Investigation of Protein Folding by Coarse-Grained Molecular Dynamics with the UNRES Force Field

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY A · FEBRUARY 2010

Impact Factor: 2.69 · DOI: 10.1021/jp9117776 · Source: PubMed

Investigation of Protein Folding by Coarse-Grained Molecular Dynamics with the UNRES Force Field

Gia G. Maisuradze,[†] Patrick Senet,[†,‡] Cezary Czaplewski,[†,§] Adam Liwo,[†,§] and Harold A. Scheraga*[,†]

Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), UMR 5209 CNRS Université de Bourgogne, 9 Avenue A. Savary, BP 47870, F-21078 Dijon, Cedex, France, and Faculty of Chemistry, University of Gdańsk, Sobieskiego 18, 80-952 Gdańsk, Poland

Received: December 13, 2009; Revised Manuscript Received: January 27, 2010

Coarse-grained molecular dynamics simulations offer a dramatic extension of the time-scale of simulations compared to all-atom approaches. In this article, we describe the use of the physics-based united-residue (UNRES) force field, developed in our laboratory, in protein-structure simulations. We demonstrate that this force field offers about a 4000-times extension of the simulation time scale; this feature arises both from averaging out the fast-moving degrees of freedom and reduction of the cost of energy and force calculations compared to all-atom approaches with explicit solvent. With massively parallel computers, microsecond folding simulation times of proteins containing about 1000 residues can be obtained in days. A straightforward application of canonical UNRES/MD simulations, demonstrated with the example of the N-terminal part of the B-domain of staphylococcal protein A (PDB code: 1BDD, a three-α-helix bundle), discerns the folding mechanism and determines kinetic parameters by parallel simulations of several hundred or more trajectories. Use of generalized-ensemble techniques, of which the multiplexed replica exchange method proved to be the most effective, enables us to compute thermodynamics of folding and carry out fully physics-based prediction of protein structure, in which the predicted structure is determined as a mean over the most populated ensemble below the folding-transition temperature. By using principal component analysis of the UNRES folding trajectories of the formin-binding protein WW domain (PDB code: 1E0L; a three-stranded antiparallel β-sheet) and 1BDD, we identified representative structures along the folding pathways and demonstrated that only a few (low-indexed) principal components can capture the main structural features of a protein-folding trajectory; the potentials of mean force calculated along these essential modes exhibit multiple minima, as opposed to those along the remaining modes that are unimodal. In addition, a comparison between the structures that are representative of the minima in the free-energy profile along the essential collective coordinates of protein folding (computed by principal component analysis) and the free-energy profile projected along the virtual-bond dihedral angles γ of the backbone revealed the key residues involved in the transitions between the different basins of the folding free-energy profile, in agreement with existing experimental data for 1E0L.

1. Introduction

Computer simulations are being carried out in many laboratories to investigate the physical properties and functions of biological macromolecules, e.g., proteins and nucleotides.[1−5] Our objective in such simulations is to gain an understanding of how inter-residue interactions determine the folding pathways of a polypeptide leading to the final native structure of the resulting protein. Such simulations are based on Anfinsen's[6] experiment on the folding of bovine pancreatic ribonuclease A, which led to the working thermodynamic hypothesis that the polypeptide chain folds to achieve the minimum free energy of the system consisting of the protein plus the solvent environment. Consequently, such techniques as energy minimization, Monte Carlo,[1] and molecular dynamics[2−5] have been applied in a search for the final native structure of a protein and the pathways leading to it.

Molecular dynamics (MD) has the potential to generate canonical ensembles to determine the time course (kinetics), structure, and thermodynamic properties of proteins and of protein−ligand complexes. MD is based on solving Newton's equations with the use of an empirical potential-energy function to determine the time-dependent trajectories for evolution of the velocities and coordinates on the way from the unfolded to the final folded thermodynamically-stable native structure.

While MD with all-atom potential-energy functions has been applied to refine X-ray and NMR structures and to investigate the initial unfolding stages of proteins, this technique has not been applicable for ab initio protein folding except for the smallest fastest-folding proteins. The difficulty in using MD to treat protein folding arises from the necessity to adopt very small (femtosecond) time steps to evolve the folding trajectory, but globular proteins usually take longer time (milliseconds and longer) to fold. To surmount this time-scale problem, coarse-grained models have been developed with which longer-time simulations can be achieved by eliminating those parts of the all-atom potential-energy functions that are responsible for very fast time-limiting and relatively unimportant motions. Such a

* To whom correspondence should be addressed. E-mail: has5@cornell.edu.

[†] Cornell University.

[‡] UMR 5209 CNRS Université de Bourgogne.

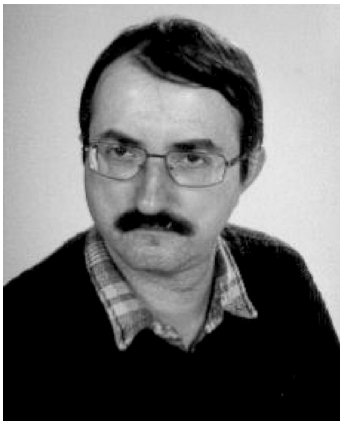
[§] University of Gdańsk.

10.1021/jp9117776
Published on Web 02/18/2010

Gia G. Maisuradze (M.S. Physics 1984, Ph.D. Physics 1990, from Tbilisi State University) is currently a Research Associate in the Department of Chemistry and Chemical Biology at Cornell University. He worked at the Institute of Inorganic Chemistry and Electrochemistry (Georgia) as a junior scientist (1985−1992) and then as a senior scientist (1993−1996). He was a postdoctoral associate at the University Pierre et Marie Curie (1992−1993), the University of Auckland (1997−1998), the Oklahoma State University (2000−2004), and the University of Nevada, Reno (2004−2006), and as a visiting scientist at Cornell University (1999−2000). In Georgia, France, and New Zealand, his research was focused mainly on the theory of resonance Raman spectroscopy. In Oklahoma, he was a pioneer in the development of one of the fitting methods (IMLS) for potential energy surfaces of unimolecular reactions. Since 2004, his research interests are centered around biological systems (proteins and peptides), particularly for understanding the thermodynamics and kinetics of protein folding.

Patrick Senet is a Full Professor of Physics at the University of Bourgogne. He joined the faculty there in 2001. He received his Ph.D. degree from the Facultés Universitaires Notre-Dame de la Paix (FUNDP) in 1993 and then completed a Postdoctoral Fellowship (1994−1997) at the Max-Planck-Institut für Dynamik and Selbstorganisation (Professor Jan-Peter Toennies' group) in Göttingen. He was a research fellow of Fonds National de la Recherche Scientifique (FNRS) at FUNDP (1997−1999) and of Fonds Wetenschappelijk Onderzoek (FWO) at the University of Antwerp (1999−2001) and at the University of Montpellier (1999) and was a visiting scientist during a sabbatic leave at Cornell University (2007−2008) in the group of Professor H. A. Scheraga. His current research interests deal with conceptual density functional theory, theoretical modeling of water, and of dynamics and folding of proteins.

Cezary Czaplewski is an Associate Professor in the Molecular Modeling Department at the Faculty of Chemistry, University of Gdansk. He received his M.Sc. (1995), Ph.D. (1998) degrees and habilitation (D.Sc.) (2006) in Chemistry from the University of Gdansk. As a postdoctoral research associate (1998−2001) and later visiting scientist, he worked with Prof. H. A. Scheraga at Cornell University. His research interests concern the development and application of methods of molecular modeling to study the structure and dynamics of polymers and biopolymers and are focused on the theoretical study of protein folding and hydrophobic interactions.



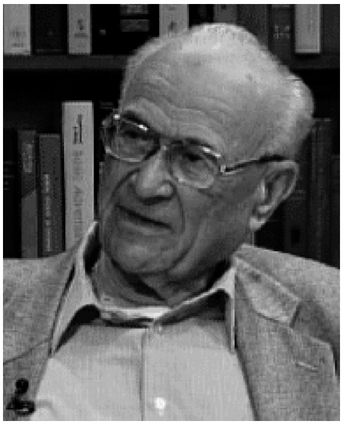
Adam Liwo is a Full Professor and Head of the Molecular Modeling Department at the Faculty of Chemistry, University of Gdansk. He received his M.Sc. (1983), Ph.D. (1989) degrees and habilitation (D.Sc.) (1997) in Chemistry from the University of Gdansk. As a postdoctoral research associate (1990−92 and 1994−95) and later visiting scientist and a senior research associate, he worked with Professor H. A. Scheraga at Cornell University. His research interests concern the development of coarse-grained force fields and algorithms for large-scale simulations of biological molecules, theoretical and experimental studies of the conformations of biologically active peptides, theoretical studies of peroxidation phenomena, and development of numerical algorithms for the analysis of experimental data.

coarse-grained united-residue (UNRES) model has been developed in our laboratory.[7−23]

This article is concerned with the description and application of UNRES to the protein-folding problem with the use of Langevin dynamics. Methods to improve the efficiency of UNRES/MD to compute folding thermodynamics, kinetics, and structures, involving generalized-ensemble methods, are discussed. Finally, methods of analysis of folding trajectories to detect folding pathways, such as principal component analysis and of characterizing conformational changes in terms of free-energy landscapes along a folding trajectory, are discussed.

