method for the same six proteins as the previous subsection. We did not add any modifications to the potential function for overlap and for bias to the gyration radius, while that was done in fragment assembly. As a measure of structural similarity to the native, we introduce an order parameter Q* defined as

$$Q = \frac{1}{N_{IJ}} \sum_{I<J} \exp\left[-\frac{(r_{IJ} - r_{IJ}^{N})^2}{3^2} \right],$$

where r_{IJ} and $r^{N}{}_{IJ}$ are the C_α distance in a given structure and in the native structure, respectively. N_{IJ} is the number of summed pairs. Judged by eye, the structures that have $Q > 0.6$ can be viewed as the native

*Note that this Q is different from the convention we use in the associative memory Hamiltonian articles.[5] This Q allows only pairs within 1Å.

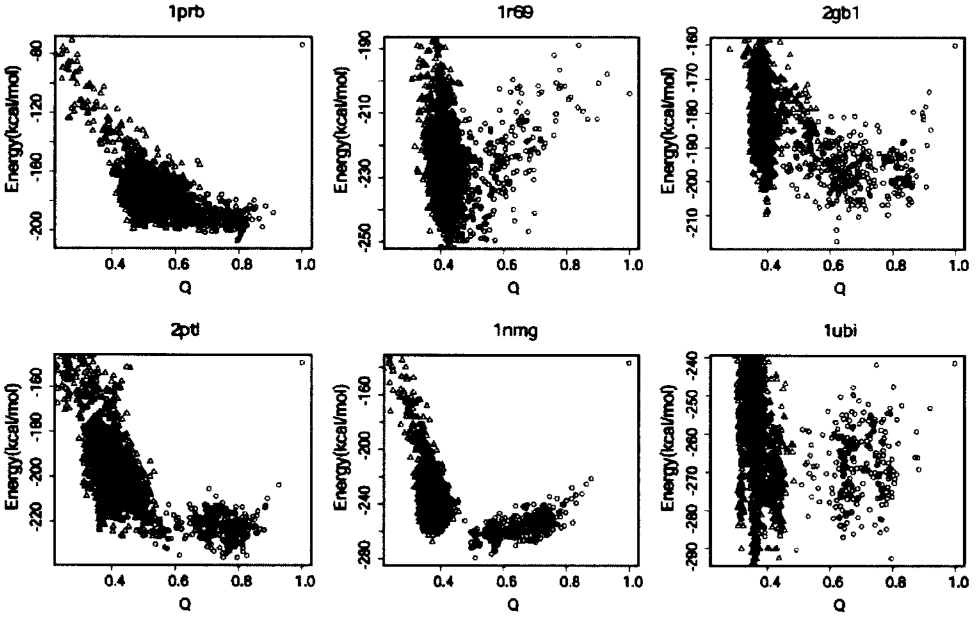

Fig. 10. Total energies and the structural similarity to the native Q for the structures sampled by the MD simulations. All the data are obtained from the replica at the lowest temperature. Circles and triangles are obtained from simulations from the native and those from the random coil, respectively. The proteins studies are the same as those in Fig. 8.

topology. On the other hand, the structures that have $Q \approx 0.4$ are compact but dissimilar to the native topology. Figure 10 depicts the relationship between energy and an order parameter Q of structures sampled at the lowest temperature. The triangle and circle are obtained from folding simulation and simulation from the native structure, respectively. Because conformational space with $Q > 0.8$ is small, to obtain these we saved structures every 0.1 ns from the initial stage of unfolding simulations.

We find that, in most of the proteins, the structures that have $Q \approx 0.7$ have the lowest energy. The averages of energy $E(Q)$ are almost unchanged in the range $0.5 < Q < 0.9$: We see $E(Q)$ has a caldera shape.[5] Some of the decoys in the caldera come from sampling a much larger conformational space that contains many structures that are on other grounds unlikely for real proteins. This problem does not arise in the conformational space of the fragment assembly method because only protein-like local structures are utilized. We note that the present Z scores' optimization procedure is not affected by this problem because decoys generated from the gapless threading do not have nonprotein-like torsion angles. We have this difficulty only after MD simulation. We also note that the native structure corresponding to $Q = 1$ has relatively higher energy because of the absence of optimal packing in our potential.

