Published in final edited form as:

J Comput Chem. 2009 October ; 30(13): 2127–2135. doi:10.1002/jcc.21215.

Exploring the Parameter Space of the Coarse-Grained UNRES Force Field by Random Search: Selecting a Transferable Medium-Resolution Force Field

YI HE^{1,2}, YI XIAO¹, ADAM LIWO^{2,3}, and HAROLD A. SCHERAGA³

¹Biomolecular Physics and Modeling Group, School of Physics, Huazhong University of Science and Technology, Wuhan 430074, China ²Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland ³Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301

Abstract

We explored the energy-parameter space of our coarse-grained UNRES force field for large-scale *ab initio* simulations of protein folding, to obtain good initial approximations for hierarchical optimization of the force field with new virtual-bond-angle bending and side-chain-rotamer potentials which we recently introduced to replace the statistical potentials. 100 sets of energy-term weights were generated randomly, and good sets were selected by carrying out replica-exchange molecular dynamics simulations of two peptides with a minimal α -helical and a minimal β -hairpin fold, respectively: the tryptophan cage (PDB code: 1L2Y) and tryptophan zipper (PDB code: 1LE1). Eight sets of parameters produced native-like structures of these two peptides. These eight sets were tested on two larger proteins: the engrailed homeodomain (PDB code: 1ENH) and FBP WW domain (PDB code: 1E0L); two sets were found to produce native-like conformations of these proteins. These two sets were tested further on a larger set of nine proteins with α or $\alpha + \beta$ structure and found to locate native-like structures of most of them. These results demonstrate that, in addition to finding reasonable initial starting points for optimization, an extensive search of parameter space is a powerful method to produce a transferable force field.

Keywords

protein folding; thermodynamic hypothesis; coarse-grained models; UNRES force field; force-field parameterization

Introduction

Protein folding is one of the most interesting problems that attracts much attention in many scientific fields. Although many years have passed, predicting protein structure and protein folding pathways based on Anfinsen's thermodynamic hypothesis¹ is still a great challenge. Knowledge-based methods, i.e., homology modeling, threading, and fragment methods, usually perform better in protein-structure prediction compared to physics-based methods.^{2–4} However, knowledge-based conformational-space search methods cannot be used in studying protein folding processes. On the other hand, knowledge-based statistical potentials

have been very successful in simulating folding pathways^{5,6}; however, in such applications,^{5,6} they are used in physics-based canonical or multicanonical simulations. Coarse-grained⁷ and all-atom⁸ statistical potentials can also locate native-like structures as the lowest in energy in unrestricted Monte Carlo simulations. The most commonly used fully physics-based methods are all-atom molecular dynamics or Monte Carlo methods.⁹ Although there are some successful examples of studying folding of small proteins by using this methodology,^{10,11} all-atom molecular dynamics with explicit solvent cannot be used for studying folding of large proteins, because of the high computational cost arising from the need to adopt very small time steps to simulate long-time folding processes. Larger proteins can be treated by all-atom molecular dynamics,¹² global-optimization methods^{13,14} with implicit-solvent models or all-atom canonical or generalized-ensemble Monte Carlo methods^{15,16} with implicit solvent; however, the maximum length of polypeptide chains is about 60 amino-acid residues. Despite size-limitations of the all-atom approach, important conclusions can be drawn from such studies, the most recent example being the unveiling of a new folding mechanism in a study of the C-terminal fragment of Top7.¹⁶

In the past decade, we have developed a physics-based united-residue force field (UNRES) for off-lattice protein structure simulations.¹⁷⁻³⁵ Because of the reduction of the number of interaction sites and variables, UNRES can be used in simulating the folding of large proteins in real time. As opposed to most united-residue force fields, which are largely knowledge-based, UNRES was carefully derived based on physics of interactions using a cluster-cumulant expansion of effective free energy of protein plus the surrounding solvent.²⁰ A hierarchical procedure was designed to parameterize the force field.²⁶⁻²⁸ Recently,³³ temperature dependence has been included in UNRES as a consequence of its definition as a restricted free-energy function; this modification of the force field enabled us to reproduce the thermodynamic characteristics of the folding transition. In addition, we have recently replaced the statistical potentials that account for the energetics of virtual-bond-angle bending³⁴ and the side-chain rotamers³⁵ with ones parameterized based on the results of quantum-mechanical calculations of model energy surfaces. We also carried out initial work on reparameterization of the temperature-dependent UNRES force field.³³ However, the optimized parameters were good for α but not for β -proteins. Later,³⁶ we included one α -(engrailed homeodomain; PDB code: 1ENH) and one β -protein (FBP28 WW domain; PDB code: 1E0L) in optimization simultaneously and were able to reproduce the experimental structure and free-energy profiles of these proteins; however, the force field had only limited transferability. One of the reason for this could be the fact that we started force-field-parameter optimization from a single set of parameters, which results in searching only a limited range of energy-parameter space.