2. Coarse-Grained United-Residue (UNRES) Force Field

2.1. The UNRES Force Field and Its Application in Molecular Dynamics. The UNRES force field[7−25] is a physics-based united-residue one for polypeptide chains derived as a restricted free energy (RFE) function (or potential of mean force), which corresponds to averaging the energy over the degrees of freedom that are neglected in the united-residue model (such as, e.g., the solvent degrees of freedom, the angles of rotation, χ, about side-chain bonds, and the angles of rotation, λ, of the peptide groups about the $C^{\alpha}\cdots C^{\alpha}$ virtual bonds);[7,10,11] this definition of effective coarse-grained energy functions is also applied by other investigators working on physics-based coarse-grained force fields.[26,27] A very important consequence of this definition is that the effective energy function is directly related to the probability of occurrence of a coarse-grained conformation; this feature is an advantage over the statistical potentials that are sums of terms corresponding to interactions



Harold A. Scheraga obtained his B.S. degree at C.C.N.Y. in 1941 with concentration in Chemistry, Physics, and Mathematics. After obtaining his M.A. (1942) and Ph.D (1946) degrees at Duke University with concentration in Chemistry and Physics, he did postdoctoral research on proteins in 1946−1947 under John T. Edsall in the Physical Chemistry Department at Harvard Medical School. He joined the Chemistry Department at Cornell University in 1947 as Instructor and advanced to Professor in 1958. In 1965, he was appointed to the Todd Professorship and became the Todd professor Emeritus in 1992. He continues to maintain his active research program in experimental and theoretical aspects of protein structure and function.

between different parts of the chain, each derived in the context of the entire protein.

In the UNRES model, a polypeptide chain is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united side chains (SC) and united peptide groups (p) located in the middle between the consecutive α-carbons (Figure 1). Only the united peptide groups and united side chains serve as interaction sites. The α-carbons serve only to define the geometry, and are not interaction sites in the UNRES model (Figure 1). The equilibrium distance of the $C^\alpha \cdots C^\alpha$ virtual bonds is taken as 3.8 Å, which corresponds to planar trans peptide groups. The energy of the virtual-bond chain is expressed by eq 1

$$
\begin{aligned}
U = {} & \sum_j \sum_{i<j} U_{\mathrm{SC}_i\mathrm{SC}_j} + w_{\mathrm{SCp}} \sum_j \sum_{i\neq j} U_{\mathrm{SC}_i\mathrm{p}_j} + \\
& w^{\mathrm{el}}_{\mathrm{pp}} f_2(T) \sum_j \sum_{i<j-1} U^{\mathrm{el}}_{\mathrm{p}_i\mathrm{p}_j} + w^{\mathrm{vdW}}_{\mathrm{pp}} \sum_j \sum_{i<j-1} U^{\mathrm{vdW}}_{\mathrm{p}_i\mathrm{p}_j} + \\
& w_{\mathrm{tor}} f_2(T) \sum_i U_{\mathrm{tor}}(\gamma_i) + w_{\mathrm{tord}} f_3(T) \sum_i U_{\mathrm{tord}}(\gamma_i, \gamma_{i+1}) + \\
& w_{\mathrm{b}} \sum_i U_{\mathrm{b}}(\theta_i, \gamma_{i-1}, \gamma_{i+1}) + w_{\mathrm{rot}} \sum_i U_{\mathrm{rot},i} + \\
& \sum_{m=2}^{N_{\mathrm{corr}}} w^{(m)}_{\mathrm{corr}} f_m(T) U^{(m)}_{\mathrm{corr}} + w^{(3)}_{\mathrm{turn}} f_3(T) U^{(3)}_{\mathrm{turn}} + w^{(4)}_{\mathrm{turn}} f_4(T) U^{(4)}_{\mathrm{turn}} + \\
& w^{(6)}_{\mathrm{turn}} f_6(T) U^{(6)}_{\mathrm{turn}} + w_{\mathrm{bond}} U_{\mathrm{bond}}(d_i) + w_{\mathrm{SS}} \sum_{\substack{\text{dilsulfide} \\ \text{bonds}}} U_{\mathrm{SS}_i} + n_{\mathrm{SS}} E_{\mathrm{SS}}
\end{aligned}
\qquad (1)
$$

with

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

where $T_0 = 300$ K; the temperature-scaling multipliers $f_n(T)$ were introduced in our recent work.[18]

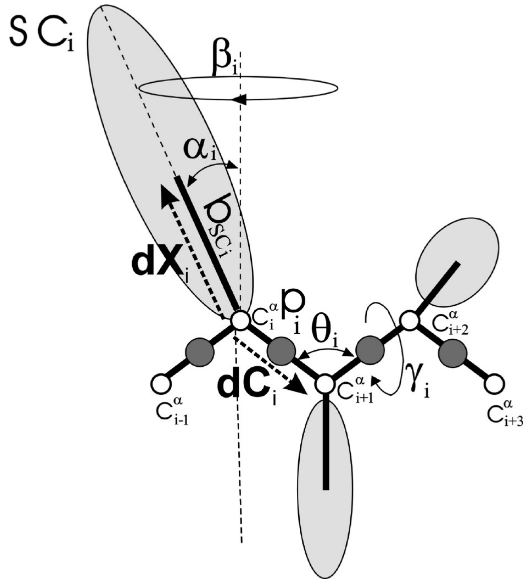


Figure 1. The UNRES model of polypeptide chains. The interaction sites are peptide-bond centers (p), and side-chain ellipsoids of different sizes (SC) attached to the corresponding α-carbons with different "bond lengths", b_{SC}. The α-carbon atoms are represented by small open circles. The equilibrium distance of the $C^\alpha \cdots C^\alpha$ virtual bonds is taken as 3.8 Å, which corresponds to planar trans peptide groups. The geometry of the chain can be described either by the virtual-bond vectors $\mathbf{dC}_i$ ($C^\alpha_i \cdots C^\alpha_{i+1}$), $i = 1, 2, ..., n - 1$ and $\mathbf{dX}_i$ ($C^\alpha_i \cdots SC_i$), $i = 2, 3, ..., n - 1$ (represented by thick dashed arrows, where n is the number of residues, or in terms of virtual-bond lengths, backbone virtual-bond angles θ_i, $i = 1, 2, ..., n - 2$, backbone virtual-bond-dihedral angles γ_i, $i = 1, 2, ..., n - 3$, and the angles α_i and β_i, $i = 2, 3, ..., n - 1$ that describe the location of a side chain with respect to the coordinate frame defined by C^α_{i-1}, C^α_i, and C^α_{i+1}.

The multipliers $f_n(T)$ account for the temperature of those UNRES energy terms which originate from the cumulants of the cluster-cumulant expansion of the RFE[11] and, consequently, scale as $T^{(n-1)}$ for the nth order cumulant.

The terms $U_{\mathrm{SC}_i\mathrm{SC}_j}$ correspond to the mean free energy of hydrophobic (hydrophilic) interactions between the side chains; at present, the Gay−Berne[28] potential is used to handle the anisotropy of these interactions.[8] These terms implicitly contain the contributions from the interactions of the side chains with the solvent. The terms $U_{\mathrm{SC}_i\mathrm{p}_j}$ correspond to the excluded-volume potential of the side-chain-peptide group interactions (and are tuned to produce reasonable bond geometry of peptide chains[7,10]). The terms $U^{\mathrm{el}}_{\mathrm{p}_i\mathrm{p}_j}$ and $U^{\mathrm{vdW}}_{p_ip_j}$represent the energy of average electrostatic and van der Waals interactions between backbone peptide groups, respectively. The terms U_{tor} and U_{tord} are the torsional and double-torsional potentials, respectively, for the rotation about a given virtual bond or two consecutive virtual bonds. The terms U_{b} and U_{rot} are the virtual-bond angle-bending and side-chain-rotamer potentials, respectively. The terms $U^{(m)}_{\mathrm{corr}}$ and $U^{(m)}_{\mathrm{turn}}$ correspond to the correlations (of order m) between peptide-group electrostatic and backbone-local interactions; the terms $U^{(m)}_{\mathrm{turn}}$ (the "turn" terms) involve consecutive segments of the chain. The terms $U_{\mathrm{bond}}(d_i)$, d_i being the length of the ith virtual bond (backbone or side-chain), present in the molecular dynamics implementation of UNRES, are Padé rational functions,[23] which take into account the presence of multiple minima in virtual-bond-stretching potentials of, for example, isoleucine or arginine side chains (earlier[29] we used simple harmonic potentials). The virtual-bond lengths are assumed fixed in other applications of UNRES. The terms $U_{\mathrm{SS}i}$[19] are the energies of distortion of disulfide bonds from their equilibrium configuration, E_{SS} is the energy of formation of an “unstrained” disulfide bond in the chain (relative to the presence of two free cysteine residues), and n_{SS} is the number of disulfide bonds. The w’s are weights of the various energy terms and have been determined by optimization of the potential-energy landscape.[12,15,18] The terms U_{bond}, U_{b}, U_{rot}, U_{tor}, U_{tord}, $U_{\mathrm{p}_i\mathrm{p}_j}^{\mathrm{el}}$, $U_{\mathrm{p}_i\mathrm{p}_j}^{\mathrm{vdW}}$, $U_{\mathrm{corr}}^{(m)}$, $U_{\mathrm{turn}}^{(m)}$, and $U_{\mathrm{S}_i\mathrm{S}_j}$ were derived from ab initio quantum mechanical calculations of the potentials of mean force of appropriate model systems in our earlier[14−16] or recent work,[17,22,23] while the terms $U_{\mathrm{SC}_i\mathrm{SC}_j}$ have been derived[8] from the statistics of side-chain−side-chain distances and orientations determined from the Protein Data Bank; however, we are now replacing these knowledge-based potentials with physics-based potentials derived from all-atom simulations of models of pairs of side chains in water.[24,25]

