A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. II. Parameterization of Short-Range Interactions and Determination of Weights of Energy Terms by *Z*-Score Optimization

A. LIWO,[1,2]* M. R. PINCUS,[3] R. J. WAWAK[1], S. RACKOVSKY,[2] S. OŁDZIEJ,[4] H. A. SCHERAGA[1]

[1]*Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301*

[2]*Department of Biomathematical Sciences, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, New York 10029*

[3]*Department of Pathology, Brooklyn Veterans Administration Medical Center, Brooklyn, New York 11209 and State University of New York, Health Science Center, Brooklyn, New York 11203*

[4]*Department of Chemistry, University of Gdańsk, ul. Sobieskiego 18, 80-952 Gdańsk, Poland*

Received 7 June 1996; accepted 11 September 1996

ABSTRACT: Continuing our work on the determination of an off-lattice united-residue force field for protein-structure simulations, we determined and parameterized appropriate functional forms for the local-interaction terms, corresponding to the rotation about the virtual bonds (U_{tor}), the bending of

*On leave from the Department of Chemistry, University of Gdańsk, Gdańsk, Poland, 1994–1995 and 1996.

Correspondence to: H. A. Scheraga; e-mail: has5@cornell.edu

This article includes Supplementary Material available from the authors upon request or via the Internet at ftp.wiley.com/public/journals/jcc/suppmat/18/874 or http://www.journals.wiley.com/jcc

Contract grant sponsor: National Institute on Aging, contract grant number: AG 00322

Contract grant sponsor: National Institute of General Medical Sciences, contract grant number: GM-14312

Contract grant sponsor: National Science Foundation, contract grant number: MCB95-13167

Contract grant sponsor: National Cancer Institute, contract grant number: CA 42500

Contract grant sponsor: Polish State Committee for Scientific Research, contract grant number: PB 190/T09/96/10

CCC 0192-8651 / 97 / 070874-14

virtual-bond angles (U_b), and the energy of different rotameric states of side chains (U_{rot}). These terms were determined by applying the Boltzmann principle to the distributions of virtual-bond torsional and virtual-bond angles and side-chain rotameric states, respectively, calculated from a data base of 195 high-resolution nonhomologous proteins. The complete energy function was constructed by combining the individual energy terms with appropriate weights. The weights were determined by optimizing the so-called Z-score value (which is the normalized difference between the energy of the native structure and the mean energy of non-native structures) of the histidine-containing phosphocarrier protein from *Streptococcus faecalis* (1PTF; an 88-residue $\alpha + \beta$ protein). To accomplish this, a database of C^{α} patterns was created using high-resolution nonhomologous protein structures from the Protein Data Bank, and the distributions of energy components of 1PTF were obtained by threading its sequence through ~ 500 randomly chosen C^{α}-patterns from the X-ray structures in the PDB, followed by energy minimization, with the energy function incorporating initially guessed weights. The resulting minimized energies were used to optimize the Z-score value of 1PTF as a function of the weights of the various energy terms, and the new weights were used to generate new energy-component distributions. The process was iterated, until the weights used to generate the distributions and the optimized weights were self-consistent. The potential function with the weights of the various energy terms obtained by optimizing the Z-score value for 1PTF was found to locate the native structures of other test proteins (within an average RMS deviation of 3 Å): calcium-binding protein (4ICB), ubiquitin (1UBQ), α-spectrin (1SHG), major cold-shock protein (1MJC), and cytochrome b_5 (3B5C) (which included α and β structures) as distinctively lowest in energy in similar threading experiments. © 1997 by John Wiley & Sons, Inc. *J Comput Chem* **18**: 874–887, 1997

Keywords: protein structure prediction; united-residue representation of a polypeptide chain; potential of mean force; inverse folding; Z-score.

Introduction

In part I of this work,[1] hereafter referred to as paper I, we outlined a procedure for the determination of a united-residue potential-energy function for polypeptide chains. The general form of the energy function is given by eq. (1) of paper I; we delay, until paper III, consideration of the multibody term, U_{corr}, of eq. (1) of paper I. The different terms of the energy function, determined by analyzing the one- and pairwise distributions calculated from protein-crystal data and averaging the all-atom potentials, are assigned appropriate weights. In paper I, we carried out a statistical analysis of the two-body distributions of side-chain centroids and, based on these distributions, we determined several forms of the long-range side-chain interaction potentials ($U_{SC,SC}$). In this article, we describe the determination of the local part of the potential: the term accounting for the rotation about the virtual bonds (U_{tor}), the virtual-bond angle bending energy (U_b), and the energy of different rotameric states of the side chains (U_{rot}), by applying the Boltzmann principle and statistical analysis to the distributions of virtual bond angles, θ, virtual torsional angles, γ, and local geometry (i.e., within the frame of three consecutive α carbons) of the side-chain centroids. We also show that, after appropriate choice of the energy terms and their weights, the potential is capable of recognizing the native folds of a number of test proteins among the structures from the Protein Data Bank (PDB).

This article is organized as follows. In the "Methods" section, we first introduce the expressions for U_{tor}, U_b, and U_{rot} and outline the procedure for determining the parameters of these expressions. Then, we describe the determination of the weights by optimization of the Z-score[2] of a test protein, based on the conformational-state distribution calculated for the structures from the PDB. In the Results section, we first describe the determination of the parameters of individual energy expressions, and then the determination of the weights. Finally, we show the performance of the potential in inverse-folding tests.

Methods

ENERGY EXPRESSIONS

Torsional Potentials

For all pairs of residues except Pro–Pro we use the same Fourier expansion for the energy of rotation about the virtual bonds, as in our earlier studies.[3,4] For the rotation about the C^{α}_{X}—C^{α}_{Y} virtual-bond axis (X and Y denote the types of neighboring amino-acid residues), the torsional energy is defined by:

$$U_{tor;\,XY}(\gamma) = a_{\circ} + \sum_{i=1}^{n} [a_{i;\,XY}(\cos n\gamma + 1) + b_{i;\,XY}(\sin n\gamma + 1)] \qquad (1)$$

where n takes a value from 3 to 6, depending on X and Y ($XY \neq$ ProPro), a_i and b_i are the coefficients of the Fourier expansion [eqs. (1) and (2)] for the torsional energy for $i = 1, 2, \ldots, n$.

For the Pro–Pro pair, the distribution of the torsional angles, γ, calculated from the PDB is zero around 0° and 120°, which means that these regions would be inaccessible. To reflect this tendency, we introduced the following energy function:

$$\begin{aligned} U_{tor;\,ProPro}(\gamma) &= a_{\circ} + \sum_{i=1}^{3} [a_{i;\,ProPro}(\cos n\gamma + 1) \\ &\quad + b_{i;\,ProPro}(\sin n\gamma + 1)] \\ &\quad + \begin{cases} a_4 \dfrac{1 + \cos 3\gamma}{1 - \cos 3\gamma} & \text{for } -60^\circ < \gamma < 180^\circ \\ 0 & \text{for } -180^\circ \leq \gamma \leq -60^\circ \end{cases} \end{aligned} \qquad (2)$$

Based on a statistical analysis of the distribution of the virtual-bond torsional angles (see the Results section), we found that three different types of neighboring amino acids must be considered, the glycine type, the proline type, and the alanine type; the last stands for all amino-acid residues except glycine and proline.

Virtual Bond-Angle Bending

Levitt[5] found that there is some correlation between the virtual-bond angle θ and the virtual-bond torsional angles adjacent to it, γ_1 and γ_2 (Fig. 1).