The aim of the work reported here was to explore the energy-parameter space to find good initial energy-parameter sets for optimization of the new UNRES force field. The primary motivation for this was our recent replacement of knowledge-based¹⁷ virtual-bond-angle bending and side-chain rotamer potentials with ones derived from quantum-mechanical calculations.^{34,35} Although these potentials control only very local properties of a polypeptide chain, their radical change led to loss of balance of the other UNRES energy terms, which are responsible for the overall fold of a polypeptide chain. Consequently, the energy-term weights found in our earlier studies^{27,28,33} are no longer good as initial approximations. We, therefore, carried out an extensive random search of energy-term weights and tested the resulting weights by performing replica-exchange (REMD) simulations^{15,37,38} on two peptides, namely the tryptophan cage³⁹ (PDB code: 1L2Y) and the tryptophan zipper 2⁴⁰ (PDB code: 1LE1), which possess a minimum stable α -helix and β -hairpin fold, respectively. Having tested the parameter sets with these two peptides, we selected the parameter sets which produced native-like structures for both of them and used each of these sets to run multiplexed replica-exchange molecular dynamics (MREMD) simulations⁴¹ on two larger proteins 1ENH⁴² and 1E0L⁴³ and

selected the sets that found the native-like structures of these two proteins. The final parameter sets were tested with nine other proteins.

Methods

UNRES Force Field

In the UNRES model,¹⁷⁻³⁵ a sequence of α -carbon atoms connected by virtual bonds with attached side chains is used to represent a polypeptide chain. Each amino-acid residue has two interaction sites, one located in the middle between two consecutive C^α atoms and the other one is in the mass center of the corresponding side chain. The C^α atoms serve only to define the geometry, as shown in Figure 1. The UNRES force field has been carefully derived as a restricted free energy (RFE) function of an all-atom polypeptide chain plus the surrounding solvent.^{19,20} The all-atom energy is averaged over the degrees of freedom which are lost when passing from the all-atom to the simplified representation. The RFE is further decomposed into factors representing the interactions within and between a given number of united interaction sites. Expansion of these factors into the Kubo generalized cumulant series⁴⁴ enabled us to obtain approximate analytical expressions for the respective terms, including the multibody terms, which are derived in other force fields from structural databases or on a heuristic basis. The details of the theoretical basis have been described in our previous work.^{19,20} The energy function of the virtual-bond chain is expressed by eq. (1).

$$\begin{aligned}
 U(\mathbf{X}, T) = & w_{sc} \sum_{i < j} U_{sc_i sc_j}(\mathbf{X}) + w_{sc_p} \sum_{i \neq j} U_{sc_i p_j}(\mathbf{X}) \\
 & + w_{pp}^{VDW} \sum_{i < j-1} U_{p_i p_j}^{VDW}(\mathbf{X}) + w_{pp}^{el} f_2(T) \sum_{i < j-1} U_{p_i p_j}^{el}(\mathbf{X}) \\
 & + w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) + w_{tord} f_3(T) \sum_i U_{tord}(\gamma_i, \gamma_{i+1}) \\
 & + w_b \sum_i U_b(\theta_i) + w_{rot} \sum_i U_{rot}(\alpha_{sc_i}, \beta_{sc_i}) \\
 & + w_{corr}^{(3)} f_3(T) U_{corr}^{(3)}(\mathbf{X}) + w_{corr}^{(4)} f_4(T) U_{corr}^{(4)}(\mathbf{X}) \\
 & + w_{corr}^{(5)} f_5(T) U_{corr}^{(5)}(\mathbf{X}) + w_{corr}^{(6)} f_6(T) U_{corr}^{(6)}(\mathbf{X}) \\
 & + w_{turn}^{(3)} f_3(T) U_{turn}^{(3)}(\mathbf{X}) + w_{turn}^{(4)} f_4(T) U_{turn}^{(4)}(\mathbf{X}) \\
 & + w_{turn}^{(6)} f_6(T) U_{turn}^{(6)}(\mathbf{X}) + w_{bond} \sum_{i=1}^{nbond} U_{bond}(d_i)
 \end{aligned} \tag{1}$$

where $\mathbf{X}$ denotes the set of coordinates of a conformation of the coarse-grained chain and T is the absolute temperature. The terms $U_{sc_i sc_j}(\mathbf{X})$ are the potentials of mean force of the interaction of isolated side chains in water (with the contribution from the solvent already included). The terms $U_{sc_i p_j}(\mathbf{X})$ are the excluded-volume potentials of interactions between side chains and peptide groups. The terms $U_{p_i p_j}^{VDW}(\mathbf{X})$ are the Lennard-Jones potentials and $U_{p_i p_j}^{el}(\mathbf{X})$ are the averaged electrostatic-interaction potentials between peptide groups. The terms $U_{tor}(\gamma_i)$ and $U_{tord}(\gamma_i, \gamma_{i+1})$ are the virtual-bond torsional and double-torsional potentials, respectively. The terms $U_b(\theta_i)$, $U_{rot}(\alpha_{sc_i}, \beta_{sc_i})$, and $U_{bond}(d_i)$ denote the energy of virtual-bond angle bending, side-chain rotamers, and virtual-bond stretching, respectively. The terms $U_{corr}^{(m)}$ represent the correlation or multibody contribution from the backbone-local and backbone-electrostatic interactions. The terms $U_{turn}^{(m)}$ denote the correlation contribution involving m consecutive peptide groups. The energy-term weights corresponding to second- and higher-order generalized cumulants are multiplied by the appropriate scaling factor, $f_n(T)$, which is defined by eq. (2), where n is the order of the cumulant.

$$f_n(T) = \frac{\ln[\exp(1) + \exp(-1)]}{\ln\left\{\exp\left[\left(\frac{T}{T_0}\right)^{n-1}\right] + \exp\left[-\left(\frac{T}{T_0}\right)^{n-1}\right]\right\}} \quad (2)$$

where T_0 is the arbitrary reference temperature (we set $T_0 = 300$ K), and T is the current temperature.