The UNRES energy terms in eq 1 arise from decomposing the RFE into factors, each of which corresponds to a particular term. If interactions between two UNRES centers or within a single UNRES center are present in a factor, that factor is the PMF of one or two isolated sites and has order one. Examples are the PMFs of virtual-bond deformation or side-chain−side-chain interaction potentials. Factors of order higher than one involve interactions between at least three UNRES centers; the lower-order contributions are subtracted from them so that they contain only the excess free energy arising from coupling between the interactions (the multibody or correlation contributions).[10,11,16] The correlation terms are essential to reproduce regular secondary structures, such as α-helices and β-sheets.[30] The sum of all factors restores the RFE; however, for tractability, only low-order factors are kept in the effective energy function. We found[15] that keeping factors of order up to 4 is sufficient to reproduce protein structures. If feasible, the factors are approximated[11] by analytical cluster cumulants introduced by Kubo.[31]

Initially UNRES was implemented in energy-based prediction of protein structure with the use of the Conformational Space Annealing (CSA) method developed in our laboratory.[32] Owing to its good performance in this task,[33] we extended UNRES to coarse-grained molecular dynamics simulations.[29,34−36] Because the solvent is implicit in UNRES, it contributes to conservative forces (through the RFE) and gives rise to nonconservative forces which originate in energy exchange of the polypeptide chain with the solvent (the stochastic and friction forces). Therefore, we developed Langevin dynamics for UNRES. Because the geometry of an UNRES chain is not uniquely defined by the Cartesian coordinates of the interacting sites, we chose the virtual-bond vectors ($\mathrm{C}^\alpha\cdots\mathrm{C}^\alpha$ and $\mathrm{C}^\alpha\cdots\mathrm{SC}$) as generalized coordinates $\mathbf{q}$. The peptide groups and side chains are represented as stretchable rods with uniformly distributed masses.[29] The Langevin equation for UNRES is given by eq 3[29,34]

$$(\mathbf{A}^{\mathbf{T}}\mathbf{M}\mathbf{A} + \mathbf{H})\ddot{\mathbf{q}} = -\nabla_{\mathbf{q}}U(\mathbf{q}) - \mathbf{A}^{\mathbf{T}}\mathbf{\Gamma}\mathbf{A}\dot{\mathbf{q}} + \mathbf{A}^{\mathbf{T}}\mathbf{f}^{\mathrm{rand}} \qquad (3)$$

where $\mathbf{A}$ is a constant matrix that transforms virtual-bond vectors into Cartesian coordinates of the interacting sites such that $a_{i(k)j} = 0$ [$i(k)$ being the index of a Cartesian coordinate of site k] if the coordinates up to j correspond to virtual-bond vectors of the part of the chain to the right of site k, $a_{i(k)j} = 1$ if the respective coordinates correspond to virtual-bond vectors to the left of site k or to a $\mathrm{C}^\alpha\cdots\mathrm{SC}$ virtual bond containing the side chain with index k, and $a_{i(k)j} = 1/2$ if the coordinate corresponds to the virtual-bond vector containing the peptide group with index i, $\mathbf{M}$ is the diagonal matrix of the masses of the sites (united peptide groups and united side chains) such that m_{ii} is the mass of the site corresponding to the ith generalized coordinate, $\mathbf{H}$ (a diagonal matrix) is the part of the inertia matrix corresponding to the internal stretching motion of the virtual bonds with $h_{ii} = (1/12)m_{\mathrm{p}}$ (m_{p} being the mass of a peptide group) for peptide groups and $h_{ii} = (1/3)m_{\mathrm{SC}_{j(i)}}$ ($m_{\mathrm{SC}_{j(i)}}$ being the mass of the side chain corresponding to the ith generalized coordinates) for side chains,[29] $\mathbf{\Gamma}$ is the diagonal friction tensor (represented by the friction matrix) acting on the interacting sites such that γ_{ii} is the Stokes coefficient of the site corresponding to the ith coordinate, $\mathbf{f}^{\mathrm{rand}}$ is the vector of random forces acting on interacting sites, U is the UNRES effective energy defined by eq 1, and $\nabla_{\mathbf{q}}$ denotes the gradient in $\mathbf{q}$. The balance between the stochastic and friction forces (which results from the fluctuation−dissipation theorem[37]) provides constant average temperature; consequently, Langevin dynamics generates canonical ensembles.

We developed[34] a stochastic analogue of the velocity Verlet algorithm.[38] Our algorithm is a simplified version of the stochastic integrator developed by Guarnieri and Still[39] that we also modified[31] to solve the nondiagonal equations of motion (eq 3). For faster generation of canonical ensembles, we also applied the velocity-Verlet algorithm with the Berendsen thermostat[40] (without explicit friction and stochastic terms); later we introduced[41] Nosé−Hoover[42,43] and Nosé−Poincaré[44] thermostats to generate canonical distributions for regular molecular dynamics (without explicit friction and stochastic terms).

2.2. Capabilities of the UNRES/MD Approach: Extension of Time Scale. Having developed the UNRES/MD approach, we subsequently determined the speed up of simulations with respect to all-atom MD. In principle, a speed up resulting from substantial reduction of computational cost and averaging out the secondary (fast-moving) degrees of freedom when passing from the all-atom representation to UNRES could be expected. Taking the Ala_{10} polypeptide in water as an example, we found that UNRES MD offers a 4000- and 60-fold speed up relative to all-atom MD simulations with explicit and implicit water, respectively.[34] Compared to all-atom molecular dynamics, the UNRES event-based time scale is 4−7 times wider.[34] The speed-up results from averaging out nonlocal interactions, which was demonstrated in our subsequent study[45] in which we compared the time scale of a simple model of Ac-Gly_2-NHMe (in which each peptide group was represented as a plate participating in only local interactions with its neighbors) with the time scale of the corresponding united-residue model (in which the PMF was obtained by numerical integration over the rotation of the plates about the $\mathrm{C}^\alpha\cdots\mathrm{C}^\alpha$ virtual-bond axes). The frequency spectra of the motion of the $\mathrm{CH}_3\cdots\mathrm{C}^\alpha\cdots\mathrm{C}^\alpha\cdots\mathrm{CH}_3$ virtual-bond−dihedral angle were nearly identical for both the plate and the corresponding coarse-grained model.[45] We also found[34] that, with UNRES, Ala_{10} folds in 0.4 ns on average, while the experimental times of α-helix formation are of the order of 0.5 μs,[46] and that the average folding time of protein A (a 46-residue three-helix bundle; PDB code: 1BDD)[47] with UNRES is 4.2 ns, while even the fastest-folding mutants of this protein fold in microseconds.[46] This means that the event-based time scale for UNRES is larger by 3 orders of magnitude than the experimental time scale. This is caused by averaging out the secondary degrees of freedom, and strongly suggests that UNRES MD can be used in ab initio studies of protein folding in real time.

To test the capability of the UNRES/MD approach to fold proteins with Langevin dynamics, we carried out test UNRES/MD simulations[35] on a number of proteins with lengths from

28 to 75 amino-acid residues for which the native-like structures were global minima as found by the CSA method. In these initial studies, we used the UNRES-4P force field[15] determined by hierarchical optimization with four training proteins: 1GAB, 1E0G, 1E0L, and 1IGD. Most of the test proteins folded to nativelike structures, although the force field was optimized using the CSA-generated and not MD-generated decoy sets. The average folding time was only 2.3 ns even for 1CLB, which was the largest protein considered (75 residues); for this protein, the folding required only about 5 wall-clock hours with a single AMD Athlon(tm) MP 2800+ processor on average, this wall-clock time being similar to that required for global optimization of this protein with the CSA method, which requires use of about 100 processors. These results demonstrated that UNRES MD is a practical approach to study folding pathways.

