JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

A Maximum-Likelihood Approach to Force-Field Calibration

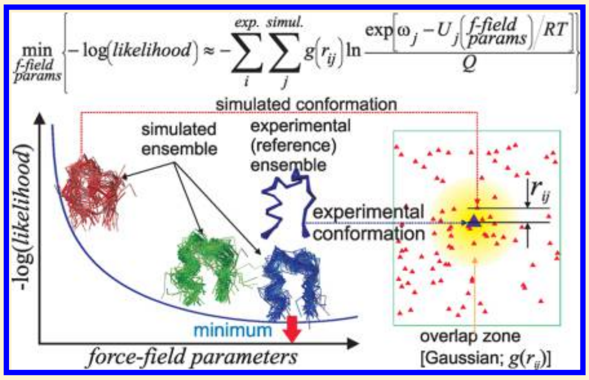
Bartłomiej Zaborowski,[†] Dawid Jagieła,[†] Cezary Czaplewski,[†] Anna Hałabis,[‡] Agnieszka Lewandowska,[‡] Wioletta Żmudzińska,[‡] Stanisław Ołdziej,[‡] Agnieszka Karczyńska,[†] Christian Omieczynski,[†] Tomasz Wirecki,[†] and Adam Liwo*,[†,§]

[†]Faculty of Chemistry, University of Gdańsk, ul. Wita Stwosza 63, 80-308 Gdańsk, Poland

[‡]Laboratory of Biopolymer Structure, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Kładki 24, 80-922 Gdańsk, Poland

[§]Center for In Silico Protein Structure and School of Computational Sciences, Korea Institute for Advanced Study, 87 Hoegiro, Dongdaemun-gu, Seoul 130-722, Republic of Korea

S Supporting Information


min f-field params {−log(likelihood) ≈ −Σ_i^exp. Σ_j^simul. g(r_ij) ln exp[ω_j − U_j(f-field params)/RT] / Q}
simulated conformation
simulated ensemble
experimental (reference) ensemble
experimental conformation
r_ij
-log(likelihood)
minimum
force-field parameters
overlap zone [Gaussian; g(r_ij)]

ABSTRACT: A new approach to the calibration of the force fields is proposed, in which the force-field parameters are obtained by maximum-likelihood fitting of the calculated conformational ensembles to the experimental ensembles of training system(s). The maximum-likelihood function is composed of logarithms of the Boltzmann probabilities of the experimental conformations, calculated with the current energy function. Because the theoretical distribution is given in the form of the simulated conformations only, the contributions from all of the simulated conformations, with Gaussian weights in the distances from a given experimental conformation, are added to give the contribution to the target function from this conformation. In contrast to earlier methods for force-field calibration, the approach does not suffer from the arbitrariness of dividing the decoy set into native-like and non-native structures; however, if such a division is made instead of using Gaussian weights, application of the maximum-likelihood method results in the well-known energy-gap maximization. The computational procedure consists of cycles of decoy generation and maximum-likelihood-function optimization, which are iterated until convergence is reached. The method was tested with Gaussian distributions and then applied to the physics-based coarse-grained UNRES force field for proteins. The NMR structures of the tryptophan cage, a small α-helical protein, determined at three temperatures (T = 280, 305, and 313 K) by Hałabis et al. (*J. Phys. Chem. B* **2012**, *116*, 6898−6907), were used. Multiplexed replica-exchange molecular dynamics was used to generate the decoys. The iterative procedure exhibited steady convergence. Three variants of optimization were tried: optimization of the energy-term weights alone and use of the experimental ensemble of the folded protein only at T = 280 K (run 1); optimization of the energy-term weights and use of experimental ensembles at all three temperatures (run 2); and optimization of the energy-term weights and the coefficients of the torsional and multibody energy terms and use of experimental ensembles at all three temperatures (run 3). The force fields were subsequently tested with a set of 14 α-helical and two $\alpha + \beta$ proteins. Optimization run 1 resulted in better agreement with the experimental ensemble at T = 280 K compared with optimization run 2 and in comparable performance on the test set but poorer agreement of the calculated folding temperature with the experimental folding temperature. Optimization run 3 resulted in the best fit of the calculated ensembles to the experimental ones for the tryptophan cage but in much poorer performance on the training set, suggesting that use of a small α-helical protein for extensive force-field calibration resulted in overfitting of the data for this protein at the expense of transferability. The optimized force field resulting from run 2 was found to fold 13 of the 14 tested α-helical proteins and one small $\alpha + \beta$ protein with the correct topologies; the average structures of 10 of them were predicted with accuracies of about 5 Å C$^\alpha$ root-mean-square deviation or better. Test simulations with an additional set of 12 α-helical proteins demonstrated that this force field performed better on α-helical proteins than the previous parametrizations of UNRES. The proposed approach is applicable to any problem of maximum-likelihood parameter estimation when the contributions to the maximum-likelihood function cannot be evaluated at the experimental points and the dimension of the configurational space is too high to construct histograms of the experimental distributions.

INTRODUCTION

The design of empirical energy functions for simulations of biologically important systems has a long history.[1] Both atomic-detailed[2−4] and coarse-grained[5−8] approaches are used. The

Received: June 20, 2015
Published: August 11, 2015

DOI: 10.1021/acs.jcim.5b00395

advantage of the first approach is greater accuracy, while the size of the system and time scale of the simulations are limited, although great progress was made in the field recently, especially by the Shaw group, as a result of the construction of the Anton supercomputer, which is dedicated to biomolecular simulations.[9,10] On the other hand, coarse-grained approaches offer a tremendous extension of the system size and time scale of the simulations,[11,12] at the expense of losing the details of the system under study.

A long-sought goal of the development of empirical force fields for biomolecular simulations is predictivity, understood as the ability to find the "native" structure of a biomolecule. This goal is particularly pursued for proteins for two reasons. One reason is the prediction of protein structures, and the other one is simulation of protein folding, functionally important motions, and conformational changes, which cannot be trusted if a force field cannot locate the native structure of any protein. Two types of approaches to protein-structure prediction are pursued. In the first one, the native structure is sought as the lowest-energy structure.[13−15] In the second one, it is understood as a family of structures that appear below the folding-transition temperature.[5,16,17] Although the first approach is more efficient in execution because efficient global-optimization methods of the target function can be used, such as the conformational-space annealing (CSA) method[18,19] and basin hopping,[20] it does not take the conformational entropy into account, which is considered in the second approach. Moreover, the force fields designed to predict protein structures understood as ensembles are good to study protein folding.

The predictive power of a force field is usually achieved by tuning of its parameters, which process is termed force-field optimization or calibration. The first such approach was proposed by Crippen and co-workers[21,22] and was based on maximization of the potential-energy gap between the lowest-energy native-like structure and the lowest-energy non-native structure. This principle was later utilized by Shakhnovich and co-workers[23] to propose a foldability criterion; however, Camacho and Thirumalai[24] subsequently demonstrated that energy-gap maximization is insufficient to produce a foldable energy landscape because it does not guarantee that folding will occur before the protein is trapped in a glassylike part of the landscape. Later, it was demonstrated that the ordering of the energies of the non-native structures according to native-likeness is critical for foldability.[25] At about the same time that energy-gap-optimization approaches were being developed, Wolynes and co-workers[26] designed a method based on maximization of the difference between the folding temperature and the glassy-transition temperature, which can be approximated by the *Z*-score, defined as the ratio of the difference between the mean energy of the native-like structures and the mean energy of non-native structures to the standard deviation of the energy of the non-native structures. These two principles, energy-gap and *Z*-score optimization, have been the basis of force-field calibration for many years and are especially efficient for threading (fold recognition),[27−29] in which case the native-like structure is unique and the set of non-native decoys is usually fixed.

For many years, we have been developing the physics-based coarse-grained UNRES force field for the simulation of protein structure and dynamics.[12,17,30−46] This force field is based on the expansion of the potential of mean force of polypeptide chains in water into Kubo cluster-cumulant functions,[47] enabling us to identify the respective effective energy terms with potentials of mean force of small model systems, which are comparatively easy to handle. The different contributions to the effective potential energy are then multiplied by weighting factors, termed energy-term weights, and added up to produce the complete effective energy function. These weights are primary targets of force-field calibration. Our first attempts at calibration of the UNRES force field were made by using *Z*-score-[31] and energy-gap optimization.[33,34] However, optimization was successful only for small simple training proteins, such as the three-helix-bundle staphylococcal protein A (PDB entry 1BDD). The reason for this was that these approaches ignore the ordering of non-native structures with different degrees of native-likeliness. This lack of ordering produces a glassylike rather than a funnel-like energy landscape, and the native structure is generally difficult to locate, even though it has a distinctively lower energy than those of non-native structures.

To overcome the problem of energy-landscape frustration, we developed the hierarchical-optimization method.[17,35,36] For each training protein, the conformations are grouped into *structural levels*. The conformations of each level contain the same native-like elements, but no conformation contains more native-like elements than the other ones in the group. The levels are subsequently ordered following the thermal-unfolding pathway determined experimentally. Optimization is directed at ordering the free energies of the structural levels according to the native-likeness of the levels. At temperatures lower than the folding-transition temperature, the higher the native-likeness of a level, the lower is its free energy. This order is opposite at temperatures higher than the folding-transition temperature, and at the folding-transition temperature, the free energies of all levels are equal. This idea has been used to design the target function, which is a sum of the penalties for violating the free-energy gaps between levels at different temperatures.[17,35,36]

By using this approach, we produced reasonably predictive versions of the UNRES force field for protein-structure prediction based on global optimization[35,36] as well as ensemble-based prediction and protein-folding and dynamics simulations.[17] The resulting force fields scored well in CASP exercises[48,49] and were applied with success to protein-folding simulations, including simulations of folding kinetics, studies of free-energy landscapes,[11,50−53] and investigations of biologically important processes such as PICK1 to BAR binding,[54] the Hsp70 chaperone cycle,[55] and, recently, modeling of the structure and stability of the complex between the iron−sulfur-binding protein 1 (Isu1) and the Jac1 Hsp40 cochaperone.[56] However, a disadvantage of the hierarchical-optimization method is that the decision as to whether a section of a simulated structure under consideration bears a resemblance to the corresponding section of the experimental structure (which is required in order to assign the conformation to a particular structural level) is based on arbitrary cutoffs, which are based on selected measures of native-likeness such as the fraction of secondary structure, the fraction of native contacts, the root-mean-square deviation (rmsd) from the native structure, etc.[35,36] For example, to decide that a conformation has a native α-helix, the fraction of $p_i \cdots p_{i+3}$ contacts (where p_i is the *i*th peptide group) in the section of the simulated conformation that corresponds to an α-helix in the experimental structure is calculated, and if that fraction is greater than a preassigned cutoff, the conformation is considered to have the α-helix (and consequently is assigned to a higher level of the hierarchy); otherwise, it is considered to be devoid of the helix (and consequently is assigned to a lower hierarchy level). Moreover, most of the target free-energy gaps between structural levels can be assigned only approximately, on the basis of the limited thermodynamic data for the fragments of the training proteins; only the free-energy gaps between the completely folded and completely unfolded structures are fully available from the experimental free energy of unfolding versus temperature [$\Delta G_f(T)$] curves.

Quite recently, by using the multiplexed replica exchange molecular dynamics (MREMD) approach[57,58] with UNRES,[59] we carried out a simulation study of the folding of protein A over a wide range of temperatures.[52] A surprising conclusion of this study was that at the folding transition temperature and about 50 K above it, the shape of the protein was largely preserved, except that about half of population had a mirror-image packing topology in which the native contacts or slightly shifted native contacts were present but the secondary structure was largely melted. We used the variant of the UNRES force field calibrated by our hierarchical method[17] with another three-α-helix-bundle protein (PDB entry 1GAB). Moreover, experimental studies of the mechanism for folding of the B1 domain of immunoglobulin-binding protein (IGG) have suggested that early structures along the folding pathway exhibit the general shape of the protein but that the secondary structure is not formed and the native contacts are shifted by one or two residues.[60] Furthermore, small-angle X-ray (SAXS) and neutron-scattering (SANS) studies of small proteins[61−63] have demonstrated that upon thermal denaturation, the radii of gyration of small proteins increase only very moderately near the folding-transition temperature[62,63] or even shrink with increasing temperature near the folding-transition point,[61] as opposed to chemical denaturation.[61−63] All of these observations are in agreement with the molten-globule concept of thermal denaturation of proteins[64] and strongly suggest that proteins achieve "rough" shape at quite high temperatures. Such a picture of the folding process is most close to the "hydrophobic collapse" mechanism of protein folding, albeit guided by favorable local interactions that initiate chain reversals in the right places;[60] these interactions can also have a hydrophobic nature.[65]

For the above reasons, we concluded that a rigid (cutoff-based) definition of structural levels does not reflect the folding pathway. Therefore, a better approach to force-field calibration than hierarchical optimization seems to be tuning of the force field so that the ensembles calculated at the temperatures of measurements fit the respective experimental ensembles. A natural method to tackle such a problem is the maximum-likelihood method.[66] However, as opposed to the textbook version of this approach, the probability-density function (pdf) cannot be evaluated exactly in the conformational analysis of proteins in the continuous space, even at a coarse-grained level. Instead, it is only available indirectly in the form of the conformational ensembles simulated at different temperatures. Because of the high dimension of the conformational space, the construction of histograms to represent the pdf is not feasible. Therefore, in this work, in order to use the simulated conformations to compute the likelihood function, we take the contribution from each of them with a Gaussian weight that depends on the distance from the respective experimental conformation. To assess the method, we use the experimental data from a recent experimental study[67] of a small α-helical protein, the tryptophan cage, at three temperatures that cover the folding transition. We test the resulting force fields (obtained by using various sections of the data and optimizing fewer or more parameters) with two sets of mostly α-helical proteins that differ in size and topology.

This paper is organized as follows. First, in the Theory section, we present the version of the maximum-likelihood fitting developed in this work and show that with a partition of the set of simulated conformations into native-like and non-native-like conformations and given a low-temperature limit, it reduces to the maximization of the energy gap between the native-like and non-native conformations. In Methods, we summarize the UNRES force field, the MREMD method[57,58] adapted to UNRES,[59,68] and the computational procedure implemented to execute the calibration of the UNRES force field by using the maximum-likelihood method. In the Results and Discussion, we first describe the tests of the maximum-likelihood approach with single- and multidimensional Gaussian distributions and subsequently the results of force-field calibrations with the tryptophan cage and tests of the resulting variants of the UNRES force field. Finally, in the Conclusions we point to further directions of the applications of the approach in force-field calibration and beyond.

THEORY

Maximum-Likelihood Approach. In the classical implementation of the maximum-likelihood approach, the target function, log L, is maximized (or −log L is minimized) to fit the parameters of the fitted pdf to the distribution of experimental points.[66] For n experimental points given in an m-dimensional space, whose coordinates are $\mathbf{X}_i = [x_{i1}, x_{i2}, ..., x_{im}]^T$ ($i = 1, 2, ..., n$), the maximum-likelihood problem can be formulated as given by eq 1:

$$\max_{z_1,z_2,...,z_\nu} \log L \equiv \min_{z_1,z_2,...,z_\nu} -\log L = \min_{z_1,z_2,...,z_\nu}\left[-\sum_{i=1}^{n} w_i \ln f(\mathbf{X}_i; z_1, z_2, ..., z_\nu)\right] \tag{1}$$

where w_i is the weight of the ith experimental point and f is the normalized pdf,

$$\int \cdots \int f(\mathbf{Y}; z_1, z_2, ..., z_\nu)\, dV_{\mathbf{Y}} = 1 \tag{2}$$

in which d$V_{\mathbf{Y}}$ is the volume element in the space of system variables.

The maximum-likelihood principle cannot be used for parameter estimation in a straightforward manner when f cannot be computed at the experimental points (analytically or numerically) but instead points can only be sampled according to f, a situation that takes place when the pdf is a function of a huge number of variables and there is no analytical expression for its integral over the configurational space. This is exactly the problem of the pdf of polymer conformations in continuous three-dimensional space, which is given by the Boltzmann law (eq 3):

$$P(\mathbf{X}_i)\, dV_{\mathbf{X}} = \frac{1}{Q}\exp\left[-\frac{U(\mathbf{X}_i)}{RT}\right] dV_{\mathbf{X}} \tag{3}$$

where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m]^T$ denotes the vector of the coordinates of the molecule (m being the dimension of this vector), $U(\mathbf{X}_i)$ denotes its energy, Q denotes the partition function, given by

$$Q = \int \cdots \int \exp\left[-\frac{U(\mathbf{X})}{RT}\right] \mathrm{d}V_{\mathbf{X}} \tag{4}$$

R denotes the universal gas constant, and T denotes absolute temperature. The Boltzmann factor $\exp[-U(\mathbf{X})/RT]$ can be evaluated analytically for any conformation, but the partition function Q cannot be evaluated in a deterministic manner unless a simple lattice representation of the polymer under study is considered. Moreover, the experimental conformations are usually high in energy because the refinement algorithms put much more emphasis on satisfying the experimental restraints than on finding a lower energy. This observation suggests that the neighborhoods of experimental conformations instead of the experimental conformations themselves should be considered in evaluating log L.