For torsional angles, γ_1 and γ_2, corresponding to extended structures, the adjacent virtual bond angle θ takes on values greater than 90°, whereas for the torsional angles γ corresponding to helical and turn structures, θ, is close to 90°. We found that this correlation can be written as:

$$\theta_c = \theta_c^{\circ} + a_1 \cos\gamma_1 + a_2 \sin\gamma_1 + b_1 \cos\gamma_2 + b_2 \sin\gamma_2 \qquad (3)$$

where θ_c is the value of θ calculated from this correlation equation, and θ_c°, a_1, a_2, b_1, and b_2 are the parameters of this correlation equation. However, we found that the deviations from this correlation are not negligible (data not shown), and the distribution of virtual-bond angles, θ, calculated by assuming that the correlation holds strictly, differs significantly from the distribution calculated from the PDB. From Figure 2, for example, we concluded that the distribution of θ can be expressed as a sum of two Gaussian distributions, the first one corresponding to normal deviations from θ_c of the correlation equation, and the second one independent of θ_c and centered approximately around $\theta_{\circ} = 90^\circ$ [see eq. (4)], with the ratio between the two distributions depending on θ_c. We also noted (data not shown) that the standard deviations, α, of the virtual-bond angles, θ, from the correlation equation are different for different θ_c: the smallest deviations are observed for θ_c close to 90° (corresponding to helical structures), moderate deviations for θ_c around 120° (corresponding to β-sheet structures), and largest deviations for θ_c located between those two bordering values, where the secondary structure is weakly defined. We, therefore, propose the following expression for the

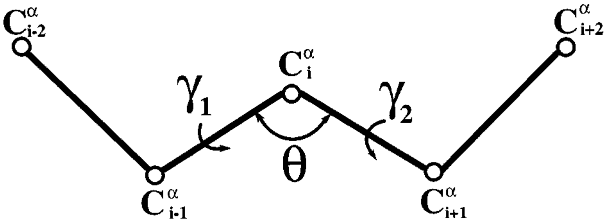

FIGURE 1. The two adjacent virtual-bond torsional angles, γ_1 and γ_2, that correlate with the central virtual-bond angle, θ.

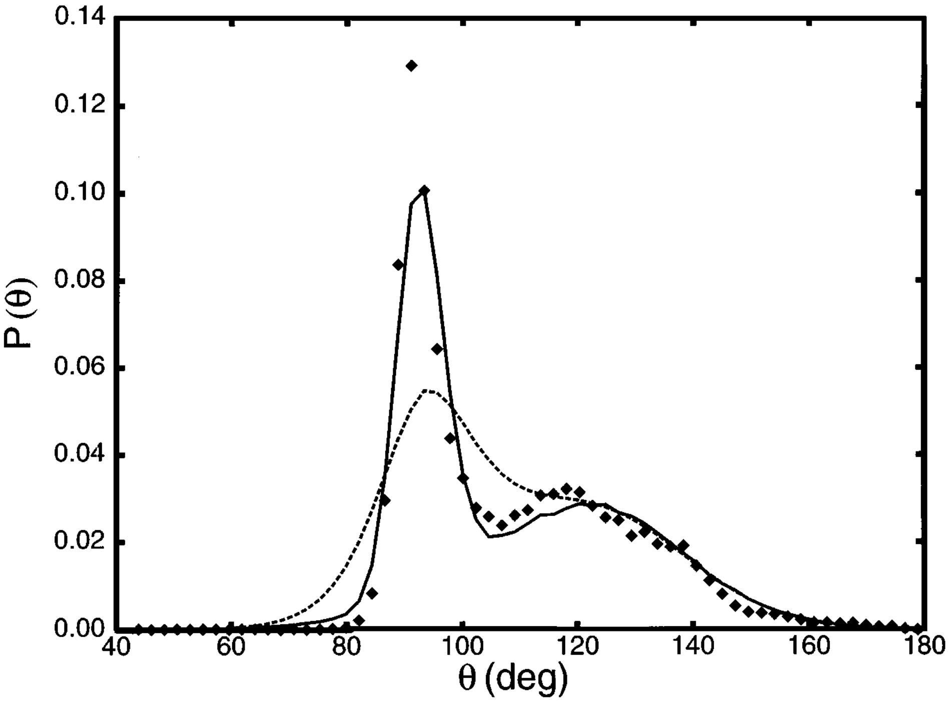

FIGURE 2. Comparison of the calculated (diamonds; based on the PDB) and simulated (lines) cumulative distribution of the virtual-bond angles, θ, of the nonglycine and nonproline residues. Solid line: distribution simulated using the full expression given by eq. (4). Broken line: distribution simulated assuming that the deviations from the correlation equation [eq. (3)] obey the normal distribution with omission of the second Gaussian in eq. (4).

distribution of the angles θ:

$$P(\theta \mid \gamma_1, \gamma_2) = \frac{1}{\sqrt{2\pi}[\sigma(\theta_c) + k(\theta_c)\sigma_\circ]} \times \left\{ \exp\left[-\frac{(\theta - \theta_c)^2}{2\sigma(\theta_c)^2} \right] + k(\theta_c)\exp\left[-\frac{(\theta - \theta_\circ)^2}{2\sigma_\circ^2} \right] \right\} \quad (4)$$

with:

$$\sigma(\theta_c) = 1 \Bigg/ \left[s_{\circ c}^2 + \left(\sum_{k=0}^{3} \alpha_k \theta_c^k \right)^2 \right] \quad (5)$$

$$k(\theta_c) = \exp\left[g_1 - \frac{(g_2 - \theta_c)^2}{2g_3^2} \right] \quad (6)$$

where $\theta_\circ$, $\sigma_\circ$, $s_{\circ c}$, $\{\alpha_k, k = 0, 1, \ldots, 3\}$, g_1, g_2, g_3 are residue-type-specific parameters determined from protein-crystal data by means of the maximum likelihood principle (see "Determination of Parameters of the Energy Function" subsection). The factor $1/\sqrt{2\pi}[\sigma(\theta_c) + k(\theta_c)\sigma_\circ]$ normalizes the probability distribution to 1.0.

The virtual-bond bending energy is calculated from this distribution by applying the Boltzmann principle:

$$U_b(\theta, \gamma_1, \gamma_2) = -RT \log P(\theta \mid \gamma_1, \gamma_2) \quad (7)$$

Local Energy of Side-Chain Rotamers

A preliminary analysis of the distributions of the side-chain centroids calculated from the PDB revealed that the distances between the side-chain centroids and the corresponding α-carbons (b_{SC}) are almost constant for each residue type, but differ from residue to residue, whereas the two angles, α_{SC} and β_{SC}, that define the orientation of a side-chain centroid with respect to the C^{α} frame (see Fig. 1 of paper I for the definition of side-chain orientation) vary considerably and tend to group into clusters corresponding to different rotamers. Thus, the simplest analytical form of the distribution of side-chain rotamers would be a sum of two-dimensional Gaussians in α_{SC} and β_{SC}. However, we found that there is some coupling between the virtual-bond angles, θ, and the angles