Because the force field used in the initial studies with UNRES/MD mentioned above was optimized using the decoys generated with the CSA method,[15] it did not reproduce the true thermodynamics of protein folding. In particular, the folding-transition temperatures were of the order of 500−900 K.[35] Moreover, the simulated folding usually occurred according to the diffusion-and-collision scenario[48] with initial formation of secondary-structure elements, which later docked to each other to form the tertiary structure. We fixed the above problems by reoptimization of UNRES using the decoys generated in replica-exchange MD (REMD) runs of the training proteins and taking into account the thermodynamic characteristics of their folding transition.[18,23]

While UNRES can be used to study the folding of proteins with size less than 100 amino-acid residues in the single-processor mode, the folding of large proteins is not possible to simulate in real time with a single processor per trajectory even with the speed-up that UNRES offers. Therefore, recently we parallelized[49] the energy and force calculations, achieving a 200-fold speed-up with 512 processors of the IBM BlueGene per conformation for proteins with size of about 800 amino-acid residues. On systems with less fast communication (but faster processors than those of IBM BlueGene), the achievable speed-up is 32 with 64 processors. This means that, with the advantage of massively-parallel machines, the folding or conformational changes of large proteins can be simulated in days, for example, 10 ns of the simulation (i.e., ∼10 μs, taking into account the extension of the time scale because of averaging of the fast degrees of freedom) of the bacterial HSP70 chaperone (600 residues, PDB code: 2KHO) takes 20 h with 128 processors of IBM BlueGene. We are currently working on further reduction of the computation time for large proteins by introducing a cutoff of nonbonded interactions and domain-decomposition parallelization of the code, as in all-atom force fields.[50]

2.3. Simple Application of UNRES/MD: the Folding Kinetics of the B-Domain of Staphylococcal Protein A. The first application of UNRES/MD was simulation of the kinetics of folding of the B-domain of staphylococcal protein A. We ran[51] 400 independent trajectories of Langevin dynamics simulations, the total duration of each trajectory being 35 ns. The force field parametrized on 1IGD using CSA-generated decoys[14] was used. The simulations were run at $T = 500$ K, which was the folding temperature with that force field. Of the 400 trajectories, 380 produced folded structures at least once during the simulation. By analysis of the trajectories, we found[51] that the C-terminal α-helix forms first, which was in agreement with some of the experimental data[52] but contradicted another[53] [later, this discrepancy was reconciled by means of all-atom MD simulations, where we demonstrated[54] that folding initiation depends on external conditions (temperature, viscosity, etc.)]. After initiation, the folding occurs through two different routes, a fast one leading directly to the folded state and a slow one passing through a misfolded kinetic trap (Figure 2). The variation of the content of the nativelike structures with time could be fitted to a sum of two exponentials: one with half-life time $\tau = 8.45$ ns and 77% contribution and the other one with $\tau = 26.9$ ns and 23% contribution. These two components corresponded to the fast- and slow-folding route, respectively.

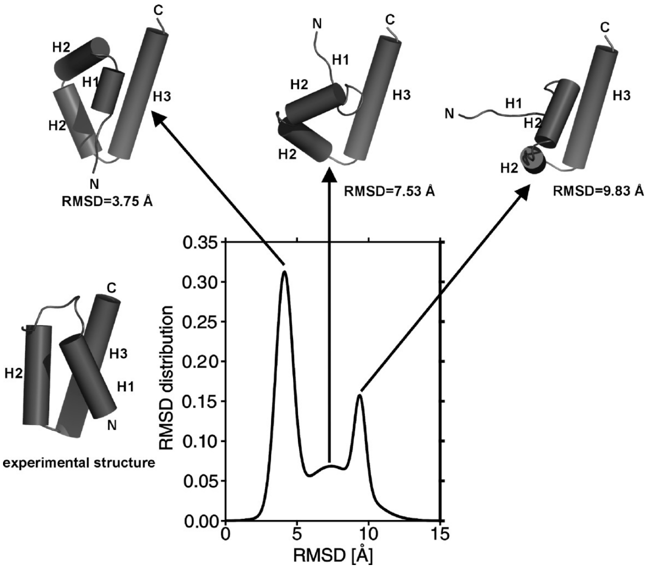


Figure 2. The average conformation of each conformational ensemble encountered during the simulated folding of protein A (top) and the experimental structure of this protein (lower left). The N- and C-termini are marked, and helices are shown as cylinders. To illustrate which structure belongs to which ensemble, the rmsd distribution from the interval of 10−10.5 ns UNRES time of simulated folding is included with arrows pointing from the peak corresponding to a particular ensemble of conformations to the average structure of this ensemble, and the actual rmsd values of each of the three structures from the experimental structure are also included. In all cases, the C-terminal helix (H3) has formed perfectly. However, in the 9.4 Å ensemble (the average structure of which has rmsd = 9.83 Å), the N-terminal helix is unfolded and extends away from the structure and the loops are distorted; such structures constitute the kinetic trap. The N terminus twists toward the core of the protein and the H2−H3 loop reaches the near-native shape in the 7.4 Å ensemble (the average structure of this ensemble has rmsd = 7.53 Å). Finally, the N-terminal domain extends upward in the 4.1 Å ensemble (the average structure of this ensemble has rmsd = 3.75 Å). As seen in the figure, even the N terminus is not fully folded in the average conformation of the native basin, and the H1−H2 loop remains helical. (From Figure 8 of ref 51, reproduced with permission).

Later,[55−57] we studied the folding of protein A and the triple β-strand WW domain from the Formin binding protein 28 (FBP) [PDB code: 1E0L][58] with a force field tuned to MD simulations and by using more powerful methods of analysis. This research is described in Section 4.

3. Extensions of UNRES/MD

3.1. Generalized Ensemble Methods with UNRES. Canonical UNRES/MD simulations can be used to estimate thermodynamic properties of proteins, as well as for a conformational search, but in practice, they tend to become trapped and thus are not effective methods for studying rough free-energy landscapes of proteins with a large number of local minima separated by high energy barriers. It is especially difficult to obtain accurate canonical distributions at low temperatures using conventional MD all-atom simulations, but it is also challenging for UNRES/MD simulations. Recently, to overcome this problem, much attention has been paid to various generalized-ensemble methods with which each state is weighted by an artificial, non-Boltzmann probability weight factor so that a random walk in potential energy space may be realized.[59] The random walk in potential energy space allows the simulation to overcome energy barriers and to sample a much wider conformational space than by conventional methods. It is important to note that kinetic information, such as folding rates, cannot be extracted directly from general ensemble simulations because of the stochastically varying temperature in such simulations.

Three of the well-known generalized-ensemble algorithms are multicanonical algorithm (MUCA)[60,61] (also known as entropy sampling[62,63]); simulated tempering (ST)[64] (also referred to as the method of expanded ensembles[65]); and the replica-exchange method (REM)[66] (also known as exchange Monte Carlo[67] or parallel tempering[68]). The MUCA algorithm directly carries out a one-dimensional random walk in energy space, while ST and REM follow a random walk in temperature space, thereby inducing a random walk in the space of potential energy. REM originated with the work carried out by Swendsen and Wang,[66] but the more familiar form of the REM algorithm was developed by Geyer[69] with his use of Metropolis-coupled Markov chain Monte Carlo.