A set of low-energy conformations can be obtained from molecular simulations in which points from the configurational space are sampled according to the Boltzmann distribution with a given potential-energy function; a more robust approach involves the use of multicanonical sampling, such as a replica-exchange[57,58] or entropic-sampling[69] method or one of their hybrids.[70−72] All of these approaches lead to points sampled from a given distribution, hereafter denoted as $\rho(\mathbf{Y})$; in particular, this distribution might have the same functional form as the target distribution, $\rho(\mathbf{Y}) = f(\mathbf{Y}; z_1^\circ, z_2^\circ, ..., z_\nu^\circ)$, where z_1°, z_2°, ..., z_ν° are the initially guessed parameters of the distribution function. Obviously, except for the situation where the configurational space is discretized to a low-resolution lattice, simulations provide a set of points $\mathbf{Y}_i$ ($i = 1, 2, ..., N$, where N is the number of snapshots from a simulation), none of which is identical to any of the experimental points (conformations).

In what follows, we will consider l_N, the negative of the maximum-likelihood function normalized by division by W, the sum of the weights for all of the conformations (eqs 5 and 6):

$$l_{\mathrm{N}} = -\frac{1}{W} \log L \tag{5}$$

where

$$W = \sum_{i=1}^{n} w_i \tag{6}$$

It should be noted that the normalization factor is independent of the force-field parameters. Therefore, minimization of l_N in eq 5 gives the same results as the minimization of −log L shown in eq 1. As will become clear in Gaussian Representation of the δ Function below, the purpose of normalization is to avoid the trivial zero minimum of the approximate l_N; the resulting equations are also simpler when l_N is used instead of −log L.

Because the Boltzmann probabilities cannot be calculated directly at the experimental conformations, the simulated conformations have to be used. Let us assume that conformations identical to the experimental ones have been sampled (a situation that is possible in principle if a lattice representation of the conformational space is used). Because the conformations are drawn from a distribution $\rho(\mathbf{X})$ (which is not necessarily the Boltzmann distribution because non-Boltzmann sampling techniques are usually more efficient than canonical sampling), the contribution to eq 5 from conformation i would be

$$\Delta l_{\mathrm{N}}(\mathbf{X}_i) = \frac{1}{W} w_i \ln f(\mathbf{X}_i^{(s)}) = \frac{1}{W} \rho(\mathbf{X}_i) w_i \ln f(\mathbf{X}_i)$$

where $\mathbf{X}_i^{(s)}$ denotes the point obtained by sampling. Therefore, l_N cannot be computed directly from the contributions at the sampled points, but instead each contribution has to be divided by $\rho(\mathbf{X}_i)$.

To handle real cases in which the simulations do not yield conformations identical to the experimental ones, we express −log L using integrals over the configurational space given the sampling probability density $\rho(\mathbf{X})$, weighted by Dirac δ functions centered at the experimental points; likewise, to normalize −log L (i.e., to obtain l_N), we express the sum of weights, W, in terms of integrals over the Dirac δ functions. The resulting expression is given by eq 7. It should be noted that eq 7 gives a result identical to that of the parent expression given by eq 5 with −log L expressed by eq 1.

$$l_{\mathrm{N}} = \frac{1}{A} \int \cdots \int \sum_{i=1}^{n} \frac{\delta(\mathbf{X}_i - \mathbf{Y})\rho(\mathbf{Y})}{\int \cdots \int \delta(\mathbf{Y} - \mathbf{Y}')\rho(\mathbf{Y}')\, \mathrm{d}V_{\mathbf{Y}'}} w_i \ln f(\mathbf{Y}; z_1, z_2, ..., z_\nu)\, \mathrm{d}V_{\mathbf{Y}} \tag{7}$$

in which

$$A = W = \int \cdots \int \sum_{i=1}^{n} \frac{\delta(\mathbf{X}_i - \mathbf{Y})\rho(\mathbf{Y})}{\int \cdots \int \delta(\mathbf{Y} - \mathbf{Y}')\rho(\mathbf{Y}') dV_{\mathbf{Y}'}} w_i\, \mathrm{d}V_{\mathbf{Y}} \tag{8}$$

where $\delta(\mathbf{Y} - \mathbf{Y}')$ is the Dirac δ function: $\delta(\mathbf{Y} - \mathbf{Y}') = \infty$ for $\mathbf{Y} = \mathbf{Y}'$ and zero elsewhere with

$$\int \cdots \int \delta(\boldsymbol{\xi} - \boldsymbol{\xi}')\, \mathrm{d}V_{\boldsymbol{\xi}'} = 1 \tag{9}$$

and

$$\int \cdots \int \delta(\boldsymbol{\xi} - \boldsymbol{\xi}') f(\boldsymbol{\xi}')\, \mathrm{d}V_{\boldsymbol{\xi}'} = f(\boldsymbol{\xi}) \tag{10}$$

In eq 7, the numerators of the components of the sum in the integrand represent sampling of the points from distribution ρ, while the denominators represent the correction for the bias introduced by sampling the points from a nonuniform distribution.

Equation 7 is the basis of the modified maximum-likelihood method introduced in this work, from which its various implementations can be constructed by using different approximate representations of the Dirac δ function. In the next section we will show that with the hypercylinder representation of the δ function, the method becomes equivalent to the method of energy-gap maximization.

Relation between Maximum-Likelihood Optimization and Energy-Gap Optimization of Force Fields. Let us consider the application of the maximum-likelihood method to a situation in which the simulated conformations are divided into two classes, native-like (nat) and non-native (non-nat), this assignment depending on, e.g., the rmsd from the crystal or the average NMR structure. A conformation is considered native-like if its rmsd does not exceed a preassigned cutoff value s and non-native otherwise. This dissection can be considered as an application of a hypercylinder representation of the Dirac δ function in eq 7, as given by eq 11:

$$\delta(\mathbf{Y}-\mathbf{Y}') = \lim_{s\to 0}\begin{cases} \dfrac{\Gamma\left(\frac{m}{2}+1\right)}{\pi^{m/2}s^m} & \text{for } \|\mathbf{Y}-\mathbf{Y}'\| < s \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where m is the dimension of the conformational space and $\Gamma(x)$ is Euler's Γ function.

When the representation of the δ function given by eq 11 with a finite hypercylinder radius s is substituted into eq 7 and it is assumed that all of the experimental conformations have weights equal to 1, the approximation to l_N, denoted as l_N^*, is given by eq 12:

$$l_\mathrm{N}^* = -\frac{1}{N'}\left\{\ln \sum_{\substack{i=1\\ i\in\{\mathrm{nat}\}}}^{N'} \exp(-\beta U_i) - \ln\left[\sum_{\substack{i=1\\ i\in\{\mathrm{nat}\}}}^{N'} \exp(-\beta U_i) + \sum_{\substack{i=1\\ i\in\{\text{non-nat}\}}}^{N-N'} \exp(-\beta U_i)\right]\right\} \tag{12}$$

where $\beta = 1/RT$, N' is the number of native-like conformations, and N is the total number of conformations. Without loss of generality, the conformations can be ordered so that the first native-like conformation is the lowest-energy native-like conformation and the first non-native conformation is the lowest-energy non-native conformation.

In the low-temperature limit (large β), with the assumption that $U_1^\mathrm{nat} < U_1^\text{non-nat}$, l_N^* is approximated by eq 13,

$$\begin{aligned} l_\mathrm{N}^* &\approx \frac{1}{N'}\exp[-\beta(U_1^\text{non-nat} - U_1^\mathrm{nat})] \\ &= \frac{1}{N'}\exp[-\beta(\min_{i\in\{\text{non-nat}\}} U_i - \min_{i\in\{\mathrm{nat}\}} U_i)] \end{aligned} \tag{13}$$

which is obtained from eq 12 by expanding the natural logarithm terms in the Taylor series about 1 and neglecting all of the terms in the expansion except the one with the energy of the lowest-energy non-native conformation. This implies that l_N^* is minimized (or log L is maximized) when the energy gap between the native structure and the lowest-energy non-native structure is maximized; this is a long-known approach to force-field optimization.[21,33] When U_1^nat is always greater than $U_1^\text{non-nat}$ (i.e., when the energy function is not optimizable), the assumption that the temperature is very low results in eq 14,

$$l_\mathrm{N}^* \approx \frac{\beta}{N}(U_1^\mathrm{nat} - U_1^\text{non-nat}) \tag{14}$$

which means that the gap between the lowest-energy native-like structure and the lowest-energy non-native structure should be minimized or the gap between the lowest-energy non-native and the lowest-energy native-like structure should be maximized, which is the same result as that obtained by assuming that $U_1^\mathrm{nat} - U_1^\text{non-nat} < 0$.

Gaussian Representation of the δ Function. The hypercylinder representation of the δ function introduces a hard cutoff s on the distances of the simulated points from the experimental points. Therefore, to obtain the final form of the working approximation to l_N, we use the Gaussian representation of the δ function (eq 15):

$$\delta(\mathbf{Y}-\mathbf{Y}') = \lim_{s\to 0}(\sqrt{2\pi}\,s)^{-m}\exp\left(-\frac{\|\mathbf{Y}-\mathbf{Y}'\|^2}{2s^2}\right) \tag{15}$$

where m is the dimension of the $\mathbf{Y}$ space. By replacing the δ function in eq 7 with its Gaussian representation shown in eq 15, we obtain eq 16:

$$l_\mathrm{N} = \frac{1}{A}\lim_{s\to 0} \int\cdots\int \sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\mathbf{X}_i-\mathbf{Y}\|^2}{2s^2}\right)\rho(\mathbf{Y})}{\lim\limits_{s\to 0}\int\cdots\int\exp\left(\frac{\|\mathbf{Y}-\mathbf{Y}'\|}{2s^2}\right)\rho(\mathbf{Y}')\,\mathrm{d}V_{\mathbf{Y}'}} w_i \ln f(\mathbf{Y}; z_1, z_2, ..., z_\nu)\,\mathrm{d}V_\mathbf{Y} \tag{16}$$

in which

$$A = \lim_{s\to 0} \int\cdots\int \sum_{i=1}^{n} \frac{\exp\left(-\frac{\|\mathbf{X}_i-\mathbf{Y}\|^2}{2s^2}\right)\rho(\mathbf{Y})w_i\,\mathrm{d}V_\mathbf{Y}}{\lim\limits_{s\to 0}\int\cdots\int\exp\left(\frac{\|\mathbf{Y}-\mathbf{Y}'\|}{2s^2}\right)\rho(\mathbf{Y}'])\,\mathrm{d}V_{\mathbf{Y}'}} \tag{17}$$

In principle, the constants s in the numerator and denominator on the right-hand sides of eqs 16 and 17 could be different; however, we keep the same s for simplicity. By removing the condition $s \to 0$, as in Relation between Maximum-Likelihood Optimization and Energy-Gap Optimization of Force Fields above, and replacing the integrations over $\mathbf{Y}$ and $\mathbf{Y}'$ with summations over the N simulated points, we obtain eq 18, which is the desired approximate expression for l_N. It should be noted that these points are drawn from the ρ distribution function, so ρ is implicit in the obtained distribution of points and therefore no longer appears in the equation.

$$\begin{aligned} l_\mathrm{N}^* &= \frac{1}{A}\sum_{j=1}^{N}\sum_{i=1}^{n}\frac{\exp\left(-\frac{\|\mathbf{X}_i-\mathbf{Y}_j\|^2}{2s^2}\right)}{\sum_{k=1}^{N}\exp\left(-\frac{\|\mathbf{Y}_j-\mathbf{Y}_k\|^2}{2s^2}\right)} w_i \ln f(\mathbf{Y}_j; z_1, z_2, ..., z_\nu) \\ &= \sum_{j=1}^{N} a_j \ln f(\mathbf{Y}_j; z_1, z_2, ..., z_\nu) \end{aligned} \tag{18}$$

in which

$$A = \sum_{j=1}^{N}\sum_{i=1}^{n}\frac{\exp\left(-\frac{\|\mathbf{X}_i-\mathbf{Y}_j\|^2}{2s^2}\right)}{\sum_{k=1}^{N}\exp\left(-\frac{\|\mathbf{Y}_j-\mathbf{Y}_k\|^2}{2s^2}\right)} w_i \tag{19}$$

and

$$a_j = \frac{1}{A}\sum_{i=1}^{n}\frac{\exp\left(-\frac{\|\mathbf{X}_i-\mathbf{Y}_j\|^2}{2s^2}\right)}{\sum_{k=1}^{N}\exp\left(-\frac{\|\mathbf{Y}_j-\mathbf{Y}_k\|^2}{2s^2}\right)} w_i \tag{20}$$

The denominators of the components of the sum over the simulated points in eq 18 are corrections of the bias introduced by drawing these points from a predefined distribution $\rho(\mathbf{Y})$

and not the uniform distribution (which would be impossible in the case of conformational sampling). Together with the terms in the numerators, they provide weights with which the contributions from the simulated points are added to l_N^* at a given experimental point. The relation between the approximate and original maximum-likelihood functions is illustrated in Figure 1.

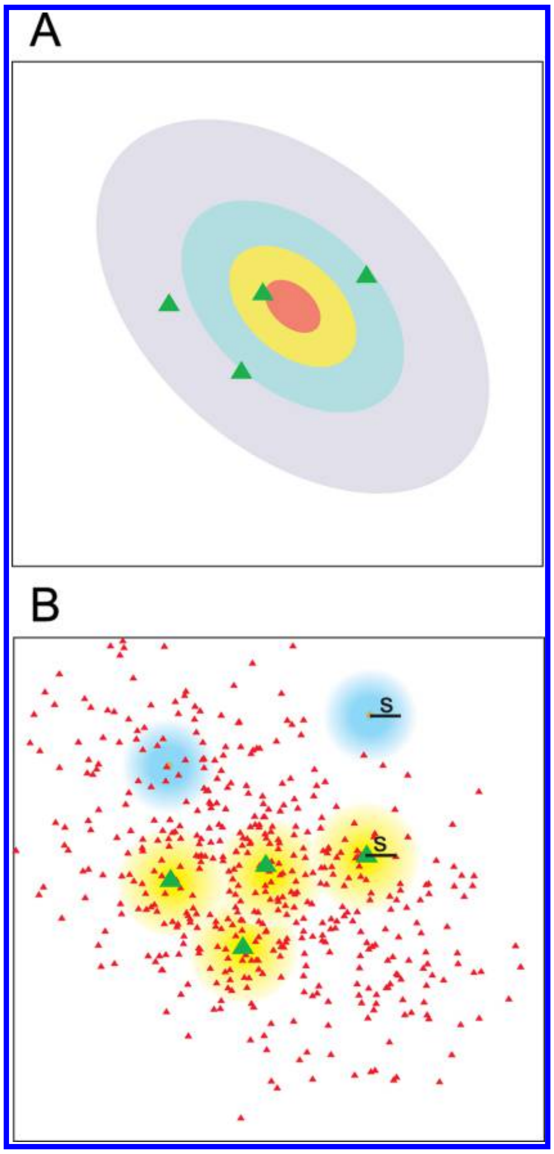


Figure 1. Illustration of the relationship between the analytical and approximate maximum-likelihood functions in the space of two variables. (A) The probability distribution function (pdf), represented as a contour plot, can be calculated exactly at the experimental points (large green triangles). (B) The pdf can be calculated only at the simulated points (small red triangles). The contributions to the approximate −log L function from the simulated points to the respective experimental points (eq 18) are weighted according to Gaussian functions centered at the experimental points and characterized by the standard deviation s (smeared yellow circles). To correct for the nonuniform sampling, the contribution from each simulated point is divided by the sum of the Gaussian weights with standard deviation s centered at a given point over all other simulated points; for two selected simulated points (marked as orange triangles), the Gaussians are illustrated by blue smeared circles.

The need to define the normalization factor in terms of A in eq 17 and not simply in terms of W (eq 6) is now apparent because the un-normalized right-hand side of eq 16 can be made arbitrarily close to zero (which is the smallest value that can be achieved given the fact that the pdf is always normalized to 1) if $\rho(\mathbf{Y})$ overlaps poorly with the distribution of the experimental points. Therefore, minimization of the un-normalized maximum-likelihood function, −log L, could easily result in a pdf that is divergent from the experimental points.

It should be noted that summation over the experimental points is on the top of the summation over the simulated points in eq 18, although the order of integration in eq 16 is the opposite. This modification is correct because the summations are carried out over a fixed range of indices. With this modification, the sums over the experimental points corresponding to given simulated points (a_i, i = 1, 2, ..., N; eq 20) can be computed at the beginning of the minimization procedure and used as constant weights when running a given iteration.