We analyzed these predicted structures with RMSD, LCS, GDT, and CE Z. We chose the predicted structure as the best (smallest RMSD) structures of the five lowest-energy structures. The results are listed in Table IV. The results of 1prb and 2ptl with MD simulation are comparable to those with fragment assembly simulation. For other test proteins, MD simulation gives poorer structures than the fragment assembly simulation. This indicates another difficulty related to the problem of sampling conformational space in the MD search. Although the conformational spaces of the small protein 1prb and the simple topology proteins 2gb1, 2ptl, and 1r69 are sufficiently explored, that of 1nmg, an all-β-sheet protein, is not. There is a gap in Q scores between the structures from folding and unfolding simulations, indicating a lack of equilibration in folding for the case of 1nmg. In the replica exchange methods, structures at a temperature lower than a certain value are almost always compact, while structures at higher temperature tend to be extended. This makes replicas separated into two groups that are not easily exchangeable. This is the limitation of conventional temperature replica exchange method. More powerful sampling methods are necessary such as the replica exchange multicanonical method, the multicanonical replica exchange method,[32] or the Hamiltonian replica exchange method.[32,33]

TABLE IV. Summary of Predictions with the Replica Exchange Method[†]

Name	Nres	RMSD (Å)	LCS 5 Å	GDT 6 Å	CE Z
1prb	47	4.4	47	42	3.7
1r69	63	9.8	28	25	2.6
2gbl	56	7.1	31	37	2.8
2ptl	60	6.6	41	39	3.5
1nmg	67	10.5	23	36	2.0
1ubi	76	9.9	40	34	2.3

[†] Nres, length of protein chain; RMSD, LCS, GDT, and CE Z, results for the best RMSD structure of the five lowest-energy structures.

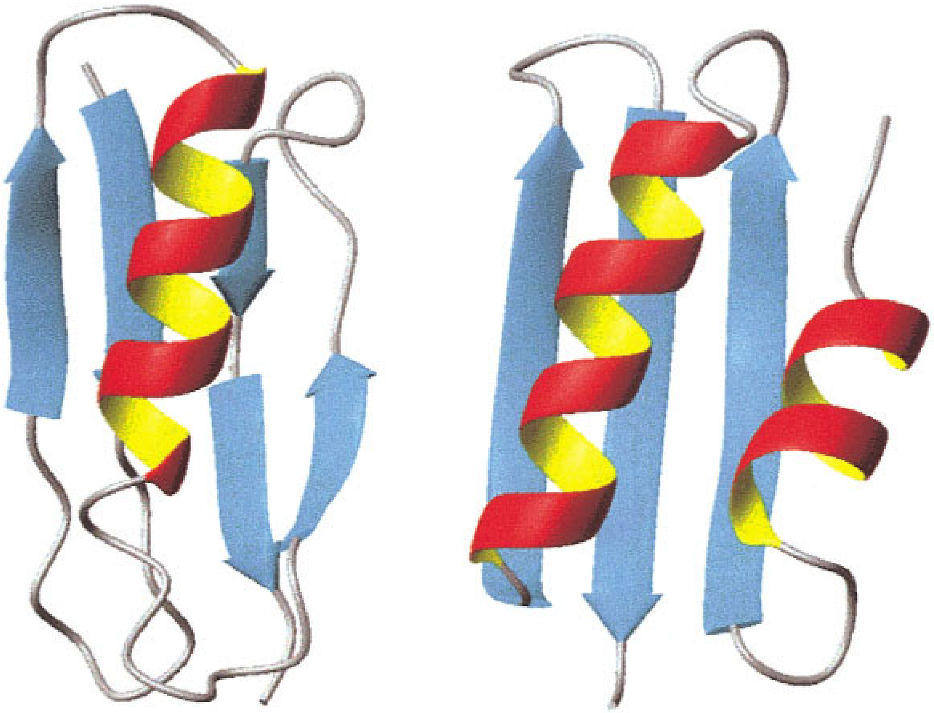

Fig. 11. Comparison of the experimental structure (left) and a representative structure obtained by the replica exchange MD that has one of the lowest energies (right) for the protein L (2ptl).