All parameters except the energy-term weights which were being determined in this work were taken from our earlier work.^{17,19,24,25,34,35} The parameters of the $U_{SC_iSC_j}$ potentials [which have the Gay-Berne functional form⁴⁵] were taken from ref.¹⁷, where they were determined from protein-crystal data. The torsional (U_{tor}) and double-torsional (U_{tord}) potentials have been determined from the RFE surfaces of model system calculated by *ab initio* molecular quantum mechanics in ref.²⁴. Likewise, the peptide-group interaction ($U_{pi pj}^{\text{el}}$ and $U_{pi pj}^{\text{VDW}}$) and correlation ($U_{\text{corr}}^{(m)}$ and $U_{\text{turn}}^{(m)}$) terms were calculated in ref.²⁵; the parameters of the correlation terms were further refined in ref.²⁷ by optimization of the energy landscape for selected training proteins. The virtual-bond-angle bending (U_b) terms were calculated using *ab initio* molecular quantum mechanics in ref.³⁴. The virtual-bond rotamer (U_{rot}) and virtual-bond stretching (U_d) terms were calculated in ref.³⁵ from the energy surfaces of terminally blocked amino-acid residues calculated with the semiempirical AM1 method of molecular quantum mechanics. The $U_{SC_{ij}}$ terms were taken directly from ref.¹⁹.

Search of Energy Parameter Space

We chose the energy-term weights [the w s of eq. (1)], because the force field is most sensitive to these parameters and they can be determined only by optimizing the force field on selected training proteins.^{21-23,28} In our previous work,^{21-23,28} we optimized the force field using old force field parameters as a starting point. This could result in too narrow a range of the search of the energy-parameter space and, consequently, reduced the performance of the force field. Because we recently modified the force field significantly by introducing temperature dependence³³ and the new physics-based U_b ³⁴ and U_{rot} ³⁵ potentials, we considered it worthwhile to perform a more extensive search of parameter space than optimization, started from a given set of parameters, can provide. The procedure was as follows:

1. Generate a number of different sets of energy-term weights at random; in this work we generated 100 sets, each weight being selected from a uniform distribution from $w_x^{\text{low}} \leq w_x \leq w_x^{\text{up}}$, where w_x denotes the weight of energy term x and the superscripts “low” and “up” mark the lower and upper boundary, respectively. There are 11 variable weights, which prohibits systematic grid search of the weight space.
2. With each set, carry out REMD or MREMD runs on a peptide with a minimal α -helix fold (we chose 1L2Y) and a protein with a minimal β -sheet fold (we chose 1LE1 whose structure is a single β -hairpin).
3. Select those sets of weights which result in the appearance of some native-like structures of the small test proteins in step 2.
4. Test the sets of weights selected in step 3 by running REMD or MREMD on larger proteins containing α - and β -folds (we chose 1ENH and 1EOL, respectively, in this work).
5. Select those sets of weights which result in the appearance of some native conformations of the test proteins in step 4.
6. Test the sets of weights selected in step 5 on a larger set of proteins.

Following our earlier work,²⁸ we considered only the correlation terms up to the fourth order and, consequently, set the weights of all correlation terms of order higher than 4 at zero. All other weights except that of the side-chain-interaction potentials, $[U_{SC_iSC_j}; \text{eq. (1)}]$ which was fixed at $w_{SC} = 1$ were variable. The value of w_{SC} was fixed, because the $U_{SC_iSC_j}$ potentials had been scaled in our earlier work¹⁷ to reproduce the experimental free energies of transfer of terminally blocked amino acids from water to *n*-octanol determined by Fauchere and Pliska;⁴⁶ consequently, they have strict connection to the energetics of side-chain interaction. The boundaries of the weights are listed in Table 1. These boundaries were chosen based on the experience from our previous work^{21-23,28,33} in which the correct geometry of regular secondary-structure elements as well as reasonable ranges of UNRES energies were produced.

Structure Comparison

We used the α -carbon root-mean-square deviation (C^α RMSD) to assess the overall similarity of the calculated and experimental structures. To assess the native-likeness of the two minimal-fold peptides: 1L2Y and 1LE1 and of fragments of larger proteins, we used the fraction of native contacts between the peptide groups as defined in ref.²⁷ and the quantity q defined by eq. (3) and introduced in our previous work,³³ which is a measure of compatibility between the residue-residue contacts in a given conformation to those in the experimental conformation.