α_{SC} and β_{SC} that belong to the same residue. Therefore, a correct expression for rotamer distribution should include three-dimensional Gaussians in θ, α_{SC}, and β_{SC}. Because θ is restricted in value[6] from 47° to 145° to avoid values of θ close to 0° and 180°, we found it more convenient to use $-\cot\theta$ instead of θ as a variable. Thus, the probability distribution of side-chain rotamers is given by:

$$\begin{aligned} P(\theta, \alpha_{SC}, \beta_{SC}) &= N^{-1} P(\theta) \\ &\times \sum_{i=1}^{nc} h_i \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_\mathbf{i}^\circ)^\mathbf{T}\mathbf{D}_\mathbf{i}(\mathbf{x}-\mathbf{x}_\mathbf{i}^\circ)\right] \end{aligned} \tag{8}$$

where nc is the number of Gaussians, h_i is the height of the ith Gaussian; for the sake of uniqueness we set $h_1 = 1$. $\mathbf{x} = (-\cot\theta, \alpha_{SC}, \beta_{SC})^T$ and $\mathbf{x}_\mathbf{i}^\circ = (-\cot\theta_i^\circ, \alpha_{SC;\,i}, \beta_{SC;\,i}^\circ)^T$ $(i = 1, 2, \ldots, nc)$ are the vectors of the variables and the vectors of the coordinates of the centers of the Gaussians, respectively, and:

$$\mathbf{D}_i = \begin{pmatrix} d_{i;11} & d_{i;12} & d_{i;13} \\ d_{i;21} & d_{i;22} & d_{i;23} \\ d_{i;31} & d_{i;32} & d_{i;33} \end{pmatrix} \quad i = 1, 2, \ldots, nc \tag{9}$$

are the symmetric dispersion matrices of the Gaussians, N is the normalization constant, $P(\theta)$ is the distribution of the virtual-bond angles, and the superscript "**T**" denotes the transposition of a matrix or a vector.

The adjustable parameters are the heights of the Gaussians, h_i $(i = 1, 2, \ldots, nc)$, the vectors of the coordinates of the centers of the Gaussians $\mathbf{x}_\mathbf{i}^\circ$ $(i = 1, 2, \ldots, nc)$, and the elements of the *symmetric* dispersion matrices $\mathbf{D}_i$ $(i = 1, 2, \ldots, nc)$ (a total of six independent elements for each matrix).

The normalization constant, N, is calculated by integrating the un-normalized distribution over the entire variable range: $[0 \le \theta \le \pi] \times [0 \le \alpha_{SC} \le \pi] \times [0 \le \beta_{SC} \le 2\pi]$. The integration over α_{SC} and β_{SC} can be carried out analytically; to simplify the expressions, we extended the respective integration ranges from $-\infty$ to ∞ because the half-widths of the Gaussians are much narrower than the range of the variation of variables. Because of the factor $P(\theta)$, the integration over θ must be carried out numerically. The cumulative probability distributions $P(\theta)$ corresponding to different types of amino-acid residues were calculated from the PDB. Therefore, the normalization constant is expressed by eq. (10), after integrating over α_{SC} and β_{SC} of eq. (8):

$$\begin{aligned} N &= 2\pi \int_{\theta=47^\circ}^{\theta=145^\circ} P(\theta) \sum_{i=1}^{nc} \frac{h_i}{\sqrt{\det \mathbf{D}'_\mathrm{i}}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[d_{11;\,i} - \mathbf{d}_{\mathbf{i};\mathbf{1}}^\mathbf{T}\mathbf{D}'_\mathbf{i}\mathbf{d}_{\mathbf{i};\mathbf{1}} \right.\right. \\ &\quad \left.\left. \times (\cot\theta - \cot\theta_i^\circ)^2\right]\right\} d\theta \\ &\approx 2\pi\Delta\theta \sum_{k=1}^{n_\theta} P(\theta_k) \sum_{i=1}^{nc} \frac{h_i}{\sqrt{\det \mathbf{D}'_\mathrm{i}}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[d_{i;\,11} - \mathbf{d}_{\mathbf{i};\mathbf{1}}^\mathbf{T}\mathbf{D}'_\mathbf{i}\mathbf{d}_{\mathbf{i};\mathbf{1}} \right.\right. \\ &\quad \left.\left. \times (\cot\theta_k - \cot\theta_i^\circ)^2\right]\right\} \end{aligned} \tag{10}$$

where

$$\mathbf{d}_{i;1} = \begin{pmatrix} d_{i;12} \\ d_{i;13} \end{pmatrix} \quad \mathbf{D}'_\mathrm{i} = \begin{pmatrix} d_{i;22} & d_{i;23} \\ d_{i;32} & d_{i;33} \end{pmatrix}$$

To carry out the numerical integration, we divided the interval from $\theta = 47^\circ$ to $\theta = 145^\circ$ into bins of width $\Delta\theta = 2^\circ$.

It should be noted that a similar description of the distribution of side-chain rotameric states in terms of Gaussians, with the use of the dihedral angles χ_1 as variables, has recently been presented by Cheng et al.[7] for use with an all-atom representation of the polypeptide chain. The backbone dihedral angles ϕ and ψ were also included to describe the local conformational states; thus, the resulting distribution functions contained three variables simultaneously. This approach is very much like that summarized by eq. (8), in which not only the side-chain variables, α_{SC} and β_{SC}, but also the virtual-bond angles, θ, appear. As in the present work, the parameters of the distribution were determined from protein-crystal data.

To calculate the local energy of side-chain rotamers, U_{rot}, we must remove from eq. (8) the factor, $P(\theta)$, which is already accounted for by eq. (4). Thus, the local energy of side-chain rotamers is calculated from eq. (11):

$$\begin{aligned} U_{rot}(\theta, \alpha_{SC}, \beta_{SC}) &= -RT \log P(\theta, \alpha_{SC}, \beta_{SC} | \theta) \\ &= -RT \log \frac{P(\theta, \alpha_{SC}, \beta_{SC})}{P(\theta)} \end{aligned} \tag{11}$$

Peptide-Group (U_{pp}), Side-Chain Peptide-Group (U_{SCp}), and Side-Chain ($U_{SC,SC}$) Potentials

For the first two energy terms, we used the energy expressions and parameters from our earlier work[3,4]: namely, U_{pp} was obtained by averaging the all-atom ECEPP/2 energy function,[8,9] and U_{SCp} was assigned an arbitrary excluded-volume repulsive potential of the functional form[4] r_{SCp}^{-6}, where r_{SCp} is the distance between the side-chain centroid and peptide-group center. The side-chain interaction potential, $U_{SC,SC}$, was assigned the radial Lennard–Jones-type potential [eqs. (2) and (3) of paper I] with parameters of Table 2a and 2b of the Supplementary Material to paper I.

DETERMINATION OF PARAMETERS OF THE ENERGY FUNCTIONS

Choice of Protein-Structure Database and Calculation of Distributions

The same database of 195 high-resolution nonhomologous protein structures was used as in Table I of the Supplementary Material to paper I. The angles θ, γ, as well as the angles defining side-chain orientation, α_{SC} and β_{SC}, were calculated from the Cartesian coordinates of the α-carbons and side-chain centroids. The data corresponding to the cis peptide groups were dicarded, because our united-residue model is restricted to trans peptide groups.[1,3,4]

Torsional Potentials U_{tor}

The distributions, $P_{tor;\,XY}(\gamma)$, of all nine types (pairs of Gly, Ala, and Pro) of virtual-bond torsional angles, $\gamma\,(-180° < \gamma < 180°)$, from PDB data were collected in 36 bins of 10° width. Then, the torsional-energy curves were calculated from eq. (12), based on the Boltzmann principle:

$$\hat{U}_{tor;\,XY}(\gamma) = -RT\log[P_{tor;\,XY}(\gamma)] \qquad (12)$$

The constants in the analytical expressions [eqs. (1) and (2)] were determined by least-square fitting of these expressions to energy values calculated from the distributions for each pair of the types of amino-acid residues:

$$\min \Phi = \sum_{i=1}^{36} \left[\hat{U}_{tor;\,XY}(\gamma_i) - U_{tor;\,XY}(\gamma_i)\right]^2 \qquad (13)$$

where $\hat{U}_{tor;\,XY}(\gamma_i)$ denotes the energy calculated from the distribution of eq. (12), and $U_{tor;\,XY}(\gamma_i)$ the energy from the appropriate analytical expression [eqs. (1) and (2)].