The MUCA method is based on an artificial distribution of states, in which the probability of occurrence of a state with energy E is scaled by the exponential of the negative of the entropy of the state, $S(E)$, so that uniform probabilities of occurrence of all states with different energies may be obtained. We can define a new variable, the multicanonical potential energy E_{mu} in the following way

$$E_{mu}(E;T_0) = T_0 S(E) = k_B T_0 \ln[n(E)] \tag{4}$$

where T_0 is the reference temperature (the temperature at which the multicanonical simulation is carried out; the sampling efficiency is affected even if thermodynamic quantities are independent of T_0), $S(E)$ is the entropy of the state with energy E, k_B is the Boltzmann constant, and $n(E)$ is the number of conformations with energy E (i.e., density of states). In the MUCA method, the probability of occurrence of a state with energy E, is defined by eq 5

$$P(E) \propto n(E)\exp\{S(E)/k_B\} = n(E)\exp\{-E_{mu}[E;T_0]/k_B T_0\} = \text{const} \tag{5}$$

The MUCA Monte Carlo simulation can be performed with the following modified Metropolis acceptance criterion, with **X** and **Y** denoting the UNRES conformation, respectively, before and after the perturbation

$$W(\mathbf{X}|\mathbf{Y}) = \begin{cases} 1 & \text{for } \Delta E_{mu} \le 0 \\ \exp(-\Delta E_{mu}/k_B T_0) & \text{for } \Delta E_{mu} > 0 \end{cases} \tag{6}$$

where $\Delta E_{mu} = E_{mu}[E(\mathbf{Y});T_0] - E_{mu}[E(\mathbf{X});T_0]$. The MUCA molecular dynamics simulation is carried out by replacing the total potential energy E by the multicanonical potential energy E_{mu} in Newton's equation of motion for the kth particle. For UNRES/MD, the multicanonical equation of motion is given by eq 7

$$\ddot{\mathbf{q}} = -\mathbf{G}^{-1}\frac{\partial E_{mu}(U;T_0)}{\partial U}\nabla_{\mathbf{q}} U[\mathbf{q}(t)] \tag{7}$$

where U is the UNRES potential energy, $\mathbf{q}(t)$ are the generalized coordinates at time t, and $\mathbf{G} = \mathbf{A}^T\mathbf{M}\mathbf{A} + \mathbf{H}$ is the inertia matrix (see eq 3).

In the ST method, temperature becomes a dynamical variable, and both the conformation and the temperature are updated during the simulation with a weight

$$W_{ST}(E;T) = \exp\{-E/k_B T + a(T)\} \tag{8}$$

where the function $a(T)$ is chosen so that the probability distribution of temperature is uniform

$$P(T) \propto n(E) W_{ST}(E;T) = n(E)\exp\{-E/k_B T + a(T)\} = \text{const} \tag{9}$$

The function $a(T)$ is the dimensionless free energy at temperature T. In practice, a discrete space for both temperature T_m $(m = 1, ..., M)$ and corresponding values of the parameters $a_m = a(T_m)$ $(m = 1, ..., M)$ are used. An ST simulation is realized by alternately performing the following two steps: (i) a canonical MC or MD simulation at fixed temperature T_m is carried out for a certain number of steps, (ii) the temperature T_m is updated to the neighboring values $T_{m\pm1}$, using the probability given by the Metropolis criterion

$$W(T_m|T_{m\pm1}) = \begin{cases} 1 & \text{for } \Delta_{ST} \le 0 \\ \exp(-\Delta_{ST}) & \text{for } \Delta_{ST} > 0 \end{cases} \tag{10}$$

where $\Delta_{ST} = E/k_B T_{m\pm1} - E/k_B T_m - (a_{m\pm1} - a_m)$.

REM also uses a discrete space of temperatures, and carries out a random walk in temperature space. In contrast to ST, M canonical simulations (MD or MC) are carried out simultaneously in the REM method, each one at a different temperature. Initially, the temperatures increase with the sequential number of replicas. After every m steps, an exchange of temperatures (or conformations, which is equivalent) between neighboring replicas is attempted, the decision about the exchange being made based on the Metropolis criterion. With a temperature-dependent UNRES force field, the Metropolis criterion is defined by eqs 11 and 12

$$W(\mathbf{X}_i, T_i|\mathbf{X}_{i+1}, T_{i+1}) = \begin{cases} 1 & \text{for } \Delta \le 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases} \tag{11}$$

$$\Delta = [U(\mathbf{X}_{i+1}, T_{i+1})/k_B T_{i+1} - U(\mathbf{X}_{i+1}, T_i)/k_B T_i] - [U(\mathbf{X}_i, T_{i+1})/k_B T_{i+1} - U(\mathbf{X}_i, T_i)/k_B T_i] \tag{12}$$

where T_i is the 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. It should be noted that eq 12 reduces to a simpler form of eq 13 if U does not depend on temperature, that is, if U is energy and not restricted free energy

$$\Delta = (1/k_B T_{i+1} - 1/k_B T_i)[U(\mathbf{X}_{i+1}) - U(\mathbf{X}_i)] \tag{13}$$

The weight factors of MUCA and ST simulations are not known a priori and they have to be estimated before starting simulations, usually by an iterative procedure. It is very difficult to obtain optimal weight factors. REM simulation can be carried out easily because no weight-factor determination is necessary as the weight factor for REM is just a product of regular Boltzmann-like factors. The only disadvantage of REM is that the required number of replicas for large systems can become quite large and computationally demanding. A REM simulation can be applied together with the multiple-histogram reweighting techniques[70] to determine the starting weight factors of MUCA and ST. MUCA, combined with REM in this way, provides a new algorithm, replica exchange multicanonical method (REMUCA).[71] Analogously, ST, started with weight factors determined by using REM, provides a new algorithm, replica exchange simulated tempering REST.[72] The ideas inherent in REM can also be combined with MUCA and ST in another way. Just as REM consists of several replicas of canonical MC or MD simulations, the multicanonical method with replica exchange (MUCAREM) consists of several replicas of multicanonical simulations.[71] The difference between REM and MUCAREM is that the replicas in REM are associated with different temperatures whereas, in MUCAREM, the replicas are associated with different energy ranges over which multicanonical simulations are carried out. The advantage of the MUCAREM approach over the traditional REM is that the probability distributions of energies of different replicas are broader in MUCAREM than in REM; therefore, a smaller number of replicas is required to cover the entire energy range. The replica exchange multicanonical-with-replica-exchange method (REMUCAREM), as in REMUCA, obtains the starting weights from REM simulations as opposed to iterative short MUCA simulations.

Recently,[73] we compared the performance of three generalized-ensemble algorithms for molecular simulations: REM, REMUCA, REMUCAREM in both MC and MD versions for efficient sampling at various temperatures to determine the thermodynamic characteristics of the UNRES force field. Of those, the REM method, especially in its multiplexed MD version (MREMD), turned out to be the most efficient. Among all these simulation methods, the calculated thermodynamic averages, such as canonical average energy and heat capacity, are in good agreement only for the simplest systems tested, poly-L-alanine and protein A. For protein A, all algorithms performed reasonably well, although some variability in the thermodynamic averages was observed, whereas for a more complicated $\alpha + \beta$ protein (1E0G) only replica exchange was capable of producing reliable statistics for calculating thermodynamic quantities.

REM is one of the most effective sampling methods and was initially developed to improve sampling in glassy systems in statistical physics.[66,67] However, following Hansmann's use of the method in simulations of a simple peptide, Met-enkephalin[68] and Sugita and Okamoto's formulation of an MD version of the algorithm (REMD),[74] the REM method has been applied extensively in biomolecular simulations. The multiplexing variation of the REMD method (MREMD)[75] differs from the REMD method in that several trajectories are run at a given temperature. Each set of trajectories run at a particular temperature constitutes a layer. Exchanges are attempted not only within a single layer but also between layers. In our very recent study,[76] we demonstrated that such a procedure increases the power of REMD considerably, and convergence of the thermodynamic quantities is achieved much faster. Intrinsic parallelism of the REM algorithm is extended effectively by multiplexing. Comparison of REMD versus MREMD shows that efficient sampling in REMD requires diffusion in temperature replica space; adding more temperature replicas means that the number of swaps grows quadratically and that either longer simulations are needed or exchanges must be attempted more frequently. On the other hand, the MREMD method takes advantage of both the multiple temperature aspect of REMD, as well as the large number of independent simulations to enhance sampling.

3.2. Application to Structure Prediction and to Compute Folding Thermodynamics. UNRES/MREMD is a robust tool to compute the thermodynamic and structural characteristics of proteins at various temperatures and, thereby, to determine the thermodynamics of protein folding. We implemented[18] the weighted histogram analysis method (WHAM)[70] method to process the results of MREMD simulations. The computed curves of heat capacity and ensemble-averaged native-likeness (e.g., rmsd from the experimental structure) as a function of temperature are good measures of the quality of the force field.[18,76]

For prediction of protein structure, we defined the native structure as the most probable conformational ensemble at a temperature below that of the folding transition. We developed a protocol[18] with which to run UNRES/MREMD simulations, then to determine the heat-capacity curve and, finally, to run a cluster analysis and select the clusters with the greatest probability at a temperature below the folding-transition temperature. Figure 3 shows the results of the implementation of this protocol to predict the structure of target T0411 in the CASP8 blind-prediction test.