Equation 18, together with eqs 19 and 20, is a general formula for the variant of the maximum-likelihood-fitting approach proposed in this work, except that it assumes that the Gaussian representation of the Dirac δ function is used. It can be applied to any type of fitted distribution function f. Obviously, the Gaussian representation of the δ function can be replaced with any other representation, as was done above in Relation between Maximum-Likelihood Optimization and Energy-Gap Optimization of Force Fields, where the relationship between the proposed method and earlier approaches to force-field optimization was explored. The procedure for energy-function calibration based on eq 18 and the use of MREMD[57−59,68] to generate the decoys is introduced in Maximum-Likelihood Calibration of Macromolecular Energy Functions in Methods.

METHODS

UNRES Force Field. In the UNRES model,[30,38] a polypeptide chain is represented by a sequence of α-carbon atoms with united side chains (SCs) attached to them and peptide groups (p) positioned halfway between pairs of consecutive α-carbons (Figure 2). The effective energy function originates from the potential of mean force (PMF) of the chain constrained to a given coarse-grained conformation plus the surrounding solvent, which is expanded into Kubo cluster-cumulant functions[47] to obtain the effective energy terms as described in our earlier work.[32] The effective energy of a coarse-grained polypeptide chain is expressed by eq 21:

$$\begin{aligned} U = {} & w_{\mathrm{SC}} \sum_{i<j} U_{\mathrm{SC}_i\mathrm{SC}_j} + w_{\mathrm{SCp}} \sum_{i \neq j} U_{\mathrm{SC}_i\mathrm{p}_j} + w_{\mathrm{pp}}^{\mathrm{vdW}} \sum_{i<j-1} U_{\mathrm{p}_i\mathrm{p}_j}^{\mathrm{vdW}} \\ & + w_{\mathrm{pp}}^{\mathrm{el}} f_2(T) \sum_{i<j-1} U_{\mathrm{p}_i\mathrm{p}_j}^{\mathrm{el}} + w_{\mathrm{tor}} f_2(T) \sum_i U_{\mathrm{tor}}(\gamma_i) \\ & + w_{\mathrm{tord}} f_3(T) \sum_i U_{\mathrm{tord}}(\gamma_i, \gamma_{i+1}) + w_{\mathrm{b}} \sum_i U_{\mathrm{b}}(\theta_i) \\ & + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot}}(\alpha_{\mathrm{SC}_i}, \beta_{\mathrm{SC}_i}) + w_{\mathrm{bond}} \sum_i U_{\mathrm{bond}}(d_i) \\ & + \sum_{k=3}^{k_{\max}} w_{\mathrm{corr}}^{(k)} f_k(T) U_{\mathrm{corr}}^{(k)} + \sum_{k=3}^{k_{\max}} w_{\mathrm{turn}}^{(k)} f_k(T) U_{\mathrm{turn}}^{(k)} \end{aligned} \tag{21}$$

where θ_i is the backbone virtual-bond angle, γ_i is the backbone virtual-bond dihedral angle, α_i and β_i are the spherical angles

DOI: 10.1021/acs.jcim.5b00395
J. Chem. Inf. Model. 2015, 55, 2050−2070

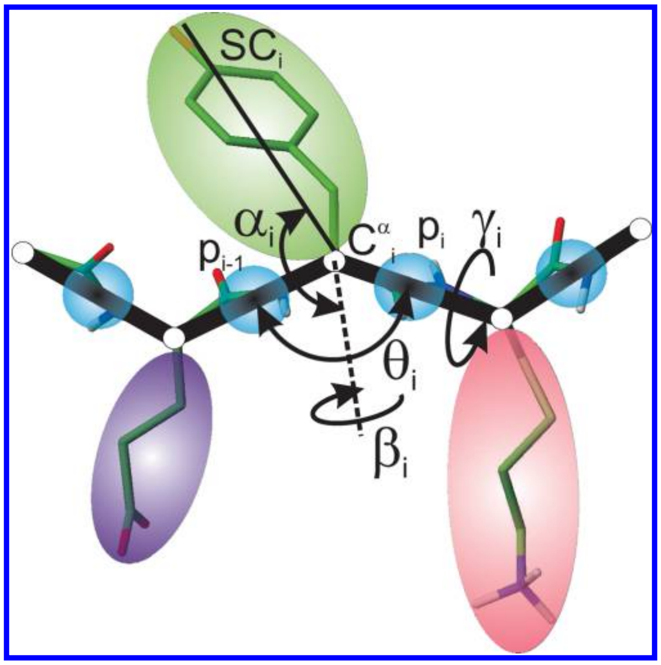


Figure 2. UNRES model of polypeptide chains. The interaction sites are united peptide groups (p) located between pairs of consecutive α-carbon atoms (light-blue spheres) and united side chains (SCs) attached to the α-carbon atoms (spheroids with different colors and dimensions). The backbone geometry of the simplified polypeptide chain is defined by the $C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond angles θ (the vertex of θ_i is at C^α_i) and the $C^\alpha \cdots C^\alpha \cdots C^\alpha \cdots C^\alpha$ virtual-bond dihedral angles γ (the axis of γ_i passes through C^α_i and C^α_{i+1}). The local geometry of the ith side-chain center is defined by the spherical angles α_i (the angle between the bisector of the respective angle θ_i and the $C^\alpha_i \cdots SC_i$ vector) and β_i (the angle of counterclockwise rotation of the $C^\alpha_i \cdots SC_i$ vector about the bisector from the $C^\alpha_{i-1} \cdots C^\alpha_i \cdots C^\alpha_{i+1}$ plane, starting from C^α_{i-1}). For illustration, the bonds of the all-atom chains, except for those to the hydrogen atoms connected to the carbon atoms, are superposed on the coarse-grained picture.

defining the location of the center of the united side chain of residue i, d_i is the length of the ith virtual bond, which is either a $C^\alpha \cdots C^\alpha$ virtual bond or a $C^\alpha \cdots SC$ virtual bond (Figure 2), and $k_{\max}$ is the maximum order of the correlation terms ($k_{max} = 4$ is used in this work, which was assessed to be sufficient in our earlier study[35,36]). Each term is multiplied by an appropriate weight, w_x, and the terms corresponding to factors of order higher than 1 are additionally multiplied by the respective temperature factors, which were introduced in our recent work[17] and reflect the dependence of the first generalized-cumulant term in those factors on temperature, as discussed in refs 17 and 73. The factors f_n are defined by eq 22:

$$f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\{\exp[(T/T_0)^{n-1}] + \exp[-(T/T_0)^{n-1}]\}} \quad (22)$$

where $T_0 = 300$ K.

The term $U_{SC_iSC_j}$ represents the mean free energy of the hydrophobic or hydrophilic interactions between the side chains, which implicitly contain the contributions from the interactions of the side chain with the solvent. The term $U_{SC_ip_j}$ denotes the excluded-volume potential of the side-chain—peptide-group interactions. The peptide-group interaction potential is split into two parts: the Lennard-Jones interaction energy between peptide-group centers ($U^{\mathrm{vdW}}_{p_ip_j}$) and the average electrostatic energy between peptide-group dipoles ($U^{\mathrm{el}}_{p_ip_j}$); the second of these terms accounts for the tendency to form backbone hydrogen bonds between peptide groups p$_i$ and p$_j$. U_{tor}, U_{tord}, U_{b}, U_{rot}, and U_{bond} are the virtual-bond dihedral angle torsional, virtual-bond dihedral angle double-torsional, virtual-bond angle bending, side-chain rotamer, and virtual-bond deformation terms, respectively; these terms account for the local propensities of the polypeptide chain. The terms $U^{(m)}_{\mathrm{corr}}$ represent correlation or multibody contributions from the coupling between backbone-local and backbone-electrostatic interactions, and the terms $U^{(m)}_{\mathrm{turn}}$ are correlation contributions involving m consecutive peptide groups and are therefore termed turn contributions. The multibody terms are critical for reproduction of regular α-helical and β-sheet structures.[32,74] The energy-term weights are determined by force-field calibration to reproduce the structures and folding thermodynamics of selected training proteins and were the primary optimization targets in this work. The effective energy terms are discussed in refs 32 and 38 and references cited therein.

Molecular Dynamics and Replica-Exchange Molecular Dynamics with UNRES. Molecular dynamics (MD), using the Langevin treatment to provide thermostatting, was implemented in UNRES to study protein-folding pathways and as the main component of conformational searching.[12,37,75] Because the coordinates of the centers of the peptide groups depend on those of the attached C^α atoms (Figure 2), the variables are the $C^\alpha \cdots C^\alpha$ and $C^\alpha \cdots SC$ virtual-bond vectors and not the Cartesian coordinates of the interacting sites. The MD implementation of UNRES is described in detail in our earlier work.[12,37] The equations of motion are integrated with the adaptive multiple-time split (A-MTS) modification[75] of the velocity Verlet algorithm.[76]

To cover the conformational space as completely as possible in runs with the tryptophan cage protein, which was used for force-field calibration, we carried out umbrella-sampling simulations with harmonic restraints on the quantity q, adapted from Wolynes and co-workers[77] in our earlier work,[17] which is a measure of the conformity between the distances in the simulated structure and in the reference structure (eq 23):

$$q = 1 - \frac{1}{n(n-2)} \sum_{i=1}^{n} \sum_{j=1}^{i-3} \left\{ \exp\left[-\frac{(d_{C_i^\alpha C_j^\alpha} - d^{\mathrm{ref}}_{C_i^\alpha C_j^\alpha})^2}{0.25(d^{\mathrm{ref}}_{C_i^\alpha C_j^\alpha})^2} \right] + \exp\left[-\frac{(d_{SC_iSC_j} - d^{\mathrm{ref}}_{SC_iSC_j})^2}{0.25(d^{\mathrm{ref}}_{SC_iSC_j})^2} \right] \right\} \quad (23)$$

where $d_{C_i^\alpha C_j^\alpha}$ and $d^{\mathrm{ref}}_{C_i^\alpha C_j^\alpha}$ are the distances between the ith and jth α-carbon atoms in the calculated and reference structures, respectively, and $d_{SC_iSC_j}$ and $d^{\mathrm{ref}}_{SC_iSC_j}$ are the distances between the centers of the ith and jth side chains in the calculated and reference structures, respectively. The reference structures are mean structures determined experimentally; details are provided below in Maximum-Likelihood Calibration of Macromolecular Energy Functions. Harmonic-restraint potentials of the form given by eq 24 are used:

$$U_i^{\mathrm{restr}}(q;\, q_i^\circ) = \frac{1}{2} k_i (q - q_i^\circ)^2 \quad (24)$$

where q_i° is the center and k_i the force constant of the ith restraint. The total potential energy is the sum of the UNRES energy and the restraint energy (eq 25):

$$V_i^{\text{tot}} = U^{\text{UNRES}} + U_i^{\text{restr}} \tag{25}$$

Canonical MD simulations, even when carried out at many different temperatures, are usually not sufficient to cover the whole conformational space, which is required for the generation of equilibrated ensemble-averaged quantities. For this reason, we used the multiplexed replica-exchange implementation of MD with respect to both temperature (MREMD)[57,58] and the Hamiltonian (HMREMD),[78] as adapted to UNRES in our earlier work.[59,68] In this approach, M canonical MD simulations are carried out simultaneously at different temperatures and with different energy functions (in this work, the energy functions differed with respect to the restraining potentials; eq 24). In regular replica exchange, each replica has a different temperature/energy function, while in k-plexed replica exchange, k trajectories share the same temperature/energy function. After every $m < M$ steps, an exchange of temperatures between neighboring trajectories ($j = i + 1$) is attempted, with the decision about whether the exchange is made being based on the Metropolis criterion, which is expressed by eq 26:

$$\begin{aligned}\Delta = {} & [\beta_j V_j(\mathbf{X}_j;\, \beta_j) - \beta_i V_i(\mathbf{X}_j;\, \beta_i)] \\ & - [\beta_j V_j(\mathbf{X}_i;\, \beta_j) - \beta_i V_i(\mathbf{X}_i;\, \beta_i)]\end{aligned} \tag{26}$$

where $\beta_i = 1/RT_i$, in which T_i is the absolute temperature corresponding to the ith trajectory, $\mathbf{X}_i$ denotes the variables of the UNRES conformation of the ith trajectory at the attempted exchange point, and V_i is the potential-energy function (including the restraining potential) corresponding to the ith trajectory (for temperature-only replica exchange, the restraining potentials are simply set equal to zero). If $\Delta \leq 0$, then $\{T_i, V_i\}$ and $\{T_j, V_j\}$are exchanged; otherwise, the exchange is performed with probability $\exp(-\Delta)$. It should be noted that the dependence of the UNRES energy on temperature is taken into account in eq 26.

To construct the conformational ensemble at a given temperature for a given energy function and to calculate ensemble-averaged quantities from the MREMD/HMREMD simulation results, the weighted histogram analysis method (WHAM)[79] is used; the implementation of this procedure to UNRES was developed in our earlier work.[17] Briefly, by solving a set of self-consistent equations (eqs 14−16 in ref 17), the quantity ω_i is calculated for each conformation, which enables us to compute the probability of the conformation at any temperature for the unrestrained energy function, as expressed by eq 27:

$$P(\mathbf{X}_i;\, T) = \frac{1}{Q(T)} \exp\left[\omega_i - \frac{U(\mathbf{X}_i;\, T)}{RT}\right] \tag{27}$$

in which $Q(T)$ is the configurational part of the partition function at absolute temperature T, given by

$$Q(T) = \sum_{k=1}^{N} \exp\left[\omega_k - \frac{U(\mathbf{X}_k;\, T)}{RT}\right] \tag{28}$$

where N is the total number of conformations in the simulations. It should be noted that the temperature should not be outside the range of temperatures used in the (H)MREMD simulation.

With eq 27, any ensemble-based property A, except for the ensemble-averaged energy and its derivatives, is calculated in a straightforward manner using eq 29:[17]

$$\langle A \rangle_T = \sum_{i=1}^{N} P(\mathbf{X}_i;\, T) A_i \tag{29}$$

The relative free energy of the ensemble is expressed by eq 30:

$$F(T) = -RT \ln Q(T) \tag{30}$$

Because the UNRES energy has the sense of the potential of mean force and consequently depends on temperature, the ensemble-averaged energy and heat capacity are not the simple average or the variance of the UNRES energy but include its temperature derivatives, as expressed by eqs 31 and 32, respectively:[17]

$$\begin{aligned}E(T) &= -RT^2 \frac{\partial}{\partial T} \ln Q(T) \\ &= \frac{1}{Q(T)} \sum_{i=1}^{N} \left[U(\mathbf{X}_i;\, T) - T\frac{\partial U(\mathbf{X}_i;\, T)}{\partial T}\right] \\ &\qquad \exp\left[\omega_i - \frac{U(\mathbf{X}_i;\, T)}{RT}\right] \\ &= \langle U - T\frac{\partial U}{\partial T} \rangle_T\end{aligned} \tag{31}$$

$$\begin{aligned}C_v(T) &= \frac{\partial}{\partial T} E(T) \\ &= -\left\langle T\frac{\partial^2 U}{\partial T^2}\right\rangle_T + \frac{1}{RT^2}\left\langle\left\langle\left[U - T\frac{\partial U}{\partial T}\right]^2\right\rangle\right\rangle_T\end{aligned} \tag{32}$$

Maximum-Likelihood Calibration of Macromolecular Energy Functions. As stated in the preceding section, we use the HMREMD method with a set of umbrella potentials (eq 24) that restrain the conformations to a given similarity to the reference structure(s)[17] in order to generate the conformations corresponding to a given set of energy parameters; processing the results with WHAM[79] enables us to calculate the probability of each simulated conformation at any temperature in the absence of restraints imposed during the simulations for sampling efficiency. Consequently, by inserting eq 27 into eq 18, we can express l_{N}^{*} by eq 33:

$$l_{\text{N}}^{*} = \ln Q(T) - \sum_{j=1}^{N} a_j \left(\omega_j - \frac{U(\mathbf{Y}_j)}{RT}\right) \tag{33}$$

where $\mathbf{Y}_j$ denotes the jth simulated conformation, $Q(T)$ is expressed by eq 28, and the coefficients a_i ($i = 1, 2, ..., n$) are defined by eq 20. It should be noted that the a_i comprise the similarity of the simulated conformations to the experimental ones. It is also worth noting that we made use of the fact that, by eq 20, the coefficients a_i are normalized to 1.