Virtual-Bond-Angle Bending and Side-Chain-Rotamer Potentials (U_b and U_{rot})

For these terms, the straightforward procedure for calculating distributions and then energies followed by least-square fitting of the analytical to the experimental energies was not feasible, because these multivariate distributions would be based on sparse data. We, therefore, applied the maximum-likelihood principle to determine parameters; in other words, the parameters were determined by maximizing the so-called log-likelihood function whose general form is given by[10]:

$$\max_{\mathbf{y}} \{\log L\} = \sum_{i=1}^{np} \log P(\mathbf{x}_i;\, \mathbf{y}) \qquad (14)$$

where $\mathbf{x}_i$ is the vector of variables defining the distribution for the ith datapoint: $\mathbf{x}_i = (\theta_i, \gamma_{1;i}, \gamma_{2;i})$ for $P(\theta \mid \gamma 1, \gamma 2)$, and $\mathbf{x}_i = (-\cot\theta_i, \alpha_{SC;i}, \beta_{SC;i})$ for $P(\theta, \alpha_{SC}, \beta_{SC})$, $\mathbf{y}$ is the vector of adjustable parameters (see subsections "Virtual-Bond Angle Bending" and "Local Energy of Side-Chain Rotamers" of section "Energy Expressions" for the specification of these parameters), and np is the number of datapoints.

This is equivalent to the minimization of $-\log L$ which, in turn, can be interpreted as the requirement that the sum of local energies over all the points collected from the protein-crystal data be the lowest. Minimization of $-\log L$ was carried out by using the secant unconstrained minimization solver (SUMSL) algorithm[11] with analytical derivatives.

DETERMINATION OF WEIGHTS OF THE ENERGY TERMS

Problem Formulation

Shakhnovich et al.,[12] Šali et al.,[13] and Goldstein et al., [14,15] as well as Hao and Scheraga,[16,17] carried out calculations that suggest that those protein sequences that fold are those for which the unique native conformation is separated from non-native folds by a sufficiently large energy gap. It can, therefore, be expected that such proteins can fold rapidly to their native conformations, even though the folding process consists of random moves. A measure of the difference between the native and non-native conformations is the so-called Z-score value defined by[2]:

$$Z = \frac{E_\circ - 1/N\,\Sigma_{i=1}^{N} E_i}{\sqrt{1/N\,\Sigma_{i=1}^{N} E_i^2 - 1/N^2(\Sigma_{i=1}^{N} E_i)^2}} \qquad (15)$$

where N is the number of conformations, $E_\circ$ is the energy of the native conformation, and E_i is the energy of the ith non-native conformation.

The value of the Z-score is the normalized difference between the energy of the native conformation and the mean energy of the quasicontinuous energy distribution corresponding to non-native structures. The more negative the Z-score values, the more the native structure is distinguished from non-native ones.

Because the individual energy terms are known in our approach, the Z-score is a function of only the weights of the energy terms. We formulate the problem of determining the weights as minimizing the Z-score of a chosen test protein, for which the native structure is known, as a function of weights:

$$\min_{\omega_{el}, w_{tor}, w_{loc}} Z(w_{el}, w_{tor}, w_{loc}) \qquad (16)$$

where w_{el}, w_{tor}, and w_{loc} are the weights of the electrostatic, torsional, and local (virtual-bond angle bending and side-chain rotamer) energy; see eq. (1) of paper I.

Generation of Energy Distribution

These calculations were needed for determining the weights and for evaluating the performance of the final potential in inverse-folding experiments (see "Results" section). The conformational space of the test protein was assumed to be represented by a number of C^α structures from the PDB, including its native structure—this is the inverse-folding approach.[2] Because our energy function is designed to work with global-minimization algorithms that require local energy minimization, we generated the energy distributions using energy-minimized structures. Energy minimization is also a convenient tool to obtain relaxed structures without steric overlaps. It should be noted that, for the latter purpose, other techniques such as Monte Carlo dynamics, as implemented in the inverse-folding algorithm of Godzik et al.,[2,18] are also appropriate. On the other hand, local energy minimization is particularly efficient to obtain a relaxed structure quickly, because it moves the structure approximately along the steepest-descent direction of the energy surface. The structures were taken from the data base of 195 protein structures described in Table 1 of the Supplementary Material to paper I, supplemented by the 38 high-resolution structures of Table 1 (of the Supplementary Material) to this article that were not considered in paper I, because these chains contain less than 100 amino acid residues, but are suitable for the generation of energy distributions. The structures were continuous portions of the chains not containing cis peptide groups. In summary, there were 507 C^α motifs in the data base.

For given weights, the energies were calculated by threading the sequence of the test protein through randomly chosen structures from the structure data base. First, a chain was chosen from the database of the C^α patterns at random. If the length of the selected chain was shorter than that of the test protein, the chain was discarded and another selection was made; otherwise, the position of the first C^α on which to superpose the test sequence on the chain was chosen at random. Then a short 1000-step Metropolis Monte Carlo simulation was carried out with sampling of the angles α_{SC} and β_{SC} from the distributions given by eq. (8), and considering the total energy of the united-residue chain in the Metropolis acceptance criterion. This Metropolis simulation can be regarded as the preliminary equilibration of the positions of the side chains given the C^α trace. To obtain the energy-minimized structure corresponding to the structure selected at random from the PDB, we carried out constrained energy minimization using the target function defined by:

$$f = U + W \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} d(C_i^\alpha C_j^\alpha) \qquad (17)$$

where U is the energy of the polypeptide chain given by eq. (1) of paper I, $d(C_i^\alpha C_j^\alpha)$ is the distance between the α-carbons of the ith and the jth residue of the selected structure, and W is the weight of the distance term.

For each selected structure, we carried out three consecutive minimizations with $W = 0.1$, 0.01, and finally 0. We will hereafter refer to this procedure as the threading-with-minimization procedure. The minimized energies were collected for 400–600 structures with the final value of $W = 0$. The minimization was carried out using the SUMSL algorithm.[11]

Because the threading-with-minimization calculations for each individual structure are independent of each other, the procedure for computing and collecting the minimized energies was an ideal target for parallelization. Typically, we ran the threading-with-minimization calculations with 8–32 processors of the IBM-SP2 supercomputer.