Classification of conformations and calculation of ensemble-averaged native-likeness is possible only when the respective experimental structure is known. However, when using the UNRES/MREMD approach for structure prediction, we also need a method to group conformations into families and to rank the families. We use the minimal-tree or minimum-variance clustering[77,78] to define families of conformations. To save computation time in clustering, we consider only those conformations whose contributions together constitute a fraction of 0.99 of the partition function at the temperature(s) of choice. This particular cutoff value can be set arbitrarily; however, setting a higher value or even including all conformations in clustering did not change the compositions and ranking of clusters, except those that have a low probability and, consequently, are unimportant. The temperature is selected as the MREMD temperature closest to the ascending part of the heat-capacity curve. After clustering is accomplished, we compute the probabilities, P_i, of the families in the conformational ensemble at the temperature of choice, from eq 14, where i ranges from 1 to the number of families

$$P_i = \frac{Z_i(T)}{Z(T)} = \frac{\sum_{k\in\{i\}} \exp[w_k - U(T,\mathbf{X}_k)/k_\mathrm{B}T]}{\sum_{k=1}^{N} [w_k - U(T,\mathbf{X}_k)/k_\mathrm{B}T]} \qquad (14)$$

where Z_i and Z are the partition functions of family i and of the entire ensemble, respectively, at temperature T, $\{i\}$ denotes the set of conformations that belong to family i, w_k is the weight factor of the kth conformation calculated using WHAM (and can be considered as the entropy of the kth conformation), $\mathbf{X}_k$ denotes the kth UNRES conformation, k_B is the Boltzmann constant. The families are then sorted according to P_i in

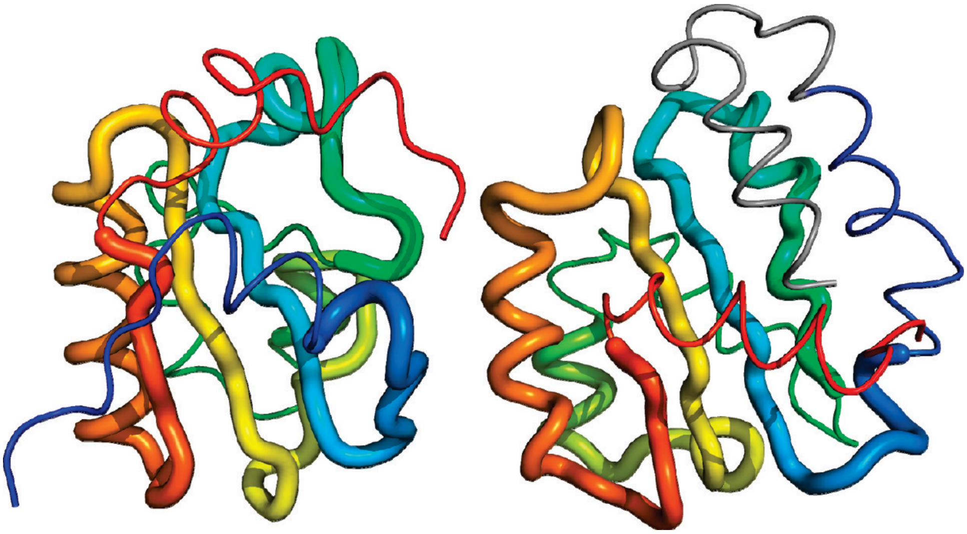


Figure 3. The predicted structure of CASP8 target T0411 using UNRES/MREMD (right) compared with the native structure (left). The correctly predicted parts of the structure are marked as thick ribbons and the incorrectly predicted parts are shown as thin ribbons. The chains are colored from blue to red from the N to the C terminus.

descending order. If a given number of candidate structures must be selected (as in the CASP exercise), the cutoff in clustering or in the clustering method can be adjusted so that the sum of the probabilities of these families is not less than a predefined threshold value. We select the representative of each cluster in the following manner. First, we superpose all conformations on the one which has the largest P_i (eq 14) among the conformations of this cluster. Then, using the superposed coordinates, we calculate a weighted average conformation. Finally, we choose that conformation of the cluster as a representative of the cluster that has the smallest rmsd from the average conformation.

Multiplexed replica exchange (MREMD) method is the method of choice for studies of the thermodynamics of protein folding. It makes various thermodynamical properties available as a function of temperature through histogram reweighting techniques (WHAM). It facilitates fully physics-based prediction because low free-energy minima are accessible through accelerated relaxation.

4. Use of Principal Component Analysis for UNRES/MD Protein-Folding Trajectories: Case Studies with 1E0L and 1BDD

4.1. Principal Component Analysis. To understand the thermodynamics and kinetics of protein folding, knowledge of the free-energy landscape (FEL), which governs the motion of a polypeptide chain, is required. The energy landscape language has emerged for experimentalists and theorists to describe how proteins fold and function.[79−81] The picture of the FEL of proteins has benefited from a variety of experimental studies[82−84] of fast-folding events, and computational studies[85−87] of small fast-folding proteins and peptides.

It should be noted that the FELs determined from canonical MD simulations at temperatures significantly lower than the folding-transition temperature are usually nonequilibrium landscapes because canonical simulations take very long to equilibrate. Generalized-ensemble algorithms,[59] in which walks in temperature or energy space are carried out, converge much faster than canonical sampling and should be used to obtain equilibrium FEL's. On the other hand, the nonequilibrium FEL's resulting from canonical simulations are also valuable, because they provide condensed information about the frequency of visiting particular regions of conformational space during the simulated folding. It must be borne in mind, however, that these FEL's are dependent on simulation setup such as trajectory length, the number of trajectories run at a given temperature, and even the starting conformation(s). In this section, we discuss the FEL's calculated from canonical trajectories which, as remarked above, are generally not equilibrated. However, because we ran our calculations close to the folding-transition temperatures for the two proteins considered in Section 4.2, which lowers the free energy barriers between conformational states, the FEL's should be close to equilibrium FEL's.

Molecular dynamics simulations based on atomic[88,89] and coarse-grained[35] models provide the atomic- and coarse-grained-level pictures, respectively, of protein motion and the connection to the underlying FEL. However, finding a relatively small and appropriate set of coordinates along which the intrinsic folding pathways can be identified still remains challenging for biological molecules containing many thousands of degrees of freedom. Commonly used reaction coordinates (radius of gyration, rmsd with respect to the native state, etc.) are arbitrary and do not necessarily capture the features of protein energy landscapes.

In a protein, out of thousands of modes, only a few modes contain more than half of the total fluctuations of the system, and the first few modes usually describe global, collective motions. Therefore, a strategy is needed to identify the most important (slow) modes. For this purpose, principal component analysis (PCA)[90] is one of the most efficient methods.

The PCA method is based on the covariance matrix with elements C_{ij}

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \qquad (15)$$

where x_1, ..., x_{3N} are the mass-weighted Cartesian coordinates of an N-particle system and $\langle \rangle$ is the average over all instantaneous structures sampled during the simulations. The symmetric $3N \times 3N$ matrix $\mathbf{C}$ can be diagonalized with an orthonormal transformation matrix $\mathbf{R}$

$$\mathbf{R}^{\mathrm{T}}\mathbf{C}\mathbf{R} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_{3\mathrm{N}}), \qquad (16)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3N}$ are the eigenvalues, and $\mathbf{R}^\mathbf{T}$ is the transpose of **R**. The columns of **R** are the eigenvectors, or the principal modes; the trajectory can be projected onto the eigenvectors to give the principal components (PCs) $q_i(t)$, i = 1, ..., 3N

$$\mathbf{q} = \mathbf{R}^\mathbf{T}(\mathbf{x}(t) - <\mathbf{x}>) \tag{17}$$

The eigenvalue λ_i is the mean-square fluctuation in the direction of the principal mode. The first few PCs typically describe collective, global motions of the system, with the first PC containing the largest mean-square fluctuation.

An alternative to calculating the principal components by using eqs 16 and 17 is the singular value decomposition (SVD).[91,92]In this technique, the matrix x(t) − <x> (n × 3N, where n is the number of snapshots) is decomposed into the matrix $\mathbf{QSV}^\mathbf{T}$, where $\mathbf{Q}_{3N\times3N}$ (a unitary matrix) is the matrix of left singular vectors, $\mathbf{S}_{3N\times3N}$ is the diagonal matrix of singular values and $\mathbf{V}_{3N\times n}$ (a unitary matrix) is the matrix of right-singular vectors. The columns of the matrix **Q** are equivalent to the principal components defined by the matrix **R** of eq 16, while the diagonal (i.e., nonzero) elements of the matrix **S** are equivalent to the square roots of the eigenvalues of the matrix **C** (eq 16). The SVD was applied in studying the dynamics of α-amylase inhibitor[91]and in the analysis of the Monte Carlo dynamics of lattice protein models.[92]

Although PCA can separate the modes of motion based on amplitude, one should be careful in interpreting the results of this analysis. First, the set of modes capturing the major fluctuations of a system depends on the width of the sampling window. In other words, with increasing width of the sampling window, more and more slower modes can acquire larger amplitudes and appear as the dominant modes.[93] Second, the principal components of multidimensional random (normal) diffusion are cosine shaped,[94] which can produce patterns that resemble collective behavior and mistakenly be interpreted as a transition of the system from one state to another. This problem exists in only short MD trajectories[94] and should not be confused with PCs of long trajectories, which also may have the shape of a cosine-like function identifying a real transition. Third, it is important to eliminate overall rotation for large-amplitude motion, on which the PCA results ultimately depend, especially for peptides and small proteins.