$$q_i = 1 - \frac{1}{N_i} \left\{ \sum_{k \in \{i\}} \sum_{\substack{l \in \{i\} \\ l \leq k-m}} \exp \left[-\frac{(d_{C_k^\alpha C_l^\alpha} - d_{C_k^\alpha C_l^\alpha}^{\text{nat}})^2}{4(d_{C_k^\alpha C_l^\alpha}^{\text{nat}})^2} \right] + \exp \left[-\frac{(d_{SC_k SC_l} - d_{SC_k SC_l}^{\text{nat}})^2}{4(d_{SC_k SC_l}^{\text{nat}})^2} \right] \right\} \quad (3)$$

where $\{i\}$ is the set of residues constituting fragment i (we note that the fragment does not have to be continuous; e.g., we can consider a β -sheet composed of two nonsequential strands as a fragment), N_i is the total number of distances (both between the C^α atoms and between the SC pseudoatoms) in fragment i , m is the minimum separation of residues in the amino-acid

sequence at which the distances are calculated (we set $m = 3$), $d_{C_k^\alpha C_l^\alpha}$ or $d_{SC_k SC_l}$ denotes the distance between the α -carbon atoms or side-chain centers of residues k and l , respectively, of the conformation under consideration, and the same quantities with the nat superscript denote the distances in the experimental structure. $q = 0$ if the conformation considered is the experimental structure and $q > 0.5$ indicates that the structure considered and the experimental structure are grossly dissimilar.

Simulation Details

All REMD and MREMD simulations were carried out at 20 temperatures $T = 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 420, 440, 460$, and 500 K. The frequency of replica exchange and the frequency of snapshot dumping was 20,000 MD steps. The Berendsen thermostat,⁴⁷ implemented in UNRES/MD in our previous work,³¹ was used to maintain constant temperature; the coupling constant was $\tau = 0.489$ fs, which was assessed in our previous work³¹ to provide a compromise between the extent of fluctuations of the potential and kinetic energy. The time increment in integrating the equations of motion was $\delta t = 4.89$ fs. For the two minimal-fold peptides: 1L2Y and 1LE1, one trajectory per temperature was run (a total of 20 trajectories; a REMD simulation) with the total of 5,000,000 MD steps per trajectory. For the other 11 systems 2 trajectories per temperature were run (a total of 40 trajectories; a MREMD simulation) with the total of 10,000,000 steps per trajectory. Thus, the total time per trajectory was 24.45 and 48.9 ns, respectively; it should be noted,

however, that the UNRES time scale is stretched compared to the all-atom time scale because of averaging of the fast-varying degrees of freedom, especially those of the solvent.³¹

The weighted-histogram analysis method (WHAM),⁴⁸ modified in our recent work³³ to treat simulations with the UNRES force field, was implemented in the computation of heat-capacity as well as ensemble-averaged RMSD and q . We analyzed snapshots from the last 1,000,000 MD steps for 1L2Y and 1LE1 and from the last 2,000,000 MD steps for the other systems. To dissect the conformational ensembles, we applied the procedure described in our recent work,³³ in which a cluster analysis is carried out for the selected batch of conformations, and the probabilities of the clusters are computed by summing up the contributions from the constituent conformations at a selected temperature calculated with WHAM. The clusters are ranked following decreasing probability. The single-link algorithm^{49,50} was applied for clustering.

Results and Discussion

Initial Selection of Weight Sets by Tests with Minimal-Fold Peptides

As stated in the Methods section, we first randomly generated 100 sets of energy-term weights, each within the boundaries specified in Table 1. The sets of conformations of 1L2Y and 1LE1 obtained in the last 1,000,000 MD steps of simulations with each of the 100 sets of energy-term weights were analyzed for the presence of native-like conformations. For 1L2Y, we required that the C^α RMSD be $< 3.0 \text{ \AA}$ and $q_{\text{Asn2-Leu10}} < 0.35$, which implies that the α -helical segment from residues Asn2 to Leu10 had a contact pattern characteristic of an α -helix. For 1LE1 we required that the C^α RMSD be $< 2.5 \text{ \AA}$ and a contact pattern between peptide groups characteristic of a β -hairpin with a tolerance of ± 3 residue shift with respect to the native β -hairpin. The native-like structures of both peptides were found in simulations carried out with eight sets of energy-term weights; these weights are summarized in Table 2. A summary of the results: the lowest RMSD values, the lowest ensemble-averaged RMSD values [computed by using the WHAM method as stated in ref. ³³], and the ensemble-averaged q values at temperatures corresponding to those of the lowest-ensemble-averaged RMSD obtained with these eight sets of weights is presented in Table 3.

It can be seen that sets 1, 3, and 6 produced not only some native-like structures but the ensemble-averaged RMSD and q values also are low for both 1L2Y and 1LE1, which means that the folding transition occurs for structures which are close to the experimental structures of these two peptides. Figures 2 and 3 present graphs of heat capacity (C_v) and ensemble-averaged RMSD vs. temperature as well as the experimental structures and representative structures at $T = 280 \text{ K}$ (i.e., lower than the folding-transition temperature). It can be seen that the heat capacity curve of 1L2Y has a peak at $T = 317 \text{ K}$ (Fig. 2A), which corresponds to the inflection point of the ensemble-averaged RMSD curve (Fig. 2B). For 1LE1, the heat-capacity peak and the inflection point of the ensemble-averaged RMSD curve occur at $T = 297 \text{ K}$. The experimental melting temperatures of these two peptides are $T_m = 315 \text{ K}$ ³⁹ and $T_m = 323 \text{ K}$,⁴⁰ respectively, which means that the calculated melting temperature of 1LE1 is significantly lower than the experimental melting temperature but the aim of our study was to find reasonable starting sets of energy-term weights and not to obtain exact correspondence between the calculated and experimental folding characteristics. Nevertheless, it is remarkable that the calculated melting temperatures, obtained not only with energy-term-weight set 1 but also with the remaining sets (Table 3), are within physiological range, a feature that occurred in our previous studies only when we explicitly requested that folding transition occur at a specified temperature.^{33,36} The force fields that we derived earlier to locate the native-like structures as the lowest minima of the UNRES energy function produced melting temperatures of the order of $T_m = 1000 \text{ K}$ or higher.³⁸ That the melting temperatures are reasonable in our current study seems to result from our freezing the weight w_{SC} at 1 and with the parameters of the $U_{\text{SC}_i\text{SC}_j}$