In the initial test of the method reported in this work, the UNRES force field was calibrated with the experimental data for the tryptophan cage (PDB entry 1L2Y; Figure 3), a 20-residue α-helical protein for which an extensive NMR study was recently carried out[67] to determine the conformational ensembles at T = 280, 305, and 313 K. The thermodynamic folding-transition temperature of this protein is T_{f} = 313 K;[80]

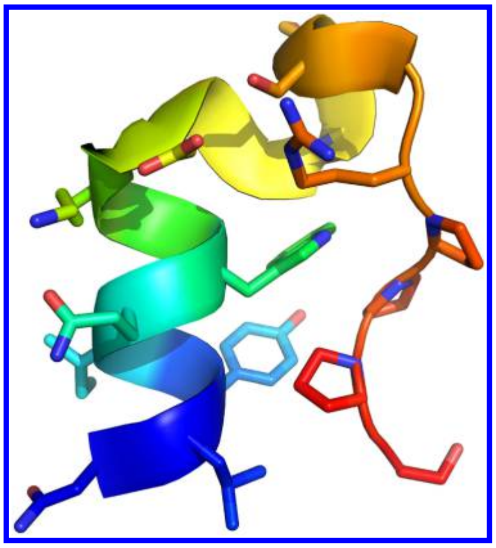


Figure 3. NMR structure of the tryptophan cage as deposited in the Protein Data Bank (entry 1L2Y). Side chains are shown in the stick representation, while the backbone is shown in the ribbon representation. The chain is colored from blue to red going from the N- to the C-terminus. The drawing was prepared using PyMOL.[88]

thus, these temperatures run up to the folding-transition temperature.

Because the ensemble of generated conformations depends on energy-function parameters, the calibration procedure has to be run in cycles consisting of ensemble generation and local minimization of l^*_N. A general iterative calibration scheme is illustrated in Figure 4. The Secant Unconstrained Minimization Solver (SUMSL) routine[81] was used for local minimization. To determine the behavior of the procedure and the dependence of the results on the data used and on the number of optimized parameters of the energy function, three calibration runs were carried out. In the first run, only the energy-term weights (the w's in eq 21) were optimized and only the ensemble of the folded protein at T = 280 K was used. In the second run, the energy-term weights were optimized but the ensembles at all three temperatures were used. Finally, in the third run, the ensembles at all three temperatures were used and all of the parameters in the torsional potentials (U_tor in eq 21) and the third-order correlation potentials ($U^{(3)}_\mathrm{corr}$ and $U^{(3)}_\mathrm{turn}$ in eq 21) were optimized. A calibration run was stopped when no significant difference between the heat-capacity and rmsd(T) curves was observed in two consecutive iterations.

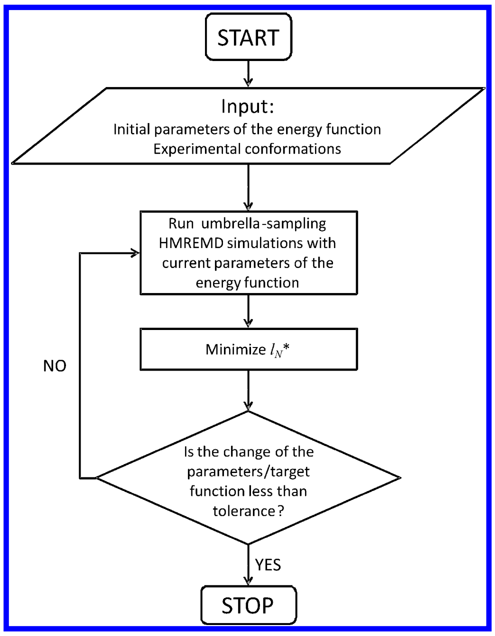

START
Input:
Initial parameters of the energy function
Experimental conformations
Run umbrella-sampling HMREMD simulations with current parameters of the energy function
Minimize l_N*
NO
Is the change of the parameters/target function less than tolerance?
YES
STOP

Figure 4. Block diagram of the maximum-likelihood force-field calibration method developed in this work.

For each of the three runs, the starting energy-term weights were generated from uniform distributions. The other parameters were as for the force field optimized in our earlier work;[39] it should be noted that for calibration run 3, the torsional-potential parameters and the correlation-potential parameters thus assigned were only initial values and were varied during the course of calibration.

In each iteration, production umbrella-sampling Hamiltonian replica-exchange simulations were carried out with restraints centered at q° = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7 with a force constant of 100 kcal/mol at T = 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 360, 380, 400, 420, 440, 460, 480, 500, 520 K (24 temperatures total), with two trajectories per temperature and restraint set (384 trajectories total). For calibration runs 2 and 3, three umbrella-sampling runs were carried out, each with the reference structure being the mean structure of the experimental ensemble at T = 280, 305, or 313 K. All three series of runs were processed together in a single WHAM calculation.

Each replica-exchange simulation was run in the Langevin mode with the water viscosity scaled down by a factor of 100, as in our previous work.[12] Each trajectory consisted of 60 000 000 steps with a step length of 4.89 fs. The A-MTS algorithm[75] was used to carry out numerical integrations. Replicas (in temperature or in temperature and restraints) were exchanged every 20 000 steps, and snapshots from the trajectories were collected at the same frequency. The results were processed with WHAM to determine the factors ω and to calculate the heat-capacity and rmsd(T) curves. For l^*_N minimization, the conformations of the last 1000 snapshots from each trajectory were considered.

Testing the Force Fields. To test the three force fields obtained as a result of optimization, we used the following 16 proteins identified by the respective PDB codes: 1BBL (α; 48 residues), 1BDD (α; 46 residues), 1BG8 (α; 85 residues), 1CLB (α; 75 residues), 1E0G ($\alpha + \beta$; 48 residues), 1E68 (α; 70 residues), 1ENH (α; 56 residues), 1FSD ($\alpha + \beta$; 28 residues), 1GAB (α; 53 residues), 1KOY (α; 62 residues), 1LQ7 (α; 67 residues), 1P68 (α; 102 residues), 1POU (α; 75 residues), 1PRU (α; 56 residues), 1VII (α; 36 residues), and 2HEP (α; 42 residues). The reason for using mostly α-helical proteins to test the force field was that in the initial test of the maximum-likelihood approach reported in this work, the force fields were calibrated with a small α-helical protein. An additional benchmark set consisting of 12 α-helical proteins that have never been used in parametrizing UNRES, namely, 1ACP (77 residues), 1BW6 (56 residues), 1EI0 (38 residues), 1FEX (59 residues), 1HYP (80 residues), 1LEA (84 residues), 1NKL (78 residues), 1RES (43 residues), 1RIJ (24 residues), 1YRF (35 residues), 2CRB (97 residues), and 1K40 (126 residues), was constructed to compare the performance of the best force field derived in this work with previous ones, including a recent force field that contains additional terms responsible for the coupling

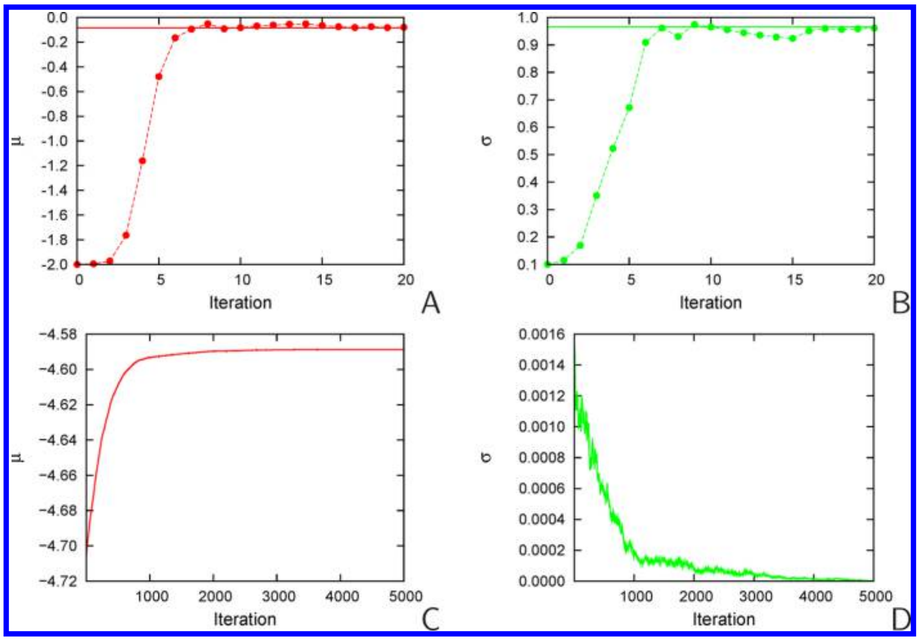


Figure 5. Variation of (A) the mean (μ) and (B) the standard deviation (σ) of the Gaussian distribution estimated from 100 "experimental" points using the maximum-likelihood-fitting method developed in this work with the iteration number. The "experimental" points were generated from a Gaussian distribution with zero mean and unit standard deviation. The horizontal lines correspond to the average and standard deviation over the 100 points (the unbiased estimators of μ and σ). The initial values were $\mu^{(0)} = -2.0$ and $\sigma^{(0)} = 0.1$, and the sampling size in the generation of the "theoretical" distribution was $M = 1000$. The overlap diameter in eq 18 was $s = 0.1$. (C, D) Same as (A) and (B), respectively, but the iteration was started with $\mu^{(0)} = -5.0$, which means that the points sampled from the initial distribution did not overlap with the "experimental" points. For clarity, the initial value of σ is not shown in (D). It can be seen that $\sigma \to 0$ and that the mean never approaches the true value even after 5000 iterations.

of the backbone-local and side-chain-local conformational states.

For each protein and force field, an unrestricted MREMD run was carried out at temperatures from $T = 200$ to 390 K in steps of 10 K (20 temperatures total) and four trajectories per temperature (80 trajectories total). Each trajectory consisted of 20 000 000 steps with a step length of 4.89 fs. Replicas were exchanged every 20 000 steps, and snapshots from the trajectories were collected at the same frequency. The conformations of the second half of each simulation (the last 500 snapshots of each trajectory) were processed with WHAM. For each protein, the heat-capacity curve was determined, and a cluster analysis was run using Ward's minimum-variance method[82] to divide the set of conformations into five clusters. The clusters were ranked according to the cumulative probabilities of the constituent conformations calculated at a temperature T_c about 20 K lower than that of the heat-capacity peak (eq 34), as in our previous work:[17]

$$P_I = \frac{\sum_{k \in \{I\}} \exp\left(\omega_k - \frac{U_k}{RT_c}\right)}{\sum_{k=1}^{N} \exp\left(\omega_k - \frac{U_k}{RT}\right)} \tag{34}$$

where P_I is the probability of cluster I and $\{I\}$ denotes the set of conformations that belong to cluster I. For each cluster, the representative conformation was defined as the conformation of the cluster with the lowest rmsd from the weighted-average conformation of the cluster, with the weights being calculated at T_c, as in our previous work.[17] All of the MREMD simulations and the postprocessing WHAM and cluster-analysis calculations were run with version 3.2 of the UNRES package (www.unres.pl).

RESULTS AND DISCUSSION

Tests with Gaussian Distributions. We initially checked the ability of the modified maximum-likelihood method to retrieve the parameters of Gaussian distributions. These simple tests also enabled us to assess the dependence of the behavior of the algorithm on the initial approximations, parameters, and dimensionality of the problem.

In the first test, we generated $n = 100$ points from the Gaussian distribution with $\mu = 0$ and $\sigma = 1$ and treated them as the experimental points. The starting parameters were $\mu^{(0)} = -2.0$ and $\sigma^{(0)} = 0.1$. The value of s was 0.1. Following eq 18, the values of $\mu^{(p)}$ and $\sigma^{(p)}$ that minimize l_{N}^{*} in the pth iteration are given by eqs 35 and 36, respectively:

$$\mu^{(p)} = \frac{\sum_{i=1}^{N} w_i^{(p-1)} y_i^{(p-1)}}{\sum_{i=1}^{N} w_i^{(p-1)}} \tag{35}$$

$$\sigma^{2(p)} = \frac{\sum_{i=1}^{N} w_i^{(p-1)} (y_i^{(p-1)})^2}{\sum_{i=1}^{N} w_i^{(p-1)}} - (\mu^{(p)})^2 \tag{36}$$

where

$$w_i^{(p-1)} = \frac{\sum_{k=1}^{n} \exp\left[-\frac{(x_k - y_i^{(p-1)})^2}{2s^2}\right]}{\sum_{k=1}^{N} \exp\left[-\frac{(y_k^{(p-1)} - y_i^{(p-1)})^2}{2s^2}\right]} \tag{37}$$

and $y_i^{(p-1)}$ denotes the ith point generated with the mean and standard deviation of the $(p - 1)$th iteration. The number of points generated per iteration was $N = 1000$.

Figure 5A,B shows the variation of $\mu^{(p)}$ and $\sigma^{(p)}$, respectively, with the iteration number. It should be noted that even though

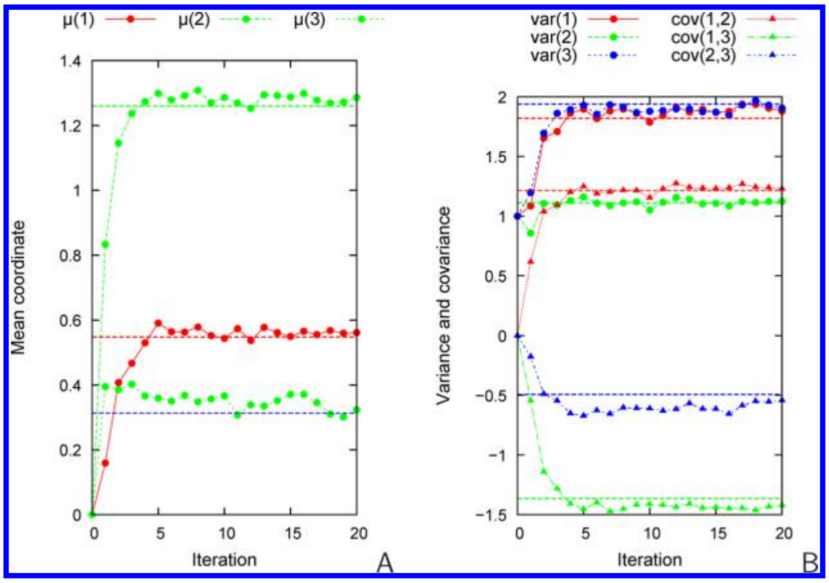


Figure 6. Variation of (A) the components of the mean and (B) the six unique elements of the variance−covariance matrix for the run of the determination of the parameters of a three-dimensional Gaussian distribution (a total of nine independent parameters) from a set of 100 points generated from the distribution given by eq 38. The horizontal lines correspond to the components of (A) the mean and (B) the elements of the variance−covariance matrix.

the initial values of μ and σ are far from the final values, there is some overlap between the initial and final distributions. It can be seen that the convergence is quick and the final values are, within the error margin, the same as the mean and the variance calculated over the points. However, if the overlap between the initial and the target distribution is poor, the iterative procedure can lead to false values of the parameters. This is illustrated in Figure 5C,D, where the initial values $\mu^{(0)} = -5$ and $\sigma^{(0)} = 0.1$ were assumed, reducing the overlap of the initial distribution with the "experimental" distribution. As a result, μ stays close to the initial value while σ tends to 0. This observation is an argument for using umbrella sampling to generate the conformational ensembles for the maximum-likelihood calibration of potential-energy functions (cf. Maximum-Likelihood Calibration of Macromolecular Energy Functions). On the other hand, the problem can be solved for the case reported in Figure 5C,D by increasing the value of the parameter s to 1.

To test the behavior of the method when the dimensions of the variable and the parameter space were increased, we considered the three-dimensional Gaussian distribution given by eq 38:

$$\rho(\mathbf{x}; \boldsymbol{\mu}; \mathbf{C}) = \sqrt{\frac{1}{(2\pi)^3 \det \mathbf{C}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{38}$$

with

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 1.2 \\ 0.3 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} 1.9434 & 1.4397 & -1.3157 \\ 1.4397 & 1.3185 & -0.5988 \\ -1.3157 & -0.5988 & 1.7780 \end{pmatrix}$$

There are nine parameters to be determined (three mean values and six unique elements of the variance−covariance matrix). The formulas for the mean values and the elements of the variance−covariance matrix of the next iteration are expressed by eqs 39 and 40, respectively:

$$\mu_j^{(p)} = \frac{\sum_{i=1}^{N} w_i^{(p-1)} y_{ij}^{(p-1)}}{\sum_{i=1}^{N} w_i^{(p-1)}} \tag{39}$$

$$\mathrm{cov}_{jk} = \frac{\sum_{i=1}^{N} w_i^{(p-1)} y_{ij}^{(p-1)} y_{ik}^{(p-1)}}{\sum_{i=1}^{N} w_i^{(p-1)}} - \mu_j^{(p)} \mu_k^{(p)} \tag{40}$$

where $j = 1, 2, 3$; $k = 1, ..., j$; and

$$w_i^{(p-1)} = \frac{\sum_{k=1}^{n} \exp\left[-\frac{\sum_{j=1}^{3}(x_{kj} - y_{ij}^{(p-1)})^2}{2s^2}\right]}{\sum_{k=1}^{N} \exp\left[-\frac{\sum_{j=1}^{3}(y_{kj}^{(p-1)} - y_{ij}^{(p-1)})^2}{2s^2}\right]} \tag{41}$$

The variation of the parameters during the course of the iteration is shown in Figure 6. It can be seen that, as for one-dimensional Gaussian distribution, the convergence is quick and that the final values are very close to the values determined directly from the "experimental" points.