DISCUSSIONS AND CONCLUSIONS

In this article, we optimized the physicochemical energy function for proteins using as input the available structure database. The derived energy parameters correlate with experimentally measured quantities reasonably. It is worth contemplating how remarkable it is that structural data alone can give numbers so well in harmony with thermodynamic measurements! We then tested the energy function by three different means: (1) a native structure recognition test against predefined decoy sets of Levitt, where we found that, for all seven proteins available in the Park and Levitt decoy sets, the native or near-native structures have the lowest energy among all decoy structure energies; (2) a structure prediction test of six small proteins with fragment assembly sampling method, where a native-like topology is obtained in one of the top five cluster centers of the predicted models for five proteins; and (3) a structure prediction test with the replica exchange MD simulations that showed poorer performance than the fragment assembly methods.

Because we performed structure prediction tests both with the fragment assembly sampling and with the replica exchange MD using basically the same energy functions, it is interesting to compare the performance of the two search algorithms and argue why the differences appear. As mentioned above, we saw that the fragment assembly sampling showed better overall performance than the MD approaches. A simple and plausible reason for this difference is the sampling efficiency: At least, for the small proteins used in this article, the fragment assembly sampling is quick and efficient, while the complete sampling by the MD simulation is not an easy task at all. This is evident from Figure 10 because structures obtained from unfolding simulations (circles) have lower energies than those sampled starting from the random initial structures. A more nontrivial and possible reason may be the importance of accurate local structures. Taking a model structure of the protein L (2ptl; Fig. 11) obtained from an MD simulation, for each of three residues, we searched for the fragment structure closest to the model within the corresponding library used in the fragment assembly. In the model obtained from MD, there are many fragment structures that do not have structural neighbors in the library. To evaluate our model structure, we further created an all-atom structure using SCWRL.[34] In the all-atom model, we find a few crucial atom clashes per protein between a side-chain atom and the main-chain oxygen atom. Most of the conformations with low energy and large RMSD from the native structure contain such steric hindrance. In our potential, the secondary structure propensity is explicitly considered as potential wells in two oval regions in the Ramachandran plot. The present result of MD simulation may indicate that the 2D space in the Ramachandran plot needs to be partitioned into more detailed subdivided regions such as the empirical potential derived by Shortle et al.[35] Also, it may be crucial, for avoiding steric hindrance, to consider the coupling between adjacent torsion angles, namely, not only the pair (ϕ_I,ψ_I) but also the pair (ϕ_I,ψ_{I+1}) may need to couple explicitly. Moreover, these potential terms may need to depend on the rotamer conformation $(\phi_I,\psi_I,r_k(A(I)))$, in which side-chain torsion angle are included. These, of course, need a more complex form of the energy function. In contrast, the fragment assembly method can circumvent this problem because of a predefined local structure library, while MD requires a still more accurate energy function for the description of local structures. This may be one of the reasons why the groups using fragment assembly have been so successful in recent CASP exercises.[36]

As mentioned already, the quality of decoy structures is crucial for better energy function derived from Z score optimization. The gapless threading method used for producing decoys in this article, albeit concise, does not give good decoys. Those obtained from the fragment assembly or MD simulations can be used for further improving the potential, which will be the basis for future work.

Although in this work we solely used the physical energy function for the single (target) sequence, for the practical purpose of the structure prediction we can also use other forms of bioinformatic guidance. For example, the sequence information of a protein's family available in the multiple sequence alignments and the profile can be utilized. Also, using results of the secondary structure predictions, in general, improves the prediction performance, as has been demonstrated by other groups. We tested including the information of the secondary structure predictions via the Ramachandran potential terms and found the accuracy of prediction to be improved, even if the bias from the secondary structure prediction is weak (data not shown). Bias from the secondary structure prediction for a residue can be considered as taking account of the coupling with adjacent residues implicitly. Restraints extracted from threading are also possible.[37] The information from the contact predictions can be converted into protein-specific pair-wise potential.