potentials at the values determined in our previous work;¹⁷ although these parameters have been determined from protein-crystal data, they were subsequently scaled to reproduce the free energies of transfer of terminally blocked amino acids from water to *n*-octanol determined by Fauchere and Pliska.⁴⁶ Therefore, the side-chain-interaction potentials are in strict correspondence with the thermodynamics of side chain-side chain interactions. The weights of the other UNRES energy terms had to accommodate to them in order to produce native-like structures of both peptides, and the thermodynamic parameters of folding are, therefore, in the range of the experimental ones even though no melting-temperature information was supplied as input to the procedure. In our previous optimizations, which did not consider thermodynamic data,^{27,28} the weight w_{SC} and the parameters of the $U_{SC;SC_j}$ potentials were varied and the connection to thermodynamics was, therefore, lost.

Selection of Sets of Energy-Term Weights by Tests with 1ENH and 1EOL

The eight sets of energy-term weights were subsequently tested on two larger proteins: 1ENH (a three-helix bundle) and 1EOL (a three-stranded antiparallel β -sheet). Sets 1 and 2 produced native-like structures of both proteins; these results are collected in Table 4 and in Figure 4D and K. In Table 4 the probabilities and average RMSD values of the clusters containing native-like structures (detected based on topology rather than low ensemble-averaged RMSD values) are listed in addition to lowest RMSD values found for the conformations corresponding to the last 2,000,000 MD steps. However, as opposed to 1L2Y and 1LE1, the ensemble-averaged RMSD values were high at all temperatures below the folding-transition temperature. For the 1EOL run with set 1 of energy-term weights, the high ensemble-averaged RMSD value was caused by misalignment of the N- and the C-terminal fragments that cap the three-stranded β -sheet in the experimental structure, while conformations with the correct three-stranded antiparallel β -sheet form the dominant cluster. For the 1ENH run with both energy-term-weight sets, and for the 1EOL run with set 2, the clusters containing native-like structures have low probabilities (Table 4).

The two sets of energy-term weights were subsequently tested on nine other proteins: 2HEP (an α -helical hairpin), 1BDD (a three-helix bundle), 1GAB (a three-helix bundle), 1LQ7 (a three-helix bundle), 1POU (α -helical structure), 1CLB (α -helical structure), 1FSD ($\alpha + \beta$ structure), 1E0G ($\alpha + \beta$ structure), and 1PGA ($\alpha + \beta$ structure). The results are also summarized in Table 4 and the best native-like structures obtained in the last 2,000,000 MD steps are shown in Figure 4. As shown in Table 4 and Figure 4, parameter set 1 produced medium-resolution native-like structures for most of these proteins except 1PGA, for which a mirror image of the native fold was obtained. Nevertheless, sets 1 and 2 of the energy-term weights produced quite transferable force fields with which to obtain medium-resolution structures. Set 1 appears better than set 2 as far as the resolution is concerned.

Conclusions

In this work, we have tested the performance of randomly generated sets of energy-term weights of the UNRES force field. Eight out of 100 sets were able to produce both α and β structures as assessed by tests with minimum-fold peptides: the tryptophan cage and the tryptophan zipper. Consequently, the space of reasonable UNRES parameters is not as restricted as it would appear from our previous studies^{33,36} in which we started force-field optimization from a given point in parameter space. Two of the energy-term-weight sets produced medium-resolution native-like structures of nine proteins with different secondary structure and size from 28 to 75 residues. Our results suggest that random search of UNRES force field parameter space is a good method to generate starting points for detailed optimization according to the procedure described in our previous work.³³ More minimal-fold

peptides or small proteins could be used as filters to select good parameters. This research is now being carried out in our laboratory.

Acknowledgments

This research was conducted by using the resources of (a) our 800-processor Beowulf cluster at the Baker Laboratory of Chemistry and Chemical Biology, Cornell University, (b) the National Science Foundation Terascale Computing System at the Pittsburgh Super-computer Center, (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Jülich, Germany, (d) the Beowulf cluster at the Department of Computer Science, Cornell University, (e) the resources of the Center for Computation and Technology at Louisiana State University, which is supported by funding from the Louisiana legislature, (f) our 45-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk, (g) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (h) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

Contract/grant sponsor: National Science Foundation of China (NSCF); contract/grant number: 30525037

Contract/grant sponsor: Polish Ministry of Science and Higher Education; contract/grant numbers: BW-R300-5-0001-8, DS/8372-4-0-138-8

Contract/grant sponsor: U.S. National Institutes of Health; contract/grant number: GM-14312

Contract/grant sponsor: U.S. National Science Foundation; contract/grant number: MCB05-41633