Choice of Protein to Determine the Weights of Energy Terms

We examined monomeric proteins from the PDB for which the chain length was shorter than 100 amino-acid residues. This was determined primarily by the feasibility of the threading-with-minimization calculations, but another reason was not to include the structures that were used to derive the individual terms $U_{SC,SC}$, U_b, U_{tor}, etc. in the potential. We eliminated from this list all structures with missing coordinates, those with resolutions exceeding 2 Å, and those containing cis peptide groups (because our united-residue model does not treat them[1,3,4]) or disulfide bonds (because the additional bridge-closure constraints could be incompatible with most of the C^{α} structures from the data base). After this selection, we concluded that the most suitable protein to determine the weights of the various energy terms is the 88-residue histidine-containing phosphotransferase from *Streptococcus faecalis*, 1PTF; its structure was determined at a resolution of 1.6 Å, and this protein contains both α-helical and β-sheet domains (Fig. 3).[19]

Iterative Calculation of Weights

After a threading-with-minimization calculation was accomplished for the phosphotransferase (1PTF), the weights of the energy terms were calculated by minimizing the *Z*-score given by eq. (15), assuming that the individual contributions to the energies, $U_{SC,SC}$, U_{SCp}, U_{pp}, U_{tor}, U_b, and U_{rot}, are constant, and the energies vary only because the weights change. However, because different weights define different energy functions, the energy-minimized structures corresponding to energy functions with different weights are not the same, although they can be assumed to be accept-

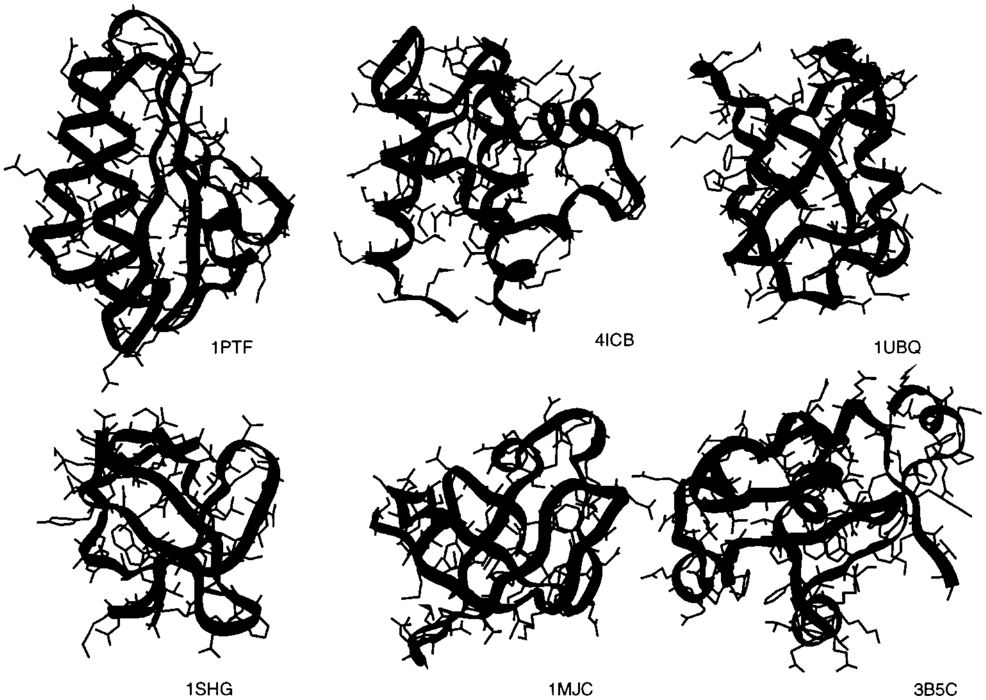

FIGURE 3. The structures of the proteins used in inverse-folding calculations: histidine-containing phosphocarrier protein (1PTF), calcium-binding protein (4ICB), ubiquitin (1UBQ), α-spectrin (1SHG), major cold-shock protein (1MJC), and cytochrome b5c (3B5C).

ably similar (we carried out only *local* energy minimizations, and the RMS deviations between the C^{α} traces of the structures corresponding to different weights were usually not greater than 3 Å). Thus, the individual contributions to the energies, $U_{SC,SC}$, U_{SCp}, U_{pp}, U_{tor}, U_{b}, and U_{rot}, will depend on the weights, although, because of the similarity of the energy-minimized structures mentioned above, it can be assumed that they vary more slowly with the weights than the total energy does. For this reason, the new weights, obtained after optimizing the Z-score assuming fixed values of individual energy contributions, were used to carry out a new threading-with-minimization calculation which gave new energies that were again used to determine the weights. The process was iterated, until the weights were self-consistent.

Results

VIRTUAL-BOND TORSIONAL POTENTIALS

To estimate the number of distinct torsional types of amino-acid residues, we calculated the correlation coefficients between the distributions of the virtual-bond torsional angles of the X—Aa pairs where X denotes any, and Aa a given, type of amino acid. We also define diagonal correlation coefficients, $r_{Aa,\,Aa}$, as the coefficients of the correlation between the distributions of the type X—Aa and type Aa—X virtual-bond torsional angles. The off-diagonal and diagonal correlation coefficients are expressed by eqs. (18) and (19), respectively:

$$r_{Aa1,\,Aa2} = \frac{\sum_{i=1}^{36}\left(P_{i;\,X-Aa1} - \bar{P}_{X-Aa1}\right)\left(P_{i;\,X-Aa2} - \bar{P}_{X-Aa2}\right)}{\sqrt{\sum_{i=1}^{36}\left(P_{i;\,X-Aa1} - \bar{P}_{X-Aa1}\right)^2 \sum_{i=1}^{36}\left(P_{i;\,X-Aa2} - \bar{P}_{X-Aa2}\right)^2}} \tag{18}$$

$$r_{Aa,\,Aa} = \frac{\sum_{i=1}^{36}\left(P_{i;\,X-Aa} - \bar{P}_{X-Aa}\right)\left(P_{i;\,Aa-X} - \bar{P}_{Aa-X}\right)}{\sqrt{\sum_{i=1}^{36}\left(P_{i;\,X-Aa} - \bar{P}_{X-Aa}\right)^2 \sum_{i=1}^{36}\left(P_{i;\,Aa-X} - \bar{P}_{Aa-X}\right)^2}} \tag{19}$$

with $P_{i;\,X-Y}$ being the probability density of finding the virtual-bond dihedral angle γ in the interval $10° \times (i-1) \leq \gamma < 10° \times i$ and

$$\bar{P}_{X-Y} = 1/36 \sum_{i=1}^{36} P_{i;\,X-Y}$$

The results are summarized in Table I. As shown, the off-diagonal correlation coefficients are particularly low when Aa=Gly and quite low (of order of 0.7–0.9) when Aa=Pro. In all other cases except His—Asn, the coefficients take a value of 0.9 or higher. We, therefore, decided to distinguish three torsional types of amino acids: glycine, proline, and alanine; the latter stands for all nonglycine and nonproline amino acids. The diagonal correlation coefficients are low when considering the X—Gly (Gly—X) and the X—Pro (Pro—X) angles, and close to 1 when considering all other types of angles. This means that the Ala—Gly distribution and, further, the Ala—Gly potential differs significantly from the Gly—Ala distribution and potential, respectively, and the Ala—Pro potential differs from the Pro—Ala potential. The same result was obtained in our earlier work[3] based on averaging the all-atom ECEPP/2 potential. It is, therefore, reasonable to distinguish nine sets of torsional potentials corresponding to the X —Y pairs, where X and Y can be any of the three torsional types of the amino acids.