The distributions of the means calculated from samples of n experimental points drawn from the same population distributed according to the Gaussian law with mean μ and variance σ^2 obey the Gaussian distribution with mean μ and variance σ^2/n. The distributions of the sums of the squares of the deviations from the experimental points from the mean scaled by $1/\sigma$ obey the χ^2 distribution for $N - 2$ degrees of freedom; this distribution can easily be transformed into the distribution of the standard deviations from the mean. To determine whether the means and standard deviations calculated with the proposed approach obey these distributions, we carried out 1000 independent optimization runs, each with a different set of 100 "experimental" points drawn from the normal distribution with $\mu = 0$ and $\sigma = 1$. The distributions of the resulting μ and σ values obtained assuming that the parameter of the method was $s = 0.1$ are shown in Figure 7A,B, respectively. As shown, the distributions conform with the

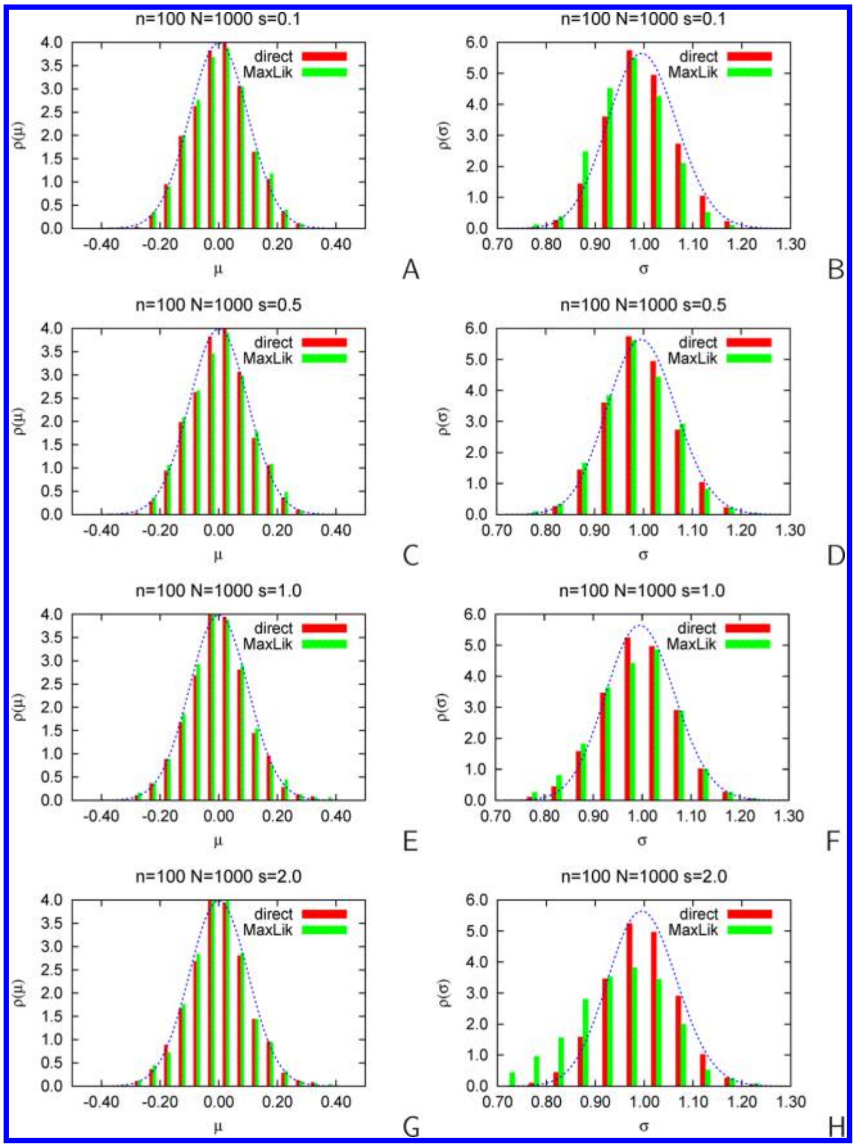


Figure 7. Distributions of (A, C, E, G) μ and (B, D, F, H) σ determined from 1000 sets of 100 points randomly generated from a Gaussian distribution with zero mean and unit standard deviation by the maximum-likelihood method developed in this work with different values of the parameter s of eq 18 (green bars) compared to the distributions of μ and σ calculated directly from the "experimental" points (red bars). The dashed lines represent the respective analytical distributions. (A, B) s = 0.1; (C, D) s = 0.5; (E, F) s = 1.0; (G, H) s = 2.0.

appropriate normal and χ^2 distributions, respectively, for the original Gaussian distribution and 98 degrees of freedom.

Finally, we analyzed the dependence of the results on the parameter s. Again, we carried out 1000 independent optimization runs with s = 0.5, 1.0, and 2.0. As shown in Figure 7C−F, the distributions of parameters do not depend on s even for s = 1.0, which is the value of the standard deviation of the distribution of points in the population. Only when s = 2.0 does the distribution of σ become visibly broader than the true distribution or the distribution of standard deviations calculated directly from the samples (Figure 7H); however, the distribution of the mean remains similar to those calculated with smaller values of s. This result suggests that that the method is robust. Given the corollary from the analysis of the results shown in Figure 5, it can be concluded that it is better to take larger values of s because this removes the danger of poor overlap of the initial distribution with the experimental points, thus preventing the method from diverting to false values of the parameters (Figure 7C,D).

Force-Field Calibration with Tryptophan Cage Experimental Data. *Single Temperature Reference.* In the first test of force-field optimization with the maximum-likelihood approach developed in this work, we used the experimental ensemble of the folded tryptophan cage miniprotein determined at T = 280 K.[67] The plots of l_N^* versus iteration number before and after local minimization are shown in Figure 8A. It can be seen that the minimized value of l_N^* in a given iteration does not differ remarkably from the initial value at the next iteration. This is a very strong argument for the robustness of the procedure because with complete coverage of the conformational space (not achievable in real simulations) these values should be identical. Thus, the use of l_N^* as the target function results in a much more stable iterative procedure than use of the energy gap, the Z-score,[33,34] the free-energy gaps between subsets of conformations with different native-likeness,[17,35,36] or the ensemble-averaged measures of native-likeness such as rmsd.[17] The reason for this is that because of the exponential dependence of the Boltzmann weights on

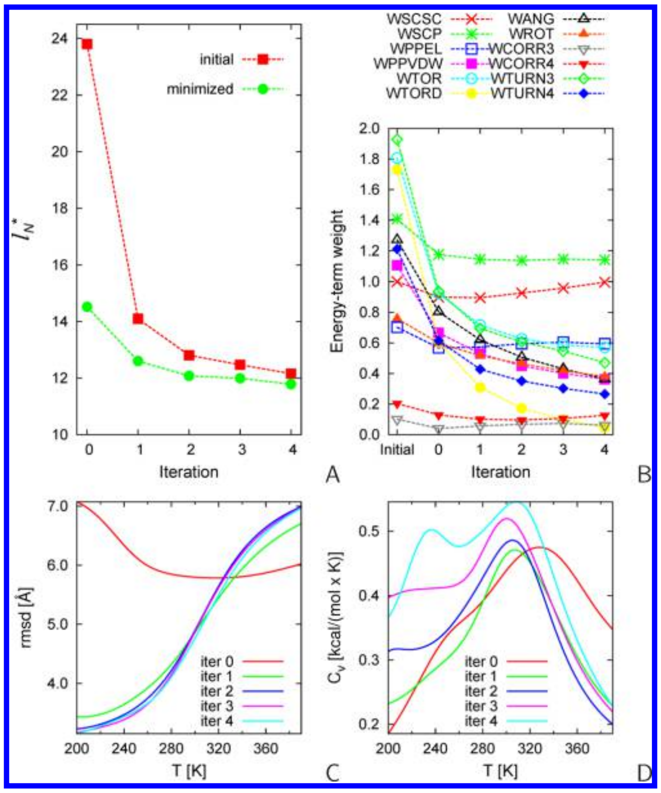


Figure 8. (A) Variation of the target function with iteration number for the calibration run of 1L2Y using the experimental conformational ensemble determined at T = 280 K. (B) Variation of the energy-term weights with iteration number. (C) Plots of the ensemble-averaged C$^\alpha$ rmsd from the 1L2Y structure vs temperature in the consecutive iterations. (D) Plots of heat capacity vs temperature in the consecutive iterations.

energy, the weights of the conformations that have, e.g., a very low rmsd from the native structure can easily be exaggerated, and a low value of the target function based on such quantities minimized with a given set of decoys can be obtained. Strict minimization of the energy gap assumes consideration of the energy of only the lowest-energy native and lowest-energy non-native conformations, thus ignoring the rest of the ensemble. As a consequence of overemphasizing selected conformations of the ensemble, the predicted averages are highly distorted and poorly correspond to those obtained in the simulation with the energy-function parameters obtained in the next iteration,[17] unless one is working with a complete set of conformations, which is possible only for simple lattice models and short chains. Conversely, l_{N}^{*} in eq 18 is a weighted sum of the logarithms of the probabilities of all simulated conformations of the ensemble, and therefore, no conformation is ignored.

From Figure 8A, it can also be seen that the convergence of the procedure is fast and stable, in contrast to the oscillatory behavior of the convergence of the previous approaches that overemphasize "good" conformations (see, e.g., ref 35). The major change is observed from iteration 0 (started from randomly generated energy-term weights) to iteration 1; some progress is observed from iteration 1 to iteration 2, and then the progress is only incremental. This observation is confirmed by the plots of the variation of the energy-term weights during the course of the iteration (Figure 8B) and the plots of ensemble-averaged rmsd versus temperature (Figure 8C). The final values of the energy-term weights are summarized in the first column of Table S1 in the Supporting Information.

The experimental (NMR) conformational ensemble of the tryptophan cage is compared with that calculated with the optimized force field in Figure 9. As shown, the calculated ensemble is more diffuse than the NMR ensemble; however, the features of the fold are preserved. The mean rmsd at the experimental temperature (T = 280 K) is 4 Å, consistent with the fact that the present UNRES is a medium-resolution force field.[49] As can also be seen from the right panel of Figure 9, the calculated mean structure superposes well on the experimental mean structure; the rmsd is 1.7 Å. Therefore, the dispersion of the conformations of the calculated set and not the deviation of the average calculated structure from the average experimental structure makes the dominant contribution to the 4 Å ensemble-averaged rmsd. Attempts to optimize more parameters with only a single protein and an experimental ensemble at a single temperature would not be reasonable.

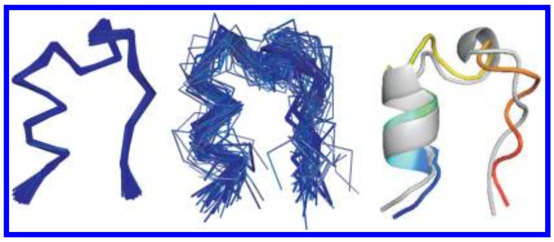


Figure 9. Comparison of the C$^\alpha$ trace of the experimental ensemble of 1L2Y at T = 280 K (left) with that obtained using the force field optimized using the experimental ensemble (center). The stick width and color saturation (from deep to light) correspond to decreasing probability of a conformation. (right) Superposition of the mean calculated structure (gray) on the mean experimental structure (colored from blue to red going from the N- to the C-terminus); the rmsd is 1.7 Å. The calculated all-atom mean structure used to draw the cartoon plot was obtained by applying the PULCHRA[89] and SCWRL[90] software to the coarse-grained mean structure. The stick drawings were prepared using MOLMOL,[91] and the cartoon drawings were prepared using PyMOL.[88]

Although thermodynamic data were not considered, the major heat-capacity peak occurs at T = 305 K (Figure 8D), a value quite close to that of the experimental heat-capacity peak, which occurs at T = 313 K.[80] The secondary peak occurs at a low temperature, and its presence suggests that single-reference calibration without thermodynamic data is not sufficient to obtain a force field with good folding properties.

Multiple Temperature References. In the next test of the optimization procedure, we used the NMR-determined conformational ensembles of the tryptophan cage at T = 280, 305, and 313 K. The plots of l_{N}^{*} and its components versus iteration number at each of these three temperatures are shown in Figure 10A. The variation of the energy-term weights with the progress of the iteration is shown in Figure 10B. The plots of the initial and final rmsd's from the reference structures corresponding to all three temperatures are shown in Figure 10C, and plots of the initial and final heat capacities are shown in Figure 10D. The final values of the energy-term weights are summarized in the second column of Table S1 in the Supporting Information. The calculated and experimental ensembles are compared in Figure 11. It can be seen that rapid progress occurs from iteration 0 to iteration 1 and that the subsequent progress is only incremental. It can also be seen that the final l_{N}^{*} value at T = 280 K (corresponding to the

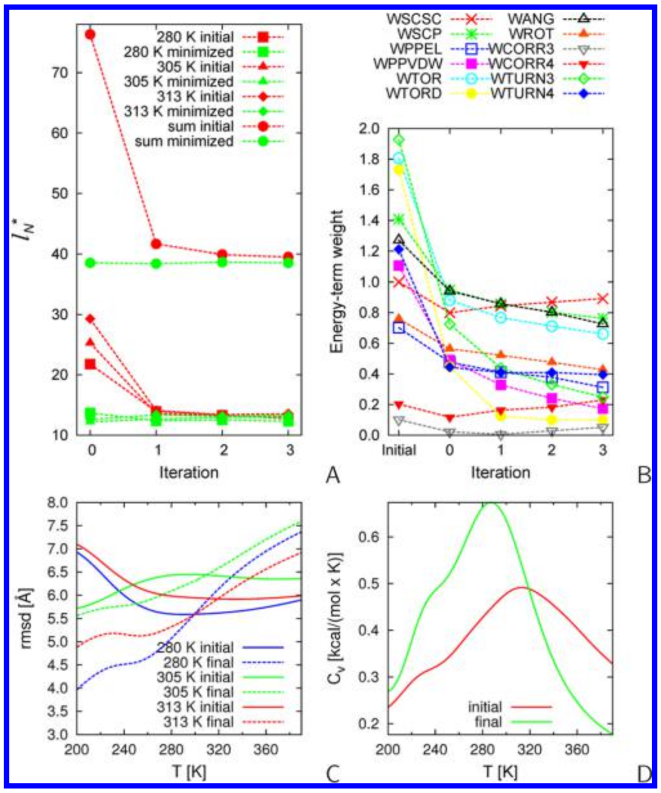


Figure 10. (A) Variation of the target function and its components corresponding to different temperatures with iteration number for the calibration run of 1L2Y using the experimental conformational ensembles determined at T = 280, 305, and 313 K. (B) Variation of energy-term weights with iteration number. (C) Plots of the ensemble-averaged C^{α} rmsd's from the 1L2Y structure vs temperature obtained with the initial (solid lines) and final (dashed lines) energy-term weights for the three temperatures considered. (D) Plots of heat capacity vs temperature obtained with the initial (solid lines) and final (dashed lines) energy-term weights for the three temperatures considered.

folded structure) is higher than that for single-reference calibration (Figure 8A) and that the rmsd's of the calculated structures from the experimental folded structure (determined at T = 280 K) calculated at T = 280 K (Figure 10C) are higher than that for the single-reference calibration (Figure 8A,C), which results from the inclusion of the experimental ensembles determined at higher temperatures in the optimization. Therefore, the result is a compromise of fitting to the ensembles at the three temperatures considered simultaneously. Consequently, the average rmsd from the mean experimental structure at T = 280 K (Figure 10C) is also higher than that for the single-reference calculations (Figure 8C), the calculated ensemble at T = 280 K is more diffuse than that obtained with the force field discussed in Single Temperature Reference (Figure 9, middle panel), and the respective mean structure has a higher rmsd from the mean experimental structure (Figure 9, right panel).

It can also be noted that the mean rmsd values for T = 305 K and T = 313 K are higher than those for T = 280 K (Figure 10C), the calculated conformational ensembles at those two temperatures are more diffuse (Figure 11B,C, middle panels), and the calculated mean structures superpose on the experimental mean structures with higher rmsd values (Figure 11E,F, left panels). The experimental ensembles at T = 305 K and T = 313 K are also more diffuse than that at T = 280 K. The poorest fit was obtained for T = 305 K, at which the extended C-terminal section (extended in the low-temperature experimental structure) is twisted with respect to the N-terminal section (α-helical in the low-temperature experimental structure), unlike the experimental structures at T = 280 and 313 K (Figure 11). Nevertheless, it can be seen from the left panel of Figure 11E that the twist of the chain is reproduced in the calculated mean structure.