References

1. Anfinsen CB. Science 1973;181:223. [PubMed: 4124164]
2. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Proteins Struct Funct Genet 2003;53:457. [PubMed: 14579334]
3. Skolnick J, Zhang Y, Arakaki AK, Koliński A, Boniecki M, Szilagy A, Kihara D. Proteins Struct Funct Genet 2003;53:469. [PubMed: 14579335]
4. Scheraga HA, Liwo A, Oldziej S, Czaplewski C, Pillardy J, Ripoll DR, Vila JA, Kaźmierkiewicz R, Arnautova YA, Jagielska A, Chinchio M, Nanas M. Front Biosci 2004;9:3296. [PubMed: 15353359]
5. Kmiecik S, Kolinski A. Proc Natl Acad Sci USA 2007;104:12330. [PubMed: 17636132]
6. Kmiecik S, Kolinski A. Biophys J 2008;94:726. [PubMed: 17890394]
7. Kolinski A, Godzik A, Skolnick J. J Chem Phys 1993;98:7420.
8. Yang JS, Chen WW, Skolnick J, Shakhnovich EI. Structure 2007;115:53. [PubMed: 17223532]
9. Zimmermann O, Hansmann UHE. Biochim Biophys Acta 2008;1784:252. [PubMed: 18036571]
10. Duan Y, Kollman PA. Science 1998;282:740. [PubMed: 9784131]
11. Simmerling C, Strockbine B, Roitberg AE. J Am Chem Soc 2002;124:11258. [PubMed: 12236726]
12. Jang S, Kim E, Shin S, Pak Y. J Am Chem Soc 2003;125:14841. [PubMed: 14640661]
13. Vila JA, Ripoll DR, Scheraga HA. Proc Natl Acad Sci USA 2003;100:14812. [PubMed: 14638943]
14. Schug A, Wenzel W. Biophys J 2006;90:4273. [PubMed: 16565067]
15. Hansmann UHE. Chem Phys Lett 1997;281:140.
16. Mohanty S, Meinke JH, Zimmermann O, Hansmann UHE. Proc Natl Acad Sci USA 2008;105:8004. [PubMed: 18408166]
17. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. J Comput Chem 1997;18:849.
18. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. J Comput Chem 1997;18:874.
19. Liwo A, Kaźmierkiewicz R, Czaplewski C, Groth M, Oldziej S, Rackovsky RJ, Wawakand S, Pincus MR, Scheraga HA. J Comput Chem 1998;19:259.
20. Liwo A, Czaplewski C, Pillardy J, Scheraga HA. J Chem Phys 2001;115:2323.
21. Lee J, Ripoll DR, Czaplewski C, Pillardy J, Wedemeyer WJ, Scheraga HA. J Phys Chem B 2001;105:7291.

22. Pillardy J, Czaplowski C, Liwo A, Wedemeyer WJ, Lee J, Ripoll DR, Arlukowicz P, Oldziej S, Arnautova YA, Scheraga HA. *J Phys Chem B* 2001;105:7299.
23. Liwo A, Arlukowicz P, Czaplowski C, Oldziej S, Pillardy J, Scheraga HA. *Proc Natl Acad Sci USA* 2002;99:1937. [PubMed: 11854494]
24. Oldziej S, Kozłowska U, Liwo A, Scheraga HA. *J Phys Chem A* 2003;107:8035.
25. Liwo A, Oldziej S, Czaplowski C, Kozłowska U, Scheraga HA. *J Phys Chem B* 2004;108:9421.
26. Liwo A, Arlukowicz P, Oldziej S, Czaplowski C, Makowski M, Scheraga HA. *J Phys Chem B* 2004;108:16918.
27. Oldziej S, Liwo A, Czaplowski C, Pillardy J, Scheraga HA. *J Phys Chem B* 2004;108:16934.
28. Oldziej S, Lagiewka J, Liwo A, Czaplowski C, Chinchio M, Nancias M, Scheraga HA. *J Phys Chem B* 2004;108:16950.
29. Liwo A, Khalili M, Scheraga HA. *Proc Natl Acad Sci USA* 2005;102:2362. [PubMed: 15677316]
30. Khalili M, Liwo A, Rakowski F, Grochowski P, Scheraga HA. *J Phys Chem B* 2005;109:13785. [PubMed: 16852727]
31. Khalili M, Liwo A, Jagielska A, Scheraga HA. *J Phys Chem B* 2005;109:13798. [PubMed: 16852728]
32. Khalili M, Liwo A, Scheraga HA. *J Mol Biol* 2006;355:536. [PubMed: 16324712]
33. Liwo A, Khalili M, Czaplowski C, Kalinowski S, Oldziej S, Wachucik K, Scheraga HA. *J Phys Chem B* 2007;111:260. [PubMed: 17201450]
34. Kozłowska U, Liwo A, Scheraga HA. *J. Phys Condens Matter* 2007;19:285203.
35. Kozłowska U, Maisuradze G, Liwo A, Scheraga HA. *Protein Sci* 2007;16(Suppl 1):92. [PubMed: 17192591]
36. Liwo, A.; Czaplowski, C.; Oldziej, S.; Kozłowska, U.; Makowski, M.; Kalinowski, S.; Kaźmierkiewicz, R.; Shen, H.; Maisuradze, G.; Scheraga, HA.; John von Neumann Institute for Computing (NIC). In: Münster, G.; Wolf, D.; Kremer, M., editors. *NIC Series, NIC Symposium 2008*; Jülich, Germany. 20-21 February 2008; 2008. 3963
37. Sugita Y, Okamoto Y. *Chem Phys Lett* 1999;314:141.
38. Nancias M, Czaplowski C, Scheraga HA. *J Chem Theor Comput* 2006;2:513.
39. Neidigh JW, Fesinmeyer RM, Andersen NH. *Nat Struct Biol* 2002;9:425. [PubMed: 11979279]
40. Cochran AG, Skelton NJ, Starovasnik MA. *Proc Natl Acad Sci USA* 2001;98:5578. [PubMed: 11331745]
41. Rhee YM, Pande VS. *Biophys J* 2003;84:775. [PubMed: 12547762]
42. Macias MJ, Gervais V, Civera C, Oschkinat H. *Nat Struct Biol* 2000;7:375. [PubMed: 10802733]
43. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. *Protein Sci* 1994;3:1779. [PubMed: 7849596]
44. Kubo RJ. *Phys Soc Jpn* 1962;17:1100.
45. Gay JG, Berne BJ. *J Chem Phys* 1981;74:3316.
46. Fauchere J-L, Pliška V. *Eur J Med Chem* 1983;18:369.
47. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. *J Chem Phys* 1984;81:3684.
48. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Osenberg JMR. *J Comput Chem* 1992;13:1011.
49. Murtagh, F. *Multidimensional Clustering Algorithms*. Physica-Verlag; Vienna: 1985.
50. Murtagh, F.; Heck, A. *Multivariate Data Analysis*. Kluwer Academic; Dordrecht: 1987.