The calculated values of the Fourier coefficients are summarized in Table 2 of the Supplementary Material and the fit of the analytical torsional potentials [eqs. (1) and (2)] to the potentials calculated from the torsion angle distributions [eq. (12)] is shown in Figure 4. As shown, the data follow quite closely the curves except for the Gly—Gly and the Gly—Pro cases for which there were sparse data for calculating the curves. For Gly—Gly, we constrained all the coefficients, b_i, in eq. (1) to zero, because the energy surface is symmetric in the virtual-bond torsional angle γ in this case.

VIRTUAL-BOND ANGLE BENDING POTENTIALS

We defined 20 types of virtual-bond angle potentials, $U_{b;\,X}$ [eq. (7)], depending on the type X of the amino-acid residue at the vertex of the angle. The initial values of the coefficients in the expan-

TABLE I.

Correlation Coefficients Between the *X*—*Aa* Virtual-Bond Torsional Angle Distributions Calculated from the PDB. *X* Denotes Any and *Aa* a Given Amino-Acid Residue.[a]

	Cys	Met	Phe	Ile	Leu	Val	Trp	Tyr	Ala	Gly	Thr	Ser	Gln	Asn	Glu	Asp	His	Arg	Lys	Pro
Cys	0.89																			
Met	0.93	0.92																		
Phe	0.96	0.98	0.98																	
Ile	0.95	0.96	0.98	0.98																
Leu	0.95	0.97	0.98	0.95	0.96															
Val	0.95	0.98	0.98	0.99	0.97	0.98														
Trp	0.93	0.95	0.95	0.93	0.95	0.95	0.93													
Tyr	0.95	0.99	0.99	0.98	0.98	0.99	0.95	0.95												
Ala	0.95	0.94	0.97	0.93	0.96	0.93	0.94	0.95	0.97											
Gly	0.26	0.23	0.25	0.14	0.26	0.17	0.13	0.21	0.27	0.41										
Thr	0.92	0.95	0.95	0.90	0.95	0.91	0.93	0.93	0.98	0.35	0.91									
Ser	0.89	0.92	0.92	0.85	0.94	0.89	0.92	0.90	0.95	0.40	0.97	0.95								
Gln	0.88	0.91	0.92	0.86	0.93	0.89	0.92	0.91	0.96	0.30	0.95	0.97	0.90							
Asn	0.90	0.92	0.91	0.88	0.90	0.90	0.91	0.91	0.90	0.33	0.90	0.92	0.89	0.69						
Glu	0.95	0.94	0.95	0.91	0.95	0.93	0.95	0.93	0.97	0.28	0.97	0.96	0.96	0.91	0.97					
Asp	0.89	0.95	0.94	0.90	0.94	0.92	0.92	0.94	0.94	0.25	0.94	0.93	0.94	0.95	0.92	0.83				
His	0.84	0.85	0.87	0.80	0.88	0.85	0.84	0.85	0.90	0.41	0.89	0.90	0.90	0.78	0.89	0.84	0.81			
Arg	0.92	0.88	0.90	0.86	0.90	0.87	0.94	0.87	0.94	0.23	0.95	0.93	0.90	0.88	0.97	0.89	0.86	0.93		
Lys	0.89	0.88	0.90	0.84	0.94	0.86	0.93	0.88	0.96	0.30	0.95	0.96	0.96	0.86	0.94	0.91	0.89	0.94	0.91	
Pro	0.77	0.74	0.77	0.66	0.80	0.70	0.80	0.72	0.87	0.36	0.87	0.89	0.88	0.75	0.87	0.81	0.89	0.90	0.93	0.88

[a] The first column represents the virtual-torsional-angle distribution for *X*—*Aa*1, and the headings at the top represent the distribution for *X*—*Aa*2. The off-diagonal entries represent the correlation coefficients between *X*—*Aa*1 and *X*—*Aa*2, whereas the diagonal entries correspond to *X*—*Aa* and *Aa*—*X*.

sion of θ_c [eq. (3)] were calculated by linear least-square fitting of eq. (3) to points $(\theta, \gamma_1, \gamma_2)$ obtained from the protein-crystal data. Then $-\log L$ of eq. (14) was minimized for each of the 20 types of amino-acid residues. The resulting parameters of the virtual-bond angle bending potential are summarized in Table III of the Supplementary Material.

The necessity of including the Gaussian, independent of θ_c, in the expression for $P(\theta \mid \gamma_1, \gamma_2)$ [eq. (4)] is illustrated in Figure 2, where the distribution of the virtual-bond angles, θ, obtained from the PDB is compared with the distribution simulated assuming that there are only normal deviations from the correlation, $\theta_c = f(\gamma_1, \gamma_2)$ [the first Gaussian in eq. (4)], and with the bimodal distribution given by eq. (4), respectively. The distribution was simulated by sampling the virtual-bond torsional angles γ_1 and γ_2 from the values calculated from the PDB. As shown, the assumption of only normal deviations from the correlation results in strong underestimation of the height of the first peak centered around $\theta = 90°$, which corresponds to α-helical structures.

SIDE-CHAIN ROTAMER POTENTIAL

The numbers of Gaussian terms in the distributions of the side-chain rotamers of the 20 types of amino acids were estimated initially by inspecting the two-dimensional distributions of the angles α_{SC} and β_{SC} defining local geometry of the side-chain centroids, taking into account the periodicity of the second angle. The positions and dispersion matrices of the Gaussians were estimated initially by carrying out a minimum-variance cluster analysis[20] of the sets of angles α_{SC} and β_{SC}. The initial values of $\theta°$ were assigned as 90°. Then, the parameters of eq. (8) were optimized by minimizing the corresponding $-\log L$ function [eq. (14)]. If the distribution of the α_{SC} and β_{SC} angles was diffuse, the number of Gaussians was difficult to determine (e.g., for Cys and Arg); hence, a series of minimizations of $-\log L$ was carried out by gradually increasing the number of Gaussians until the $-\log L$ function did not decrease by more than 10%. The parameters of the side-chain distributions and, consequently, the energy expressions for the 20 amino acids are summarized in Table 4 of

TABLE II.
The Values of b_{SC} for the 20 Types of Amino-Acid Residues.

Type	b_{SC} (Å)
Cys	1.237
Met	2.142
Phe	2.299
Ile	1.776
Leu	1.939
Val	1.410
Trp	2.605
Tyr	2.484
Ala	0.743
Gly	0.000
Thr	1.393
Ser	1.150
Gln	2.240
Asn	1.684
Glu	2.254
Asp	1.709
His	2.113
Arg	3.020
Lys	2.541
Pro	1.345

the Supplementary Material. The values of the "virtual side-chain bond lengths," b_{SC} (averaged over the PDB for *each* type of residue), are summarized in Table II.

DETERMINATION OF WEIGHTS

The history of the iterative determination of the weights of the various energy terms by threading-with-minimization calculations on the phosphocarrier protein (1PTF) is shown in Table III. The initial weights were chosen to correspond to the relations between the energy terms of our earlier force field.[3,4] In each threading-with-minimization calculation, about 500 randomly chosen structures, including the native structure of the phosphocarrier protein (1PTF), were considered. As shown, the Z-score decreased quickly and, moreover, the energy difference between the native structure and the lowest-energy non-native structure amounts to 14 kcal/mol with the final weights. The energy distribution for the final weights, shown in Figure 5, also indicates that the native structure is well separated from the quasicontinuous ensemble of non-native structures. The RMS deviation of the energy-minimized 1PTF structure from the native structure is 2.1 Å. For the non-native structures, we found that the final RMS deviations of the energy-minimized structures from the starting structure ranged from 6 to 14 Å. This indicates that, although energy minimization can move a structure quite far away from the starting PDB structure, the native structure remains stable in our potential.