On the other hand, it can be observed that, as opposed to the results obtained by fitting to conformational ensembles at a single temperature (Figure 8C), the heat-capacity profile has a single peak (Figure 10C). This observation suggests that using multiple references helps in reproducing the correct folding-transition thermodynamics. However, the heat-capacity peak occurs at a lower temperature and is broader and higher than the experimental one,[80] which suggests that explicit use of thermodynamic data in force-field optimization is necessary.

Optimization of the Torsional and Correlation Potentials. To find out whether optimization of other parameters can lead to additional progress, we included the parameters of the torsional potentials (12 parameters per residue-type pair, except for the Gly-Gly pair, where there are six parameters because of symmetry; it should be kept in mind that only three residue types—Gly, Ala, and Pro, with Ala representing every residue except for Gly and Pro—are considered with regard to local interactions). The initial energy-term weights were taken from iteration 2 of the run described above in Multiple Temperature References. The plots of l_N^* and its components versus iteration number at each of the three temperatures (T = 280, 305, and 313 K) are shown in Figure 12A, and the variation of the energy-term weights with the progress of iteration is shown in Figure 12B. The plots of the initial and final rmsd's from the reference structures corresponding to all three temperatures are shown in Figure 12C, and the plots of the initial and final heat capacities are shown in Figure 12D. The final values of the energy-term weights and torsional- and correlation-potential parameters are summarized in Tables S1−S3 in the Supporting Information, respectively. As could be expected, the final l_N^* values and the rmsd's are lower than those corresponding to calibration with only energy-term weights as parameters. The conformational ensembles calculated with the optimized parameters are also less diffuse than those calculated with the force field in Multiple Temperature References, in which only the energy-term weights were optimized (Figure 11A−C, right panels). The calculated mean structures superpose on the respective experimental mean structures with lower rmsd values (Figure 11, right panels) and have chain shapes closer to those of the experimental structures, especially at T = 305 and 313 K (Figure 11D−F, right panels). In particular, the conformations of the ensemble calculated at T = 280 K have the fine features of the chain trace in the top loop very close to those of the experimental ensemble and the chain twist for the calculated structures of the ensemble at T = 305 K is reproduced. On the other hand, the heat-capacity peak is too high (Figure 10D) compared with the experimental peak.[80]

Tests of the Derived Force Fields. The minimum rmsd's obtained during the simulations, the rmsd's, and the GDT-TS[83] and TMscore[84] values for the most native-like clusters of all 16 proteins tested for the three force fields described in Force-Field Calibration with Tryptophan Cage Experimental Data are shown in Figure 13. The numerical values of these quantities as well as the ranks and probabilities of the most native-like clusters are shown in Tables S4−S6 in the Supporting

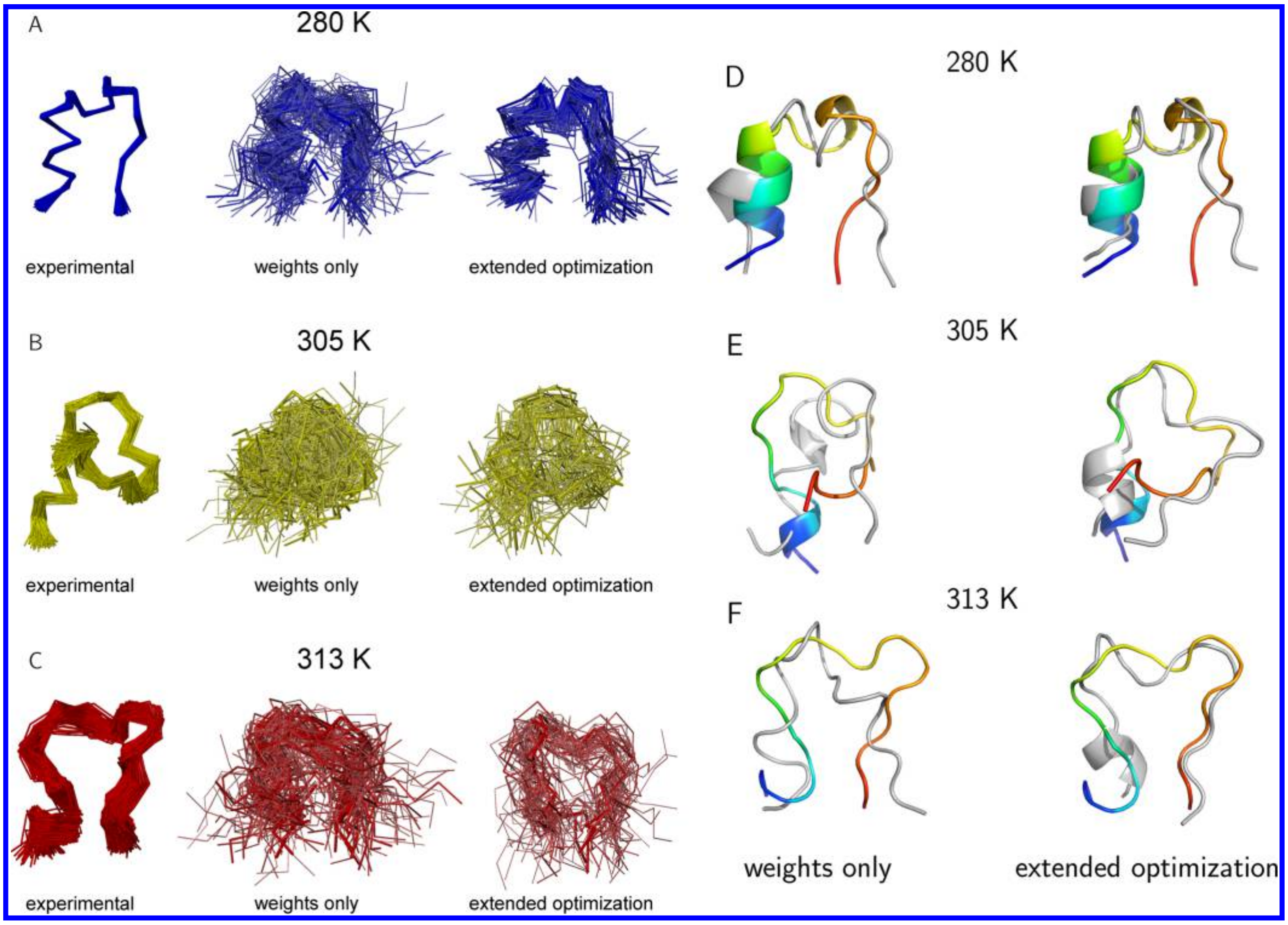


Figure 11. (A−C) Comparison of C$^{\alpha}$ traces of the experimental ensembles of the tryphtophan cage[67] at the three temperatures at which the measurements were performed (left) with the ensembles calculated using the force field obtained by optimizing the energy-term weights in eq 21 (center) and the force field for which the torsional and correlation potentials were additionally optimized (right). The ensembles are colored according to temperature: blue, T = 280 K (A); yellow, T = 305 K (B); red, T = 313 K (C). For the calculated ensembles, the stick thickness and color saturation (from deep to light) correspond to the decreasing probability of the conformations. (D−F) Superpositions of the calculated mean structures (gray cartoons) on the experimental mean structures of the tryptophan cage (cartoons colored from blue to red going from the N- to the C-terminus) at (D) 280 K (rmsd = 3.6 and 2.2 Å, respectively), (E) 305 K (rmsd = 6.0 and 4.1 Å, respectively), and (F) 315 K (rmsd = 5.7 and 3.2 Å, respectively). The calculated all-atom mean structures used to draw the cartoon plots were obtained by applying the PULCHRA[89] and SCWRL[90] software to the coarse-grained mean structures. The stick drawings were prepared using MOLMOL[91] and the cartoon drawings were prepared using PyMOL.[88]

Information. It can be seen that native-like conformations were found by MREMD for all of the proteins and all of the force fields. However, the clustering procedure did not select the native-like conformations for 1E0G (α + β) and 1BG8 (α-helical). For 1KOY and 1PRU, the rmsd values are comparatively high; however, the structures of the most native-like clusters still have native-like topologies, especially for 1PRU, which has a largely unfolded N-terminal part that contributes the most to the high rmsd. Compared with the results obtained in our earlier work with hierarchical optimization,[17] the force field is significantly more transferable, even though it was calibrated only with a small α-helical protein.

As shown in Figure 13, the results obtained with the force field calibrated with the reference structures at three temperatures are slightly better than those obtained with the force field calibrated with a single reference temperature, but the difference is not substantial. It can also be seen that even though the force field obtained by optimizing the torsional and correlation-term parameters in addition to the energy-term weights produced a better fit of the calculated ensembles of the tryptophan cage to the experimental ensembles (cf. Optimization of the Torsional and Correlation Potentials), it resulted in rmsd values that were higher by about 1 Å and TMscore and GDT-TS values that were lower by about 0.08 on average compared with those obtained using the force field for which only the energy-term weights were optimized. Some improvement was achieved only for 1ENH, 1KOY, and 1PRU. This observation suggests that the force field was overfitted to a single protein at the expense of transferability. Consequently, more proteins must be used when more parameters are optimized, which in turn is necessary to obtain a force field with better transferability and better resolution.

Representative conformations of the most native-like clusters of the test proteins corresponding to the force field of calibration run 2 (reference structures from three temperatures, only energy-term weights optimized) superposed on the experimental structures are shown in Figure 14. Three

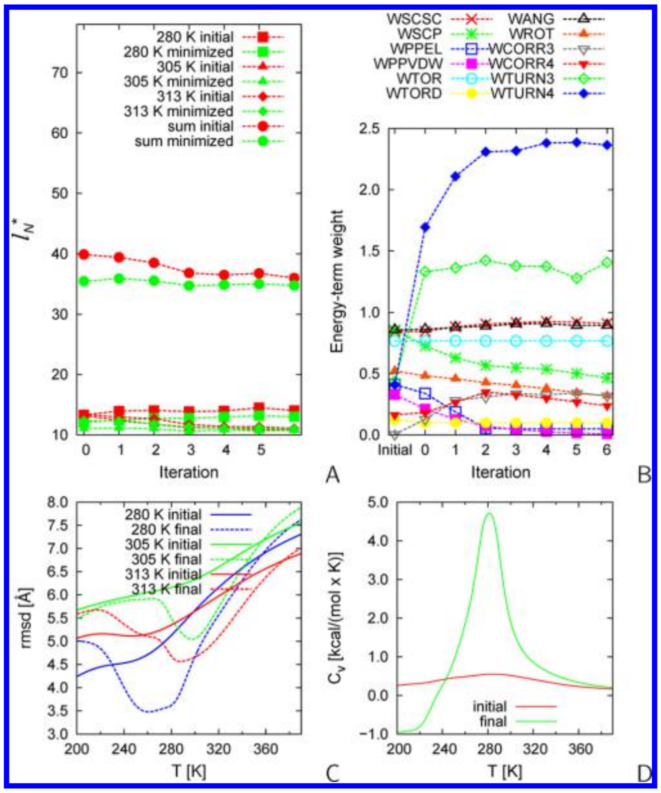


Figure 12. Same types of plots as in Figure 10 but for optimization of the energy-term weights, torsional parameters, and the parameters of the correlation terms.

examples of proteins whose structures were well-predicted with the force field, namely, 1ENH, 1E68, and 1P68, are shown in Figure 14A−C, respectively. 1ENH has a distorted three-α-helix bundle topology that was reproduced in the calculated structure, as shown in Figure 14A, even though the rmsd from the experimental structure is about 6 Å. The five-helix-bundle topology of 1E68 was reproduced quite accurately except for a slightly different angle of the perpendicular α-helix preceding the C-terminal helix (Figure 14B). The four-α-helix-bundle protein 1P68 was the largest protein considered in this series of tests. As can be seen from Figure 14C, the predicted structure overlaps very well with the experimental structure.

As an example of predictions for which relatively high rmsd values were obtained, the calculated and experimental structures of 1KOY are superposed in Figure 14D. It can be seen that the calculated structure traces the experimental structure; however, the angles between the long C-terminal α-helix and the N-terminal section are different in the calculated and experimental structures, which results in the relatively high rmsd. A similar situation occurs for 1POU. Only for 1BG8 did none of the five calculated models have a topology similar to that of the experimental structure.

The calculated and experimental structures of the two $\alpha + \beta$ proteins studied in this work, 1FSD and 1E0G, are superposed in Figure 14E,F, respectively. For 1FSD, it can be seen that a model with a distorted β-hairpin was obtained (Figure 14D). For 1E0G, the central α-helical section superposes well on the respective section of the experimental structure, but the N- and the C-terminal strands are α-helical in the calculated structure.

For the variant-2 force field, we did additional tests with 12 α-helical proteins of the set used in our latest work[45] to evaluate a version of the UNRES force field that includes new torsional terms involving side-chain centers; these terms correspond to coupling between the backbone-local and side-chain-local conformational states. These proteins have never been used in parametrization of any variant of UNRES. The bar diagrams of the rmsd's for the best models compared with the rmsd's for the best models computed with the version of the UNRES force field of ref 39 and that with the new terms[45] are shown in

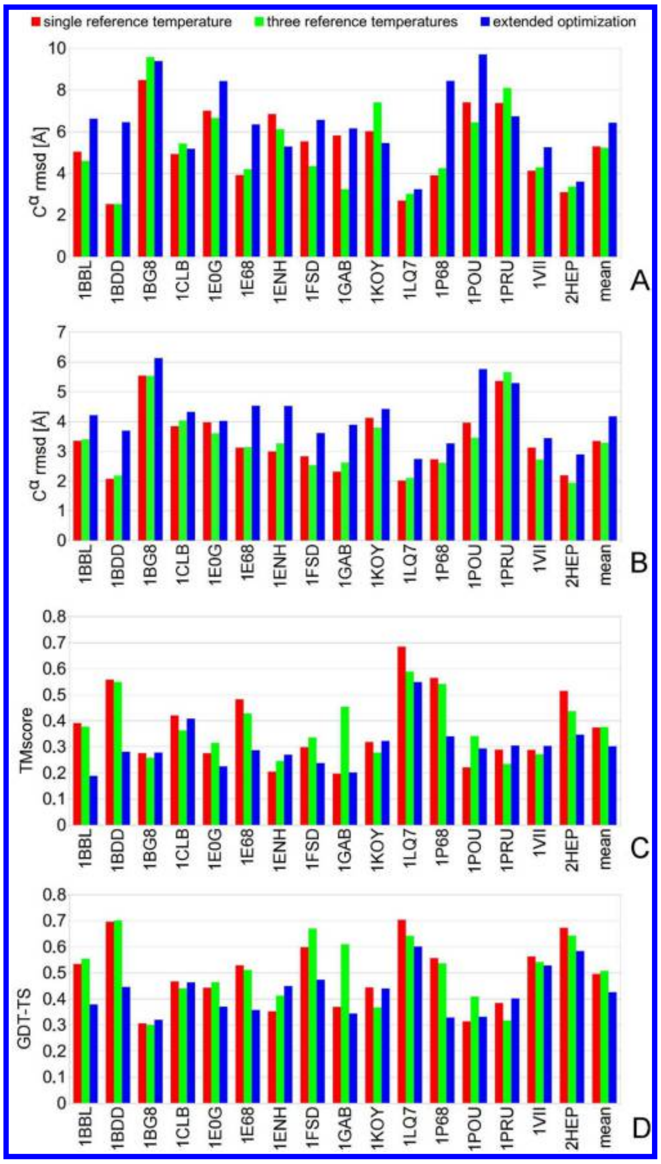


Figure 13. Bar diagrams of the four quantities that characterize the agreement between the calculated and experimental structures obtained with the three variants of the UNRES force field generated in this work. Red bars: variant 1 (only the energy-term weights were optimized, and only the experimental ensemble of the tryptophan cage determined at T = 280 K was used). Green bars: variant 2 (only the energy-term weights were optimized, and the experimental ensembles of the tryptophan cage determined at T = 280, 305, and 313 K were used). Blue bars: variant 3 (the energy-term weights, parameters of the torsional potentials, and parameters of the correlation terms were optimized, and the experimental ensembles of the tryptophan cage determined at T = 280, 305, and 313 K were used). (A) C^α rmsd of the mean structure of the most native-like cluster from the experimental structure. (B) Minimum C^α rmsd from the experimental structure over the entire MREMD run. (C) TMscore.[84] (D) GDT-TS.[83]