The cause of the first two problems in PCA is insufficient simulation time for complete sampling. Thus, determination of a minimum MD simulation length, which is required for the convergence of sampling, is still an actively-studied topic. Thus far, there is no unique solution of this problem. The length of a minimum MD simulation can change from system to system and depends on the size of the system. For small peptides, 1 ns all-atom MD simulation is sufficient to achieve convergence of sampling;[95] proteins require much longer simulation times, but how much longer is still not clear. Several years ago, Hess introduced the cosine content of PCs,[96] which is a good indicator of bad sampling; however, accurate study of the convergence behavior in proteins is impossible because current computers are not fast enough to probe all available conformations. Thus, because all-atom MD simulations that must achieve convergence are generally insufficiently long when treating large proteins, it is not easy to satisfy the basic motivation for using PCA in the analysis of all-atom MD trajectories, which is the identification of slow modes and their use for prediction of long-time dynamics.

To overcome these problems and study larger proteins, coarse-grained MD trajectories are required. Therefore, in our recent study,[56] of the folding dynamics of the 1E0L protein with the UNRES force field, these problems have been addressed. In particular, we determined the approximate value of the cosine content, as a threshold, separating the times of insufficient and sufficient sampling, which is ∼0.5 for proteins and lowers to ∼0.2 for peptides.[95] In addition, we illustrated[56] (not shown here) the evolution of the PCs with MD simulation time, which was classified into the following three categories: (i) the cosine-shaped projections for the unfolded state, emerging from simple Brownian motion[94] encountered in short-time simulations; (ii) the projections identifying the end of random diffusion and the beginning of the region in the free-energy landscape in which a potential barrier is encountered; and (iii) projections of trajectories that have already overcome random diffusion and have reached the region of the potential barriers on the free-energy landscape, which becomes independent of the starting structure on any segment of a folding trajectory.

For the solution of the third problem, regarding discrimination of the internal motion from the overall rotation, we used the approach proposed by Mu et al.[97] and Altis et al..[98] In this PCA approach, Cartesian coordinates are replaced by internal coordinates, which are the backbone coordinates (θ_i,γ_j) in UNRES. To avoid potential problems due to the periodicity of the angles, the space of backbone angles is transformed to a linear metric coordinate space, that is

$$\begin{aligned} x_i &= \cos(\theta_i), \quad x_{i+1} = \sin(\theta_i) \\ x_j &= \cos(\gamma_j), \quad x_{j+1} = \sin(\gamma_j) \end{aligned} \tag{18}$$

where i and j are the numbers of θ and γ angles, respectively.

4.2. Free Energy Landscape of 1E0L and 1BDD. With the above solutions of possible problems, which may be encountered in PCA, we constructed FELs along PCs to study protein folding dynamics. However, in spite of fact that PCA drastically reduces the dimensionality of a complex system, the low-dimensional representation [one-dimensional (1D) and two-dimensional (2D)] of an FEL is not always correct and may lead to serious artifacts.[99,100]

In our recent studies of 1E0L and 1BDD proteins,[55−57] we have investigated the adequateness of low-dimensional FELs for the description of protein folding kinetics and diffusive behavior. The important aspect is to find the criterion for the selection of PCs, along which an FEL can be constructed. Recently, based on the fact that the subspace formed by multiply hierarchical PCs[101] contains the most important molecular conformations, Hegger et al.[102] defined the dimension of the free energy landscape by the number of multiply hierarchical PCs for peptides. In other words, each peak of the probability distribution function of a multiply hierarchical PC corresponds to a different conformational state of the peptide, and PCs with unimodal probability distribution (approaching a Gaussian shape with increasing PC index) describe the fluctuations of the peptide within the specific conformational state. A multiple hierarchical PC is one that is characterized by a highly rugged, anharmonic FEL with many local minima within a multiple number of coarse-grained minima.[101] We employed the approach of Hegger et al.[102] in our studies of coarse-grained trajectories of proteins, which in general nicely described the folding dynamics for most of the proteins. However, for some proteins with complex dynamics, not all peaks of the probability distribution function of multiply hierarchical PCs correspond to conformational states; they may correspond to conformational substates in a large

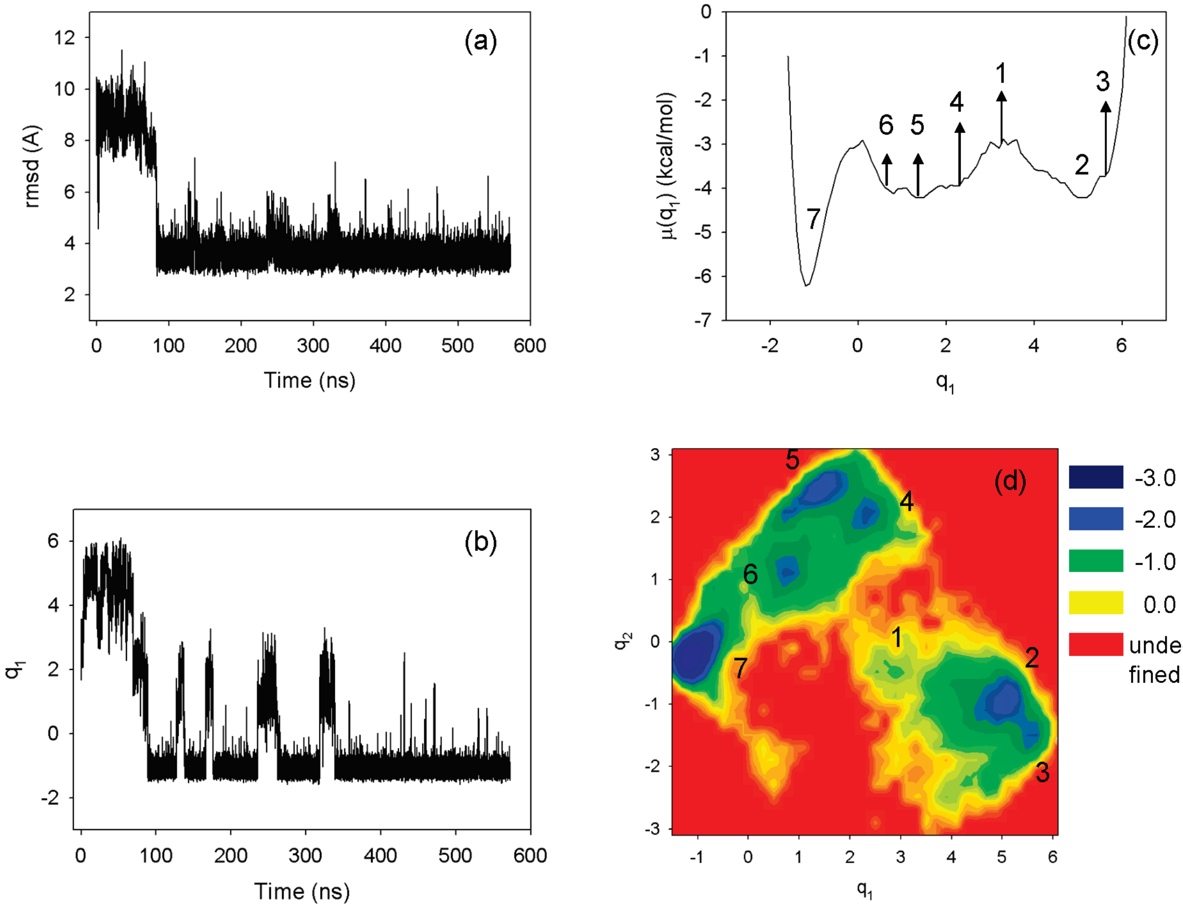


Figure 4. (a) rmsd and (b) first PC as a function of time; (c) 1D and (d) 2D FELs (in kcal/mol) of 1E0L (T = 330 K).

basin[57] and, therefore, careful examination of the structures in each minimum is necessary.