INVERSE-FOLDING TESTS OF THE POTENTIAL

Using the weights determined from the inverse-folding calculations on the phosphocarrier protein (1PTF), we checked the ability of the potential to locate the native structures of other proteins correctly, using the threading-with-minimization approach. Table IV summarizes the results of these calculations for a number of monomeric proteins with lengths exceeding 50 amino-acid residues. As shown, in each case, the native structure is the lowest in energy and is separated from non-native structures by a significant energy gap. The table also includes weights optimized specifically for each protein after the appropriate series of threading-with-minimization calculations. As shown, optimization of weights in this manner did not result in major decreases of the Z-score values or the energy differences between the native and lowest energy non-native conformation, which means that weights obtained by determining the weights of the energy terms using the phosphocarrier protein (1PTF) are also relevant for other proteins. It should be noted that none of the above proteins was used in parameterization of the potential.

We note that, as far as inverse folding is concerned, the discrimination of the native structure from non-native structures is included in the "hydrophobic-interaction" term, $U_{SC,SC}$ [cf. eq. (1) of paper I]. For 1PTF and all the proteins summarized in Table IV, this term makes the dominant contribution to the energy. For example, for the energy-minimized structure of 1PTF, the second dominant electrostatic-interaction term, U_{pp}, is only 18% of $U_{SC,SC}$, U_{tor} is 9%, $U_{SC,p}$ is 7%, U_b is 2%, and U_{rot} is 1%. None of the *individual* energy terms alone except $U_{SC,SC}$ distinguishes the native structure from non-native structures, although they enlarge the energy gap when combined together.

It should also be noted that the Z-score values obtained in our calculations are higher (approximately − 4) compared with the values obtained with the potentials designed for on-lattice inverse folding[18] in which Z-score values corresponding to native structures range from −10 to −20. This

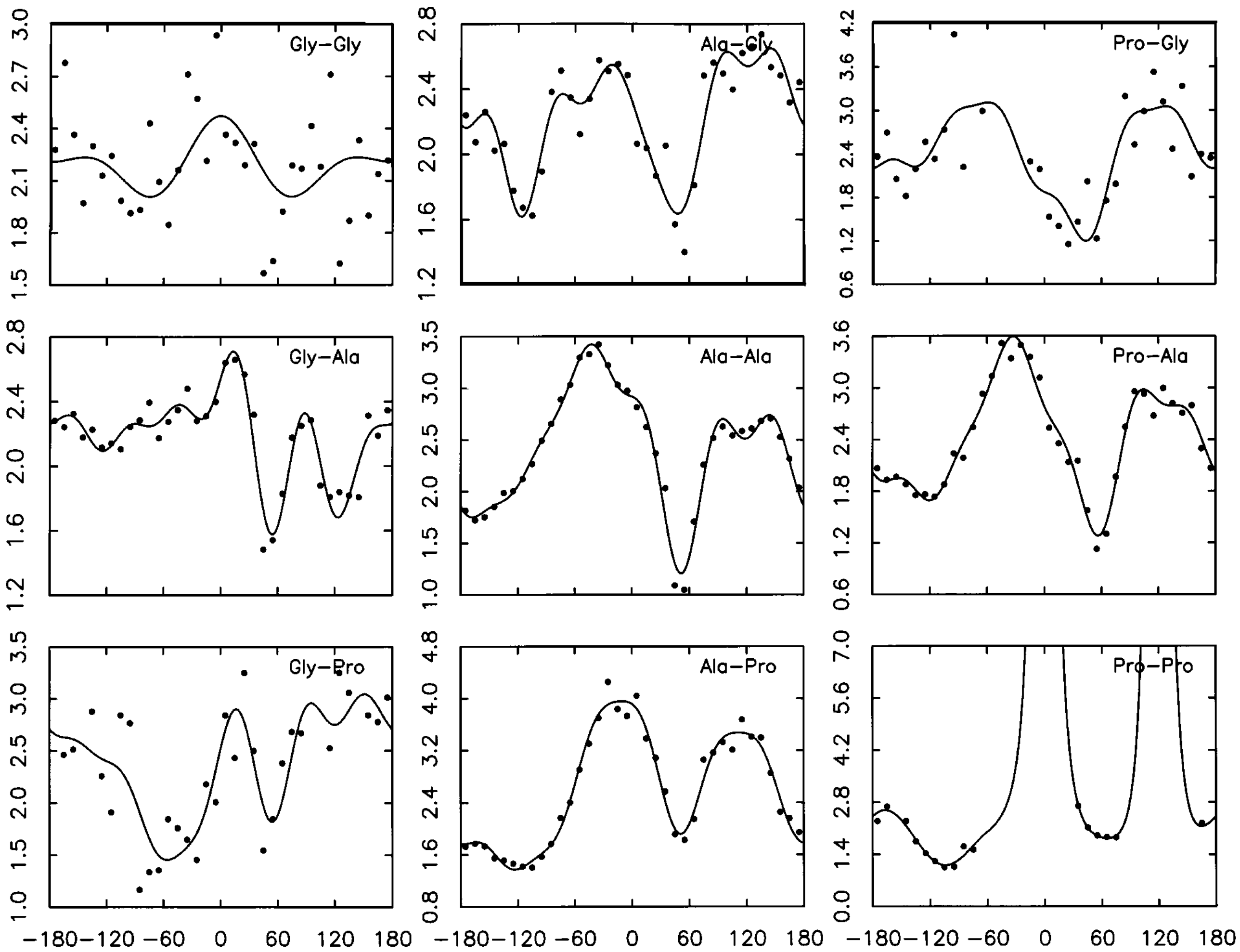

FIGURE 4. Fit of the analytical torsional energy curves [eqs. (1) and (2)] (solid lines) to torsional energies calculated from the virtual-bond torsional angle γ distribution calculated from the PDB (filled circles). Abscissae: γ; ordinates: torsional energy in RT units.

TABLE III.
Iterative Determination of Self-Consistent Weights of Energy Terms for 1PTF.

Iteration	$w_e^{\circ\,a}$	$w_l^{\circ\,a}$	$w_t^{\circ\,a}$	$\Delta E_{\mathrm{nat}}{}^{b}$ (kcal / mol)	Z	$w_e^{*\,a}$	$w_l^{*\,a}$	$w_t^{*\,a}$	$\Delta E_{\mathrm{nat}}{}^{b}$ (kcal / mol)	Z
0	1.450	0.626	1.692	+20.0	−1.88	0.249	0.050	0.271	−25.8	−4.67
1	0.610	0.186	0.846	−16.8	−4.26	0.341	0.112	0.403	−21.0	−4.66
2	0.341	0.117	0.403	−23.5	−4.41	0.650	0.190	0.000	−26.4	−4.74
3	0.495	0.150	0.201	−14.4	−3.80	0.347	0.201	0.091	−12.7	−3.87
4	0.421	0.175	0.147	−13.4	−3.78	0.444	0.144	0.042	−14.0	−3.80

[a] w°, w^{*} denote the initial weights used to carry out the threading-with-minimization calculations of a given iteration and the final weights optimized in this iteration, respectively. To avoid oscillations, the initial weights of the next iteration were arithmetic means of the weights optimized in two preceding iterations.

[b] $\Delta E_{nat} = E_{nat} - \min_{i \neq \mathrm{nat}}\{E_i\}$, where the latter term is the minimal element in the set of energies of all non-native structures.