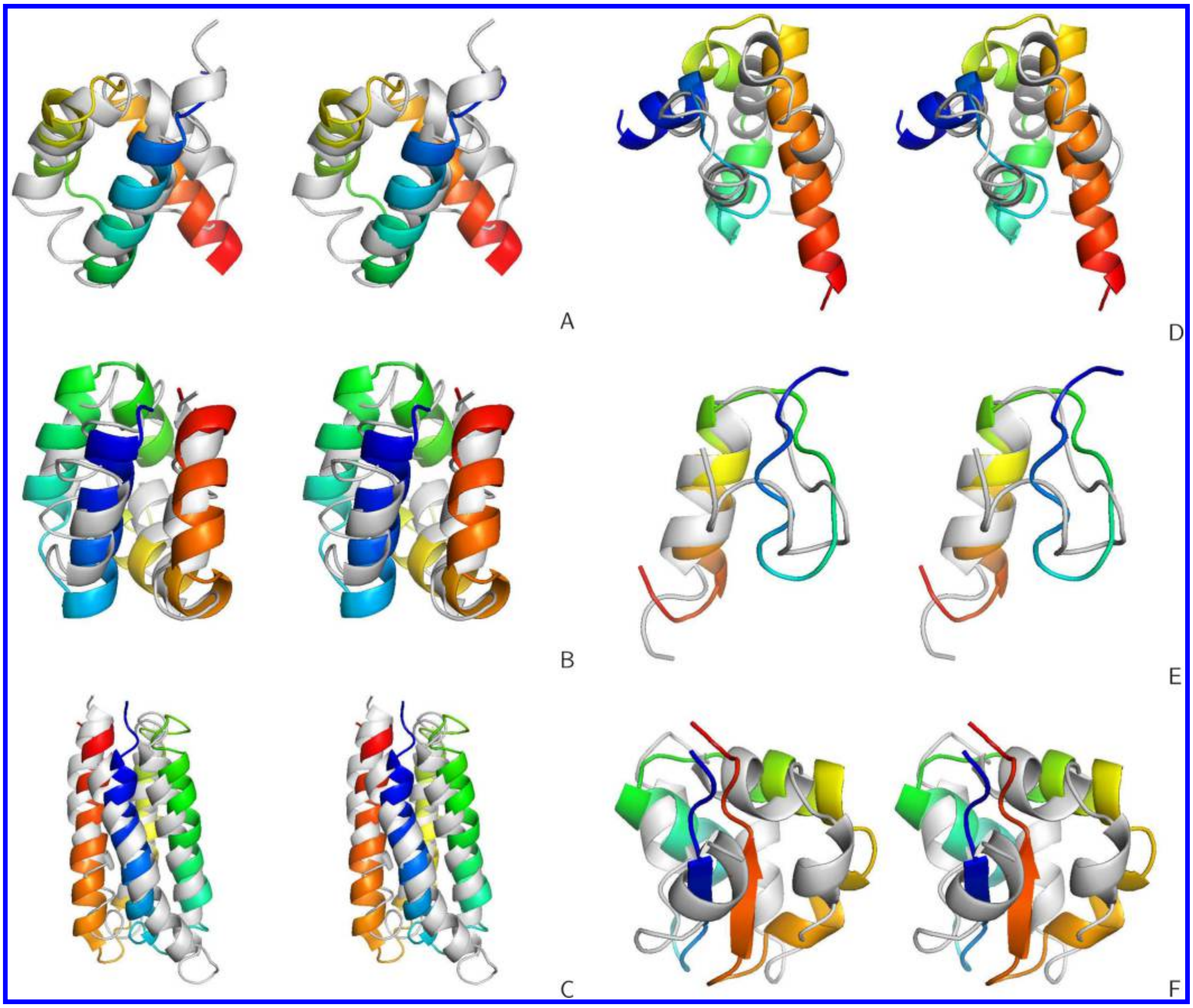


Figure 14. Cartoon diagrams of mean structures from the most native-like clusters of conformations of selected proteins studied in this work calculated with variant 2 of the UNRES force field calibrated in this work (light-gray cartoons) superposed on those of the experimental structures (identified by PDB code; chains colored blue to red going from the N- to the C-terminus): (A) 1ENH, (B) 1E68, (C) 1P68, (D) 1KOY, (E) 1FSD, (F) 1E0G. The drawings were prepared using PyMOL.[88]

Figure 15. As shown, the force field derived in this work gives lower rmsd values as well as higher TMscore and GDT-TS values on average than those of the two other force fields.

It can be noted that although the force field obtained by maximum-likelihood optimization with the use of the conformational ensembles of the tryptophan cage determined at three temperatures performs well on α-helical proteins, the calculated structures are of medium quality. This feature might result from the use of only a single small α-helical protein for calibration, as a result of which the force field was not trained sufficiently to recognize the details of the folds. On the other hand, because the quality of the calculated structures is comparable to that of the calculated structures of the training protein, it might also result from the use of rmsd as a criterion of similarity, which emphasizes the overall fold and is less sensitive on the details of the local structure.

■ CONCLUSIONS

The new variant of the maximum-likelihood method developed in this work appears to be a robust and stable approach to force-field calibration. As in our previous approaches,[17,31,33,34] it involves iterations consisting of simulations with current energy-function parameters to generate decoys and minimization of the target function with the obtained set of decoys. Its major advantage over the existing approaches is that neither the reference nor calculated structures must be classified according to the degree of native-likeness, which always results in some arbitrariness. This feature also enables us to use multiple reference conformations, which is especially important for proteins near or at the folding-transition temperature. The calculated structures are matched with the experimental structures in a flexible manner by means of the Gaussian-overlap function introduced as an approximation to the Dirac δ function (eq 7). The only remaining arbitrariness is in the choice of the standard deviation of the Gaussian (s); however,

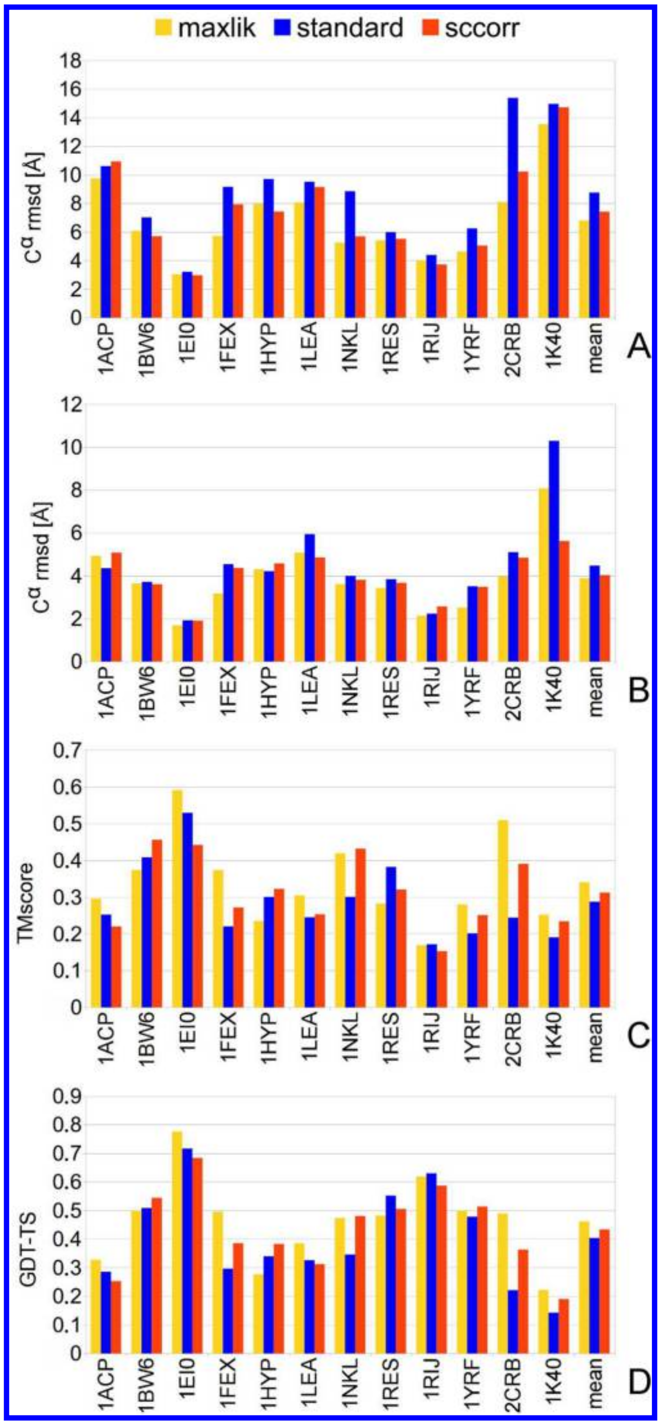


Figure 15. Bar diagrams of the four quantities that characterize the agreement between the calculated and experimental structures corresponding to the results of simulations for the 12 additional α-helical benchmark proteins with variant 2 of the force field obtained in this work (yellow bars and the "maxlik" label) compared with the results obtained for the standard UNRES force field derived in ref 39 (blue bars and the "standard" label) and the latest variant of the force field enhanced with additional torsional terms that involve the side-chain centers[45] (red bars and the "sccor" label). (A) C^{α} rmsd of the mean structure of the most native-like cluster from the experimental structure. (B) Minimum C^{α} rmsd from the experimental structure over the entire MREMD run. (C) TMscore.[84] (D) GDT-TS.[83]

the test runs described in Tests with Gaussian Distributions demonstrated that the method is not very sensitive to this parameter. Another advantage of the approach is that in contrast to those designed previously,[17,22,26,33,34,85−87] the target function does not overemphasize selected conformations either explicitly (as in the energy-gap[22,33,34,87] and Z-score optimization approaches[26,85,86]) or implicitly (e.g., as in hierarchical optimization by overemphasizing "good" conformations through large Boltzmann weights). A consequence of this fact is that the iterative procedure exhibits steady convergence (Figures 8 and 12). The proposed approach is related to previously developed methodologies of force-field calibration: when a hypercylinder representation of the Dirac δ function (eq 11) is used instead of the Gaussian representation and there is only a single reference structure (e.g., the crystal structure), its application leads to maximization of the energy gap between the lowest-energy native-like and lowest-energy non-native structures in the low-temperature limit.

The results of the evaluations of the force fields obtained in this work strongly suggest that the new approach leads to well-transferable force fields within the structural class of the training protein(s) used. Even though the force fields were calibrated only with a small α-helical protein, clusters of structures with native-like topology were located for 13 out of 14 α-helical proteins of the first test set, with sizes ranging from 36 to 102 residues (as shown in Tests of the Derived Force Fields). The force field also located a cluster of native-like structures of 1FSD (a minimum $\alpha + \beta$ fold; Figure 14E). However, it should be noted that because the force field was trained on an α-helical protein, it can be used to study only α-helical proteins and not $\alpha + \beta$ or β-proteins. Once experimental data for $\alpha + \beta$ and β-proteins are available, the method will be used to produce a force field applicable to all structural classes of proteins. On the other hand, the calculated structures are of medium resolution, which reflects the medium resolution of the calculated structures of the training protein (the tryptophan cage). Better resolution can most likely be achieved by the introduction of more training proteins, the use of a measure of similarity that emphasizes the local details of the structure better than the rmsd, and possibly modification or inclusion of new terms in the force field. Another important point is the inclusion of thermodynamic data, presumably the heat-capacity curves, in the target function to reproduce the thermodynamics of folding. This work is now underway in our laboratory.

As mentioned in the Introduction, previous attempts to optimize UNRES were successful only when training proteins with simple topologies were used. The method proposed in this work has good numerical behavior (steady convergence). Moreover, it uses partially unfolded experimental structures measured at higher temperatures, and thus, force-field training includes thermal-unfolding pathways, which should prevent us from obtaining nonergodic force fields in which the native-like structures have the lowest energies but there is virtually no pathway to the native-like region from the unfolded-structure region. Therefore, the new method should be able to handle training proteins with complicated topologies (subject to the availability of the corresponding experimental data) and thus to produce force fields with enhanced transferability.

Finally, it should be noted that the variant of the maximum-likelihood approach introduced in this work can be applied to solve any maximum-likelihood-fitting problem in which the probability-density function cannot be evaluated at the experimental points but can be obtained only by simulations and the dimension of the configurational space is too high for the construction of histograms of the distributions of the experimental points.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00395.

> Optimized energy-term weights for the three variants of the UNRES force field obtained in this work by maximum-likelihood optimization (Table S1); optimized torsional-potential coefficients for the third variant of the UNRES force field obtained in this work by maximum-likelihood optimization (Table S2); parameters of the correlation terms of the UNRES force field obtained by maximum-likelihood optimization with tryptophan-cage experimental data (Table S3); results of tests of the UNRES force field obtained by maximum-likelihood optimization using the experimental conformational ensemble of the tryptophan cage at T = 280 K (Table S4); results of tests of the UNRES force field obtained by maximum-likelihood optimization using the experimental conformational ensembles of the tryptophan cage at T = 280, 305, and 313 K with only energy-term weights optimized (Table S5); and results of tests of the UNRES force field obtained by maximum-likelihood optimization using the experimental conformational ensembles of the tryptophan cage at T = 280, 305, and 313 K with energy-term weights, torsional coefficients, and correlation-term parameters optimized (Table S6) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +48 58 523 5124. Fax: +48 58 523 5012. E-mail: adam@sun1.chem.univ.gda.pl.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by Grant DEC-2012/06/A/ST4/00376 from the Polish National Science Centre and Grant Mistrz 7./2013 from the Foundation for Polish Science. Computational resources were provided by (a) the supercomputer resources at the Informatics Center of the Academic Computer Center in Gdańsk TASK (CI TASK) (Gdańsk, Poland); (b) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) of the University of Warsaw (Warsaw, Poland); (c) our 184-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk; (d) the 624-processor Beowulf cluster at the Baker Laboratory of Chemistry, Cornell University (Ithaca, NY, U.S.A.); and (e) computing resources at the KIAS Center of Advanced Computations. This research was also supported by an allocation of advanced computing resources provided by the U.S. National Science Foundation (http://www.nics.tennessee.edu/) and by the U.S. National Science Foundation through TeraGrid resources provided by the Pittsburgh Supercomputing Center.

■ REFERENCES

(1) Mackerell, A. D., Jr. Empirical Force Fields for Biological Macromolecules: Overview and Issues. *J. Comput. Chem.* **2004**, *25*, 1584−1604.

(2) Némethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for Use in the ECEPP/3 Algorithm with Application to Proline-Containing Peptides. *J. Phys. Chem.* **1992**, *96*, 6472−6484.

(3) Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham, T., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(4) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1615.

(5) Kolinski, A. Protein Modeling and Structure Prediction with a Reduced Representation. *Acta Biochim. Pol.* **2004**, *51*, 349−371.

(6) Kolinski, A.; Skolnick, J. Reduced Models of Proteins and their Applications. *Polymer* **2004**, *45*, 511−524.

(7) Tozzini, V. Coarse-Grained Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.

(8) Czaplewski, C.; Liwo, A.; Makowski, M.; Ołdziej, S.; Scheraga, H. A. In *Multiscale Approaches to Protein Modeling*; Koliński, A., Ed.; Springer: New York, 2010; Chapter 3, pp 1−18.

(9) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; Mcleavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **2008**, *51*, 91−97.

(10) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98−105.

(11) Liwo, A.; Khalili, M.; Scheraga, H. A. Ab Initio Simulations of Protein-Folding Pathways by Molecular Dynamics with the United-Residue Model of Polypeptide Chains. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 2362−2367.

(12) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. Molecular Dynamics with the United-Residue Model of Polypeptide Chains. II. Langevin and Berendsen-Bath Dynamics and Tests on Model α-Helical Systems. *J. Phys. Chem. B* **2005**, *109*, 13798−13810.

(13) Wales, D. J.; Scheraga, H. A. Global Optimization of Clusters, Crystals, and Biomolecules. *Science* **1999**, *285*, 1368−1372.

(14) Scheraga, H. A.; Pillardy, J.; Liwo, A.; Lee, J.; Czaplewski, C.; Ripoll, D. R.; Wedemeyer, W. J.; Arnautova, Y. A. Evolution of Physics-Based Methodology for Exploring the Conformational Energy Landscape of Proteins. *J. Comput. Chem.* **2002**, *23*, 28−34.

(15) Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. Advances in Protein Structure Prediction and De Novo Protein Design: A Review. *Chem. Eng. Sci.* **2006**, *61*, 966−988.

(16) Kolinski, A.; Godzik, A.; Skolnick, J. A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J. Chem. Phys.* **1993**, *98*, 7420−7433.

(17) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. Modification and Optimization of the United-Residue (UNRES) Potential Energy Function for Canonical Simulations. I. Temperature Dependence of the Effective Energy Function and Tests of the Optimization Method with Single Training Proteins. *J. Phys. Chem. B* **2007**, *111*, 260−285.

(18) Lee, J.; Scheraga, H. A.; Rackovsky, S. New Optimization Method for Conformational Energy Calculations on Polypeptides:

Conformational Space Annealing. *J. Comput. Chem.* **1997**, *18*, 1222–1232.