Figure 1.

The UNRES model of the polypeptide chain. The united peptide groups (p) are shown as dark circles, and open circles represent the C^α atoms which serve only as geometric points. The side chains are represented by ellipsoids, SC's; each is located at the mass center b_{SC} connected to the corresponding C^α . The p 's are located in the middle of two consecutive C^α atoms. The virtual-bond angles θ , the virtual-bond dihedral angles γ , and the angles α_{SC} and β_{SC} that define the location of a side chain with respect to the backbone are also indicated.

Figure 2.

Plots of (A) heat capacity, (B) ensemble-averaged RMSD from the native structure obtained in the MREMD simulation of 1L2Y with set 1 of energy-term weights, (C) C^α traces of the experimental structure, and (D) the representative structure of the dominant cluster of this peptide. The last 1,000,000 MD steps were taken to compute averages and clustering. The N-terminus is marked with “N” in parts C and D.

Figure 3. Plots of (A) heat capacity, (B) ensemble-averaged RMSD from the native structure obtained in the MREMD simulation of 1L2Y with set 2 of energy-term weights, (C) C^α traces of the experimental structure, and (D) the representative structure of the dominant cluster of this peptide. The last 1,000,000 MD steps were taken to compute averages and clustering. The N-terminus is marked with “N” in parts C and D.

Figure 4.

Most native-like structures obtained in MREMD simulations (left) and the experimental structures (right) of the proteins of the test set obtained with energy-term weights set 1. (A) 2HEP, (B) 1BDD, (C) 1GAB, (D) 1ENH, (E) 1LQ7, (F) 1POU, (G) 1CLB, (H) 1FSD, (I) 1E0G, (J) 1PGA, (K) 1E0L. Except for 1PGA, all calculated structures are the experimental structures; for 1PGA the structure with the best contact pattern which is the mirror image of the experimental structure is shown. The N-termini of the proteins are marked with “N” and the PDB codes are shown above the experimental structures. The RMSD values are shown as $\rho_{\min}$ in Table 4.

Table 1

Boundaries of Energy-Term Weights

	Lower bound	Upper bound
w_{SCp}	1.0	1.8
w_{pp}^{el}	0.6	1.2
w_{ang}	0.6	2.0
w_{rot}	0.1	2.0
w_{tor}	1.0	2.0
w_{tord}	1.0	1.8
$w_{corr}^{(4)}$	0.0	0.3
$w_{corr}^{(3)}$	0.1	0.4
$w_{turn}^{(3)}$	1.0	2.0
$w_{turn}^{(4)}$	0.4	1.4
w_{pp}^{vdW}	0.1	0.6

Table 2
Sets of Energy-Term Weights that Produced Native-Like Structures of Both 1L2Y and 1LE1 In MREMD Simulations

Set	w_{scp}	w_{el}^{pp}	w_{ang}	w_{rot}	w_{tor}	w_{fird}	$w_{corr}^{(4)}$	$w_{corr}^{(3)}$	$w_{turn}^{(3)}$	$w_{turn}^{(4)}$	w_{vdw}^{pp}
1	1.23315	0.84476	0.62954	0.10554	1.84316	1.26571	0.19212	0.37367	1.40323	0.64673	0.23173
2	1.78996	1.19576	1.40891	0.35725	1.68959	1.23368	0.26828	0.22620	1.85748	0.57574	0.19938
3	1.74206	1.18097	1.77232	0.36656	1.86689	1.78732	0.08580	0.32061	1.36552	0.74055	0.21898
4	1.59928	1.15712	1.99574	0.51033	1.58651	1.43755	0.20836	0.11872	1.41638	1.15683	0.29410
5	1.22216	1.02765	0.96511	0.18784	1.87436	1.77764	0.07738	0.39796	1.43672	0.52335	0.12415
6	1.35948	0.85084	1.78132	0.15290	1.60146	1.41820	0.19152	0.27342	1.55878	0.72243	0.10444
7	1.26246	1.16824	1.34348	1.55169	1.89678	1.74862	0.05212	0.33001	1.65033	0.87174	0.33966
8	1.16545	0.88564	0.90563	1.05048	1.19211	1.27904	0.21272	0.38661	1.74187	1.02105	0.57488