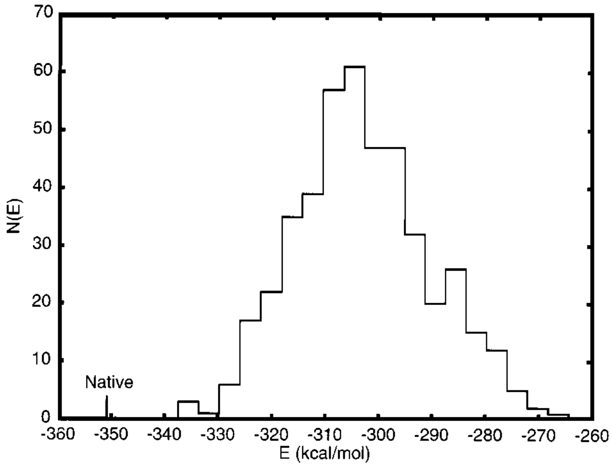

FIGURE 5. Distribution of energy (kilocalories per mole) of phosphocarrier protein (1PTF) obtained from the threading-with-minimization calculations with the final weights of the various energy terms.

might be due to the fact that we have energy-minimized all structures, regardless of the *Z*-score value (in other works, structures are not relaxed or only the best-scoring structures are relaxed by Monte Carlo dynamics[18]). Therefore, the energies of non-native structures approach closer to the energy of the native structure, thus resulting in increasing the *Z*-score value.

Discussion and Conclusions

The results reported in this study show that it is possible to combine the individual terms of the united-residue energy function determined from protein-crystal data or by averaging the all-atom potentials to produce a folding potential, at least in the inverse-folding simulations. However, the distribution of conformational states used to derive the weights was restricted in our calculations to the structures from the PDB, and need not be representative of the entire conformational space which also includes structures that never occur in native proteins. Especially, the weights of local-interaction terms can be wrong, because the PDB structures contain mostly organized structures, such as α-helices and *β*-sheets, for which the local interactions are already favorable. This was the reason that we could neglect, for the moment, the multibody term in eq. (1) of paper I, which is responsible for the stabilization of secondary structure.[2] On the other hand, the choice of PDB structures for the preliminary tests of the potential was justified, because they should correspond to low-energy structures; hence, the inverse-folding experiments would most probably fail to locate the native structures of the test proteins as distinctively lower in energy than the non-native structures if the component energy terms were completely wrong and could not be combined to produce a folding potential. In conclusion, the force field presented in our two articles, at the present stage of development, is suitable for inverse-folding calculations, but its use in *de novo* simulations of protein structure should be considered with caution.

To obtain a folding potential for *de novo* simulations, the distribution of conformational states used to derive the energy-term weights must include patterns devoid of secondary structure, calculated by Monte Carlo sampling capable of extensive covering of the conformational space. The recently

TABLE IV.
Summary of Control Threading Experiments Using All Test Proteins with Weights Determined Using the Phosphocarrier Protein (1PTF) (Initial Weights of Iteration 4 of Table III).

				Initial[c]		Optimized [d]		
Protein[a]	N[b]	Type	Cofactor	ΔE_{nat}	*Z*-score	ΔE_{nat}	*Z*-score	RMS[e]
4ICB	76	α	Ca^{2+}	–24.6	–4.84	–27.6	–5.08	4.5
1UBQ	76	$\beta + \alpha$	none	–15.5	–3.29	–10.1	–3.68	3.1
3B5C	85	$\alpha + \beta$	heme, Fe^{2+}	–13.5	–3.56	–14.8	–3.99	3.1
1SHG	57	*β*	none	–5.2	–3.06	–5.2	–3.78	2.5
1MJC	69	*β*	none	–3.9	–3.40	–7.8	–3.82	2.5

[a] See Table I for the names of these proteins.
[b] The number of amino-acid residues.
[c] Values calculated using the final weights obtained for the phosphocarrier protein (1PTF) ($\omega°$ of iteration 4 from Table III).
[d] Values calculated from the weights optimized using the energies obtained in threading-with-minimization calculations for a given protein.
[e] RMS deviation from the native structure (Å).

developed entropic-sampling method,[21] as implemented by Hao and Scheraga,[22] seems to be suitable for this purpose. The multibody term, U_{corr}, must also be included to obtain a sufficiently stable secondary structure. These two issues are now being addressed in our laboratory. The anisotropic forms of the potential of paper I, which were not used in inverse-folding calculations, might also be necessary to obtain proper packing patterns.

Acknowledgments

This work was supported by Grant AG 00322 from the National Institute on Aging (to S.R.), by Grants GM-14312 from the National Institute of General Medical Sciences (to H.A.S.), by Grant MCB95-13167 from the National Science Foundation (to H.A.S.), by Grant CA 42500 from the National Cancer Institute (to M.R.P.), and by grant PB 190/T09/96/10 from the Polish State Committee for Scientific Research (KBN) (to A.L. and S.O.).

The computations were carried out with up to 32 processors of the IBM-SP2 computer at the Cornell National Supercomputer Facility, a resource of the Center for Theory and Simulation in Science and Engineering at Cornell University, which is funded by the National Science Foundation, New York State, the IBM Corporation, and members of its Corporate Research Institute, with additional funds from the National Institutes of Health.

We thank Dr. M.-H. Hao for helpful discussions and advice regarding *Z*-score optimization.

References

1. A. Liwo, S. Ołdziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. S. Scheraga, *J. Comput. Chem* (accompanying article).
2. A. Godzik, A. Koliński, and J. Skolnick, *J. Comput.-Aid. Mol. Des.*, **7**, 397 (1993).
3. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *Prot. Sci.*, **2**, 1697 (1993).
4. A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, *Prot. Sci.*, **2**, 1715 (1993).
5. M. Levitt, *J. Mol. Biol.*, **104**, 59 (1976).
6. K. Nishikawa, F. A. Momany, and H. A. Scheraga, *Macromolecules*, **7**, 797 (1974).
7. B. Cheng, A. Nayeem, and H. A. Scheraga, *J. Comput. Chem.*, **17**, 1453 (1996).
8. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361 (1975).
9. G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883 (1983).
10. R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988, p. 13.
11. D. M. Gay, *Assoc. Comput. Math. Trans. Math. Software*, **9**, 503 (1983).
12. E. I. Shakhnovich and A. M. Gutin, *Proc. Natl. Acad. Sci. USA*, **90**, 7195 (1993).
13. A. Šali, E. Shakhnovich, and M. Karplus, *Nature*, **369**, 248 (1994).
14. R. A. Goldstein, Z. A. Luthey-Shulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, **89**, 4918 (1992).
15. R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, **89**, 9029 (1992).
16. M.-H. Hao and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, **93**, 4984 (1996).
17. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **100**, 14540 (1996).
18. A. Godzik, A. Koliński, and J. Skolnick, *J. Mol. Biol.*, **227**, 227 (1992).
19. Z. Jia, M. Vandonselaar, J. W. Quail, and L. T. J. Delbaere, *Nature*, **361**, 94 (1993).
20. H. Späth, *Cluster Analysis Algorithms*, Halsted Press, New York, 1980, p. 170.
21. J. Lee, *Phys. Rev. Lett.*, **71**, 211 (1993).
22. M.-H. Hao and H. A. Scheraga, *J. Phys. Chem.*, **98**, 9882 (1994).