We, however, emphasize that having a physical energy function usable for a single sequence has several advantages over those containing bioinformatic innovations even where the latter is helpful for the practical purposes of structure prediction. First, bioinformatic score functions would seem to be inherently applicable only to natural sequences, while a purely physical energy function can be applied for any sequence that, for example, includes random sequences that do not fold to a unique structure. More important, the physical energy function can be applied for designing artificial sequences that fold to a target structure even though target structures have not yet been observed. Indeed, the previous version of our energy function[7] has already been used for the de novo design of a small globular protein that was synthesized in the laboratory and showed both secondary and tertiary structures compatible with the target ones.[38] Second, ultimately only a physical energy function can give a completely microscopic understanding to the principles of protein architecture, one of our ultimate goals. Third, the physical energy function, possibly after significant modification, can be used for analyzing protein–protein interactions, protein–substrate binding, and membrane–protein folding.

ACKNOWLEDGMENTS

The work of S.T. was supported, in part, by a Grant-in-Aid for scientific research on priority areas (C) ”Genome Information Science” from the Ministry of Education, Science, Sports and Culture of Japan and ACT, Japan Science and Technology Corp. The work of Z.A.L-S. and P.G.W. was supported by NIH Grant PHS 2R01GM44557-10. The authors also thank NSF and JSPS for the U.S.–Japan cooperative science program.

REFERENCES

1. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84: 7524–7528.
2. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
3. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
4. Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. Proc Natl Acad Sci USA 1998;95:2932–2937.
5. Hardin C, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Associative memory hamiltonians for structure prediction without homology: alpha-helical proteins. Proc Natl Acad Sci USA 2000;97:14235–14240.
6. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations. II. Parameterization short-range interactions and determination of weights of energy terms by Z-score optimization. J Comput Chem 1993;18:353–366.
7. Takada S. Protein folding simulation with solvent-induced force field: folding pathway ensemble of three-helix-bundle proteins. Proteins 2001;42:85–98.
8. Takada S, Luthey-Schulten Z, Wolynes PG. Folding dynamics with nonadditive forces: a simulation study of a designed helical protein and a random heteropolymer. J Chem Phys 1999;110: 11616–11629.
9. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
10. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
11. Samudrala R, Levitt M. Decoys ’R’ Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
12. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
13. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314:141–151.
14. Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
15. Miyazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
16. Tobias DJ, Sneddon SF, Brooks CL III. Reverse turns in blocked dipeptides are intrinsically unstable in water. J Mol Biol 1990;216: 783–796.
17. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
18. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79–96.
19. Fauchere JL, Pliska V. Hydrophobic parameters π of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides. Eur J Med Chem 1983;18:369–375.

20. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 1978;47:45–148.
21. Simons KT, Strauss C, Baker D. Prospects for ab initio protein structural genomics. J Mol Biol 2001;306:1191–1199.
22. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001;45(suppl 5):127–132.
23. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258: 367–392.
24. Fain B, Xia Y, Levitt M. Design of an optimal Chebyshev-expanded discrimination function for globular proteins. Protein Sci 2002;11:2010–2021.
25. 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.
26. Henikoff S, Henikoff JG. Amino acid substitution matrices. Adv Protein Chem 2000;54:73–97.
27. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
28. Betancourt MR, Skolnick J. Finding the needle in a haystack: educing protein native folds from ambiguous ab initio folding predictions. J Comput Chem 2001;22:339–353.
29. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;45(suppl 5):13–21.
30. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
31. Hess B, Bekker H, Berendsen HJ, Fraaije JG. LINCS A linear constraint solver for molecular simulatons. J Comput Chem 1997;18:1463–1472.
32. Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 2001;60:96–123.
33. Fukunishi H, Watababe O, Takada S. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. J Chem Phys 2002;116: 9058–9067.
34. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
35. Shortle D. Composites of local structure propensities: evidence for local encoding of long-range structure. Protein Sci 2002;11:18–26.
36. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): Round IV. Proteins 2001;45(suppl 5):2–7.
37. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;(suppl 3):177–185.
38. Jin W, Kambara O, Sasakawa H, Tamura A, Takada S. De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. Structure 2003;11: 581–590.