(19) Lee, J.; Liwo, A.; Scheraga, H. A. Energy-based *de novo* protein folding by conformational space annealing and an off-lattice united-residue force field: Application to the 10–55 fragment of staphylococcal protein A and to *apo*-calbindin D9K. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 2025–2030.

(20) Prentiss, M. C.; Wales, D. J.; Wolynes, P. G. Protein Structure Prediction Using Basin-Hopping. *J. Chem. Phys.* **2008**, *128*, 225106.

(21) Crippen, G. M.; Snow, M. E. A 1.8 Å Resolution Potential Function for Protein Folding. *Biopolymers* **1990**, *29*, 1479–1489.

(22) Seetharamulu, P.; Crippen, G. M. A Potential Function for Protein Folding. *J. Math. Chem.* **1991**, *6*, 91–110.

(23) Sali, A.; Shakhnovich, E.; Karplus, M. How Does a Protein Fold? *Nature* **1994**, *369*, 248–251.

(24) Camacho, C. J.; Thirumalai, D. A Criterion that Determines Fast Folding of Proteins: A Model Study. *Europhys. Lett.* **1996**, *35*, 627–632.

(25) Liwo, A.; Arłukowicz, P.; Ołdziej, S.; Czaplewski, C.; Makowski, M.; Scheraga, H. A. Optimization of the UNRES Force Field by Hierarchical Design of the Potential-Energy Landscape. I: Tests of the Approach Using Simple Lattice Protein Models. *J. Phys. Chem. B* **2004**, *108*, 16918–16933.

(26) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. Protein Tertiary Structure Recognition Using Optimized Hamiltonians with Local Interactions. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 9029–9033.

(27) Bryant, S. H.; Lawrence, C. E. An Empirical Energy Function for Threading Protein Sequence through the Folding Motif. *Proteins: Struct., Funct., Genet.* **1993**, *16*, 92–112.

(28) Miller, R. T.; Jones, D. T.; Thornton, J. M. Protein Fold Recognition by Sequence Threading: Tools and Assessment Techniques. *FASEB J.* **1996**, *10*, 171–178.

(29) Buchete, N. V.; Straub, J. E.; Thirumalai, D. Development of Novel Statistical Potentials for Protein Fold Recognition. *Curr. Opin. Struct. Biol.* **2004**, *14*, 225–232.

(30) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. I. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *J. Comput. Chem.* **1997**, *18*, 849–873.

(31) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. II: Parameterization of Local Interactions and Determination of the Weights of Energy Terms by Z-Score Optimization. *J. Comput. Chem.* **1997**, *18*, 874–887.

(32) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Cumulant-Based Expressions for the Multibody Terms for the Correlation between Local and Electrostatic Interactions in the United-Residue Force Field. *J. Chem. Phys.* **2001**, *115*, 2323–2347.

(33) Lee, J.; Ripoll, D. R.; Czaplewski, C.; Pillardy, J.; Wedemeyer, W. J.; Scheraga, H. A. Optimization of Parameters in Macromolecular Potential Energy Functions by Conformational Space Annealing. *J. Phys. Chem. B* **2001**, *105*, 7291–7298.

(34) Pillardy, J.; Czaplewski, C.; Liwo, A.; Wedemeyer, W. J.; Lee, J.; Ripoll, D. R.; Arłukowicz, P.; Ołdziej, S.; Arnautova, Y. A.; Scheraga, H. A. Development of Physics-Based Energy Functions that Predict Medium-Resolution Structure for Proteins of the α, β, and α/β Structural Classes. *J. Phys. Chem. B* **2001**, *105*, 7299–7311.

(35) Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. Optimization of the UNRES Force Field by Hierarchical Design of the Potential-Energy Landscape: 2. Off-Lattice Tests of the Method with Single Proteins. *J. Phys. Chem. B* **2004**, *108*, 16934–16949.

(36) Ołdziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. Optimization of the UNRES Force Field by Hierarchical Design of the Potential-Energy Landscape. 3. Use of Many Proteins in Optimization. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.

(37) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. Molecular Dynamics with the United-Residue Model of Polypeptide Chains. I. Lagrange Equations of Motion and Tests of Numerical Stability in the Microcanonical Mode. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.

(38) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Rojas, A. V.; Kaźmierkiewicz, R.; Makowski, M.; Murarka, R. K.; Scheraga, H. A. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G., Ed.; CRC Press: Boca Raton, FL, 2008; Chapter 8, pp 1391–1411.

(39) He, Y.; Xiao, Y.; Liwo, A.; Scheraga, H. A. Exploring the Parameter Space of the Coarse-Grained UNRES Force Field by Random Search: Selecting a Transferable Medium-Resolution Force Field. *J. Comput. Chem.* **2009**, *30*, 2127–2135.

(40) Kozłowska, U.; Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. Determination of Side-Chain-Rotamer and Side-Chain and Backbone Virtual-Bond-Stretching Potentials of Mean Force from AM1 Energy Surfaces of Terminally-Blocked Amino-Acid Residues, For Coarse-Grained Simulations of Protein Structure and Folding. II. Results, Comparison with Statistical Potentials, and Implementation in the UNRES Force Field. *J. Comput. Chem.* **2010**, *31*, 1154–1167.

(41) Sieradzan, A. K.; Scheraga, H. A.; Liwo, A. Determination of Effective Potentials for the Stretching of $C^{\alpha} \cdots C^{\alpha}$ Virtual Bonds in Polypeptide Chains for Coarse-Grained Simulations of Proteins from ab Initio Energy Surfaces of N-Methylacetamide and N-Acetylpyrrolidine. *J. Chem. Theory Comput.* **2012**, *8*, 1334–1343.

(42) Sieradzan, A. K.; Hansmann, U. H. E.; Scheraga, H. A.; Liwo, A. Extension of UNRES Force Field to Treat Polypeptide Chains with D-Amino Acid Residues. *J. Chem. Theory Comput.* **2012**, *8*, 4746–4757.

(43) Krupa, P.; Sieradzan, A. K.; Rackovsky, S.; Baranowski, M.; Ołdziej, S.; Scheraga, H. A.; Liwo, A.; Czaplewski, C. Improvement of the Treatment of Loop Structures in the UNRES Force Field by Inclusion of Coupling between Backbone- and Side-Chain-Local Conformational States. *J. Chem. Theory Comput.* **2013**, *9*, 4620–4632.

(44) Sieradzan, A. K.; Niadzvedtski, A.; Scheraga, H. A.; Liwo, A. Revised Backbone-Virtual-Bond-Angle Potentials to Treat the L- and D-Amino Acid Residues in the Coarse-Grained United Residue (UNRES) Force Field. *J. Chem. Theory Comput.* **2014**, *10*, 2194–2203.

(45) Sieradzan, A. K.; Krupa, P.; Scheraga, H. A.; Liwo, A.; Czaplewski, C. Physics-Based Potentials for the Coupling between Backbone- and Side-Chain-Local Conformational States in the United Residue (UNRES) Force Field for Protein Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 817–831.

(46) Sieradzan, A. K. Introduction of Periodic Boundary Conditions into UNRES Force Field. *J. Comput. Chem.* **2015**, *36*, 940–946.

(47) Kubo, R. Generalized Cumulant Expansion Method. *J. Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.

(48) Ołdziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Makowski, M.; Schafroth, H. D.; Kaźmierkiewicz, R.; Ripoll, D. R.; Pillardy, J.; Saunders, J.; Kang, Y.; Gibson, K.; Scheraga, H. Physics-Based Protein-Structure Prediction using a Hierarchical Protocol Based on the UNRES Force Field: Assessment in two Blind Tests. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7547–7552.

(49) He, Y.; Mozolewska, M. A.; Krupa, P.; Sieradzan, A. K.; Wirecki, T. K.; Liwo, A.; Kachlishvili, K.; Rackovsky, S.; Jagieła, D.; Ślusarz, R.; Czaplewski, C. R.; Ołdziej, S.; Scheraga, H. A. Lessons from Application of the UNRES Force Field to Predictions of Structures of CASP10 Targets. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 14936–14941.

(50) Khalili, M.; Liwo, A.; Scheraga, H. A. Kinetic Studies of Folding of the B-Domain of Staphylococcal Protein A with Molecular Dynamics and a United-Residue (UNRES) Model of Polypeptide Chains. *J. Mol. Biol.* **2006**, *355*, 536–547.

(51) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. How Adequate are One- and Two-Dimensional Free Energy Landscapes for Protein Folding Dynamics? *Phys. Rev. Lett.* **2009**, *102*, 238102.

(52) Maisuradze, G. G.; Senet, P.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. Investigation of Protein Folding by Coarse-Grained Molecular Dynamics with the UNRES Force Field. *J. Phys. Chem. A* **2010**, *114*, 4471–4485.

(53) Zhou, R.; Maisuradze, G. G.; Sunol, D.; Todorovski, T.; Macias, M. J.; Xiao, Y.; Scheraga, H. A.; Czaplewski, C.; Liwo, A. Folding Kinetics of WW Domains with the United Residue Force Field for Bridging Microscopic Motions and Experimental Measurements. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 18243−18248.

(54) He, Y.; Liwo, A.; Weinstein, H.; Scheraga, H. PDZ Binding to the BAR Domain of PICK1 is Elucidated by Coarse-grained Molecular Dynamics. *J. Mol. Biol.* **2011**, *405*, 298−314.

(55) Gołaś, E. I.; Maisuradze, G. G.; Senet, P.; Ołdziej, S.; Czaplewski, C.; Scheraga, H. A.; Liwo, A. Simulation of the Opening and Closing of Hsp70 Chaperones by Coarse-Grained Molecular Dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 1750−1764.

(56) Mozolewska, M.; Krupa, P.; Scheraga, H. A.; Liwo, A. Molecular Modeling of the Binding Modes of the Iron-Sulfur Protein to the Jac1 Co-Chaperone from Saccharomyces cerevisiae by All-Atom and Coarse-Grained Approaches. *Proteins: Struct., Funct., Genet.* **2015**, *83*, 1414−1426.

(57) Hansmann, U. H. E.; Okamoto, Y. Comparative Study of Multicanonical and Simulated Annealing Algorithms in the Protein Folding Problem. *Phys. A* **1994**, *212*, 415−437.

(58) Rhee, Y. M.; Pande, V. S. Multiplexed-Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. *Biophys. J.* **2003**, *84*, 775−786.

(59) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. Application of Multiplexing Replica Exchange Molecular Dynamics Method to the UNRES Force Field: Tests With α and $\alpha + \beta$ Proteins. *J. Chem. Theory Comput.* **2009**, *5*, 627−640.

(60) Lewandowska, A.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. β-hairpin-Forming Peptides; Models of Early Stages of Protein Folding. *Biophys. Chem.* **2010**, *151*, 1−9.

(61) Chodankar, S.; Aswal, V. K.; Kohlbrecher, J.; Vavrin, R.; Wagh, A. G. Small Angle Neutron Scattering Studies on Protein Denaturation Induced by Different Methods. *Pramana* **2008**, *71*, 1021−1025.

(62) Appavou, M.-S.; Gibrat, G.; Bellissent-Funel, M.-C. Temperature Dependence on Structure and Dynamics of Bovine Pancreatic Trypsin Inhibitor (BPTI): A Neutron Scattering Study. *Biochim. Biophys. Acta, Proteins Proteomics* **2009**, *1794*, 1398−1406.

(63) Meersman, F.; Atilgan, C.; Miles, A. J.; Bader, R.; Shang, W.; Matagne, A.; Wallace, B. A.; Koch, M. H. J. Consistent Picture of the Reversible Thermal Unfolding of Hen Egg-White Lysozyme from Experiment and Molecular Dynamics. *Biophys. J.* **2010**, *99*, 2255−2263.

(64) Ptitsyn, O. B. Molten Globule and Protein Folding. *Adv. Protein Chem.* **1995**, *47*, 83−229.

(65) Warme, P. K.; Momany, F. A.; Rumball, S. V.; Tuttle, R. W.; Scheraga, H. A. Computation of Structure of Homologous Proteins. α-Lactalbumin from Lysozyme. *Biochemistry* **1974**, *13*, 768−782.

(66) Seber, G. A.; Wild, C. J. *Nonlinear Regression*; Wiley: New York, 1989; pp 228−236.

(67) Hałabis, A.; Żmudzińska, W.; Liwo, A.; Ołdziej, S. Conformational Dynamics of the Trp-Cage Miniprotein at Its Folding Temperature. *J. Phys. Chem. B* **2012**, *116*, 6898−6907.

(68) Nanias, M.; Czaplewski, C.; Scheraga, H. A. Replica Exchange and Multicanonical Algorithms with the Coarse-Grained United-Residue (UNRES) Force Field. *J. Chem. Theory Comput.* **2006**, *2*, 513−528.

(69) Lee, J. New Monte Carlo Algorithm: Entropic Sampling. *Phys. Rev. Lett.* **1993**, *71*, 211−214.

(70) Sugita, Y.; Okamoto, Y. Replica-Exchange Multicanonical Algorithm and Multicanonical Replica-Exchange Method for Simulating Systems with Rough Energy Landscape. *Chem. Phys. Lett.* **2000**, *329*, 261−270.

(71) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Replica-Exchange Multicanonical and Multicanonical Replica-Exchange Monte Carlo Simulations of Peptides. I. Formulation and Benchmark Test. *J. Chem. Phys.* **2003**, *118*, 6664−6675.

(72) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Replica-Exchange Multicanonical and Multicanonical Replica-Exchange Monte Carlo Simulations of Peptides. II. Application to a More Complex System. *J. Chem. Phys.* **2003**, *118*, 6676−6688.

(73) Shen, H.; Liwo, A.; Scheraga, H. A. An Improved Functional Form for the Temperature Scaling Factors of the Components of the Mesoscopic UNRES Force Field for Simulations of Protein Structure and Dynamics. *J. Phys. Chem. B* **2009**, *113*, 8738−8744.

(74) Kolinski, A.; Skolnick, J. Discretized Model of Proteins. I. Monte Carlo Study of Cooperativity in Homopolypeptides. *J. Chem. Phys.* **1992**, *97*, 9412−9426.

(75) Rakowski, F.; Grochowski, P.; Lesyng, B.; Liwo, A.; Scheraga, H. A. Implementation of a Symplectic Multiple-Time-Step Molecular Dynamics Algorithm, Based on the United-Residue Mesoscopic Potential Energy Function. *J. Chem. Phys.* **2006**, *125*, 204107.

(76) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clus ters. *J. Chem. Phys.* **1982**, *76*, 637−649.

(77) Eastwood, M. P.; Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. Statistical Mechanical Refinement of Protein Structure Prediction Schemes: Cumulant Expansion Approach. *J. Chem. Phys.* **2002**, *117*, 4602−4615.

(78) Fukunishi, H.; Watanabe, 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.

(79) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.

(80) Streicher, W. W.; Makhatadze, G. I. Unfolding Thermodynamics of Trp-Cage, a 20 Residue Miniprotein, Studied by Differential Scanning Calorimetry and Circular Dichroism Spectroscopy. *Biochemistry* **2007**, *46*, 2876−2880.

(81) Gay, D. M. Algorithm 611. Subroutines for Unconstrained Minimization Using a Model/Trust-Region Approach. *ACM Trans. Math. Software* **1983**, *9*, 503−524.

(82) Murtagh, F.; Heck, A. *Multivariate Data Analysis*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1987.

(83) Zemla, A.; Venclovas, C.; Moult, J.; Fidelis, K. Processing and Evaluation of Predictions in CASP4. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 13−21.

(84) Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins: Struct., Funct., Genet.* **2004**, *57*, 702−710.

(85) Hao, M.-H.; Scheraga, H. A. Optimizing Potential Functions for Protein Folding. *J. Phys. Chem.* **1996**, *100*, 14540−14548.

(86) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimizing Physical Energy Functions for Protein Folding. *Proteins: Struct., Funct., Genet.* **2004**, *54*, 88−103.

(87) Lee, J.; Park, K.; Lee, J. Full Optimization of Linear Parameters of a United Residue Protein Potential. *J. Phys. Chem. B* **2002**, *106*, 11647−11657.

(88) *PyMOL Molecular Graphics System*, version 1.3; Schrödinger, LLC: New York, 2010; http://www.pymol.org (accessed Sept 14, 2010).

(89) Rotkiewicz, P.; Skolnick, J. Fast Procedure for Reconstruction of Full-Atom Protein Models from Reduced Representations. *J. Comput. Chem.* **2008**, *29*, 1460−1465.

(90) Wang, Q.; Canutescu, A. A.; Dunbrack, R. L., Jr SCWRL and MolIDE: Computer Programs for Side-Chain Conformation Prediction and Homology Modeling. *Nat. Protoc.* **2008**, *3*, 1832−1847.

(91) Koradi, R.; Billeter, M.; Wüthrich, K. MOLMOL: a Program for Display and Analysis of Macromolecular Structures. *J. Mol. Graphics* **1996**, *14*, 51−55.