Table 3
Results of MREMD Simulation of 1L2Y and 1LE1 Obtained with Energy-Term Weights of Table 2

Set	1L2Y				1LE1			
	$\rho_{\min}^a$	$\langle \phi \rangle^b$	$\langle q \rangle^c$	$T_f [K]^d$	$\rho_{\min}^a$	$\langle \phi \rangle^b$	$\langle q \rangle^c$	$T_f [K]^d$
1	2.126	2.920	0.316	317	1.939	2.724	0.271	297
2	2.508	4.361	0.350	337	2.245	4.465	0.473	299
3	2.656	3.097	0.290	308	2.206	3.198	0.335	279
4	2.272	4.625	0.372	347	1.807	4.450	0.486	303
5	2.409	4.360	0.348	354	2.204	5.013	0.502	316
6	2.373	3.107	0.321	294	2.141	2.537	0.262	300
7	2.197	4.836	0.417	348	1.591	2.364	0.255	314
8	2.120	3.097	0.369	323	1.764	4.176	0.470	316

^aLowest RMSD value in the last 1,000,000 steps of the MREMD simulation for a specific parameter set and peptide.

^bLowest ensemble-averaged RMSD in the last 1,000,000 steps of the MREMD simulation for a specific parameter set and peptide.

^cEnsemble-averaged q at the temperature corresponding to the lowest ensemble-averaged RMSD in the last 1,000,000 steps of the MREMD simulation for a specific parameter set and peptide.

^dFolding temperature.

Table 4
MREMD Results Obtained with Energy-Term-Weight Set 1 and 2 for all Test Proteins

PDB ID	Size	Secondary structure	Set 1					Set 2				
			$\rho_{\min}^a$	Fragment q^b	$\langle \rho \rangle^c$	p^d	n^e	$\rho_{\min}^a$	Fragment q^b	$\langle \rho \rangle^c$	p^d	n^e
2HEP	42	<i>aa</i>	3.33	0.09/0.15	4.74	0.84	1	4.81	0.39/0.21	6.14	0.17	3
1BDD	46	<i>aaa</i>	3.17	0.24/0.24/0.13	4.82	0.89	1	4.66	0.28/0.42/0.14	7.40	0.06	5
1GAB	47	<i>aaa</i>	4.25	0.34/0.29/0.45	6.22	0.34	2	4.88	0.52/0.09/0.29	6.43	0.11	2
1ENH	54	<i>aaa</i>	4.52	0.24/0.29/0.22	6.39	0.16	3	4.86	0.65/0.28/0.43	6.31	0.08	4
1LQ7	67	<i>aaa</i>	3.39	0.14/0.12/0.11	5.73	0.11	4	2.40	0.17/0.20/0.14	3.51	0.38	2
1POU	71	<i>aaaa</i>	5.13	0.23/0.41/0.25/0.27	6.22	0.27	2	5.37	0.26/0.38/0.67/0.25	6.61	0.21	2
1CLB	75	<i>aaaa</i>	4.36	0.27/0.22/0.15/0.27	6.53	0.24	1	5.85	0.35/0.66/0.13/0.28	7.67	0.16	1
1FSD	28	<i>$\beta\alpha$</i>	3.47	0.43/0.12	5.21	0.07	6	3.13	0.36/0.18	5.41	0.03	2
1E0G	48	<i>$\beta\alpha\alpha\beta$</i>	4.79	0.05/0.27/0.19/0.05	6.00	0.32	1	5.15	0.05/0.19/0.70/0.07	6.88	0.28	1
1PGA	56	<i>$\beta\alpha\beta$</i>	6.29	0.61/0.17/0.18	<i>f</i>	<i>f</i>	<i>f</i>	7.67	0.41/0.56/0.58	<i>f</i>	<i>f</i>	<i>f</i>
			9.68	0.63/0.15/0.27 ^g	10.1	0.88	1	10.6	0.59/0.22/0.47 ^h	10.9	0.99	1
1E0L	37	<i>$\beta\beta$</i>	5.65	0.56/0.50	10.0	0.98	1	4.62	0.47/0.37	7.91	0.10	2

^aLowest RMSD found in the last 2,000,000 MD steps.
^bThe q values corresponding to the secondary-structure elements specified in column 2.
^cEnsemble-averaged RMSD of the most native-like cluster at $T = 280$ K.
^dProbability of the most native-like cluster at $T = 280$ K.
^eRank of the most native-like cluster according to decreasing probability at $T = 280$ K.
^fNative-like conformations had negligible weight at $T = 280$ K.
^gFor 1PGA the lowest-RMSD structure does not have the fully developed N-terminal hairpin and the structure best matching the topology of the native image which is listed in this entry (see Fig. 4J) with a higher RMSD value.