By studying the trajectories of two different proteins, we illustrate here how efficient and correct the approach of Hegger et al.[102] is for the description of protein folding dynamics. In the folding trajectory of 1E0L at 330 K (T_f = 339 K), the first PC (q_1), which exhibited only the multiply hierarchical shape, not only nicely captures the motion of the protein during the entire trajectory, but also contains about half of the overall fluctuations (panels a and b of Figure 4). The second and higher indexed PCs of this trajectory (not shown) belong to either the singly hierarchical category or the harmonic category, which does not contribute significantly to the total fluctuation because it involves low-amplitude local minima and corresponds to local motions.[101] A singly hierarchical PC is one that is characterized by an anharmonic FEL with a number of local minima within only a single coarse-grained minimum.[101] The percentage of fluctuations captured by the singly hierarchical and harmonic PCs is much smaller. Thus, based on the definition by Hegger et al.,[102] a 1D representation of the FEL, that is, a free energy profile (FEP), of 1E0L should suffice to describe its main features correctly. Panels c and d of Figure 4 illustrate 1D and 2D FELs constructed along the first PC, $\mu(q_1) = -k_BT \ln P(q_1)$, and along the first two PCs, $\mu(q_1,q_2) = -k_BT \ln P(q_1,q_2)$, respectively, where P, T and k_B are the probability distribution function, the absolute temperature, and the Boltzmann constant, respectively. Indeed, the 1D FEL clearly illustrates not only all conformational states (three-state folding), which is in agreement with biphasic kinetics for folding, observed in experiment,[103] but also all conformational substates (local minima) of each conformational state can more or less be identified. Since the second PC belongs to the singly hierarchical category, the 2D FEL does not reveal any new conformational state (panel d). Also, except for making the local minima more distinguishable than they are in the 1D FEL with slight rearrangements of the coordinates, no further changes are observed in the 2D FEL. The numbers in panels c,d indicate the minima of each conformational state. No new local minima or major change in the folding kinetics were revealed by higher-dimensional (≥3D) FELs (not shown here).[57]

Unlike the 1E0L trajectory, the first four PCs exhibit the multiply hierarchical shape (not shown here) in the MD simulation of 1BDD at 310K (T_f = 320 K), and the percentage of the fluctuations captured by these PCs are ∼14%, 12%, 7%, and 6%, respectively.[55] Thus, the folding dynamics is more complicated and a multidimensional FEL is required. Figure 5 shows the rmsd as a function of time (a), and the FEL of the MD trajectory along the first (b), the first two (c) and the first three (d,e) PCs, $\mu(q_1,q_2,q_3) = -k_BT \ln P(q_1,q_2,q_3)$, respectively. Panel d shows all points in the 3D FEL space with $\mu \leq 0$ kcal/mol. Since the folding−unfolding pathways are not clearly illustrated in this plot because of strong overlapping of points corresponding to diverse energies, we plotted the same 3D FEL with only the lowest free energy points in panel e. The numbers in each panel indicate the conformational states of the folding/unfolding trajectory. Figure 5 illustrates how insufficient the low-dimensional FELs are (panels b and c) for a correct description of the folding dynamics. The 3D representation of the FEL (d,e) is necessary to illustrate the complete characterization of the MD trajectory. Since the fourth PC also exhibits a multiply hierarchical shape, the complete FEL must be four-dimensional. Since it is impossible to plot the 4D FEL, we

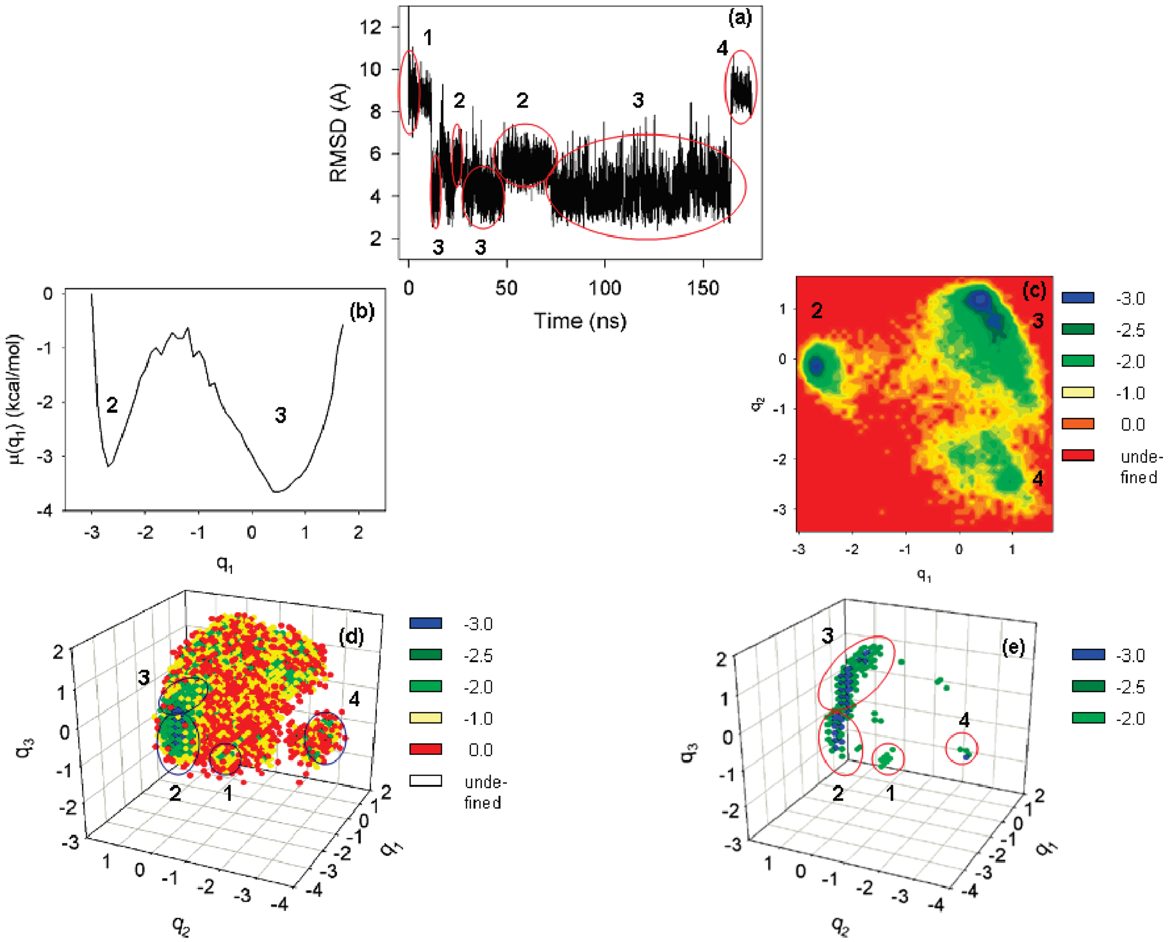


Figure 5. (a) rmsd as a function of time, (b) 1D, (c) 2D, and (d,e) 3D FELs (in kcal/mol) of 1BDD (T = 310 K). (From Figure 1 of ref 55, reproduced with permission).

represented the 4D FEL[55] in tabular form (not shown here); however, we could not find any new major basins in the 4D FEL, which might have been hidden in the 3D FEL. The reason for this absence of any new major basins can be a slightly pronounced second minimum along the fourth PC (not shown here).

The activation barrier between non-native and native states in the multidimensional FELs (3D and higher) is ~2.2 times lower than in the 1D and 2D FELs. This means that not only can the folding pathway and kinetics be incorrect in a low-dimensional representation, but the diffusive behavior can also be misinterpreted. The point is that, after studying the diffusion in the folding dynamics of UNRES trajectories,[56] we observed that the diffusion of a protein in conformational space is anomalous and of two types: subdiffusion and superdiffusion. Since subdiffusion indicates that a system is trapped in local minima in conformational space, and superdiffusion emerges when the system makes long jumps in conformational space, the drastic change of the activation barrier height may cause the change of diffusion type.

After investigating several different trajectories of different proteins, we observed that the percentages of the total fluctuations captured by PCs, which were necessary for a correct description of the folding dynamics, are ~40% or higher.[55−57] Thus, the FEL constructed along PCs is correct if these PCs can capture at least 40% of the total fluctuations. This finding can be considered as another criterion for determining the minimal dimensionality for a correct FEL.

4.3. Sequence of FEP $\mu(\gamma)$ along the Primary Amino-Acid Sequence of 1E0L. Another view of the folding thermodynamics and pathways is provided by analysis of sections of the FEP along the virtual-bond-dihedral angles γ of the backbone (see Figure 1). Although, as opposed to PCA, such an analysis does not enable us to extract a few collective variables that capture most of the conformational changes during folding, it provides a more detailed insight into the conformational changes of chain segments. Analysis of the substates of the main chain along the primary sequence for different successful and unsuccessful UNRES folding trajectories may provide insight into the role of different residues in the large conformational changes observed in the folding process (unpublished results). In addition, as shown below, such an analysis enables us to identify key residues in the transition between the basins of the FEL.

A segment of the main-chain is defined here as four successive virtual $C^{\alpha}\cdots C^{\alpha}$ bonds, and its conformation is measured by the sole dihedral angle γ built by these bonds. The different conformational substates[101,104,105] of the main chain can be visualized by projecting the full free-energy landscape along the coarse-grained dihedral coordinates γ.[106,107] The FEP constructed along each dihedral angle coordinate γ, that is, $\mu(\gamma) = -k_BT \ln P(\gamma)$, where $P(\gamma)$ is the residential probability of each dihedral angle,[107] was computed for the folding trajectory of 1E0L (panel a in Figure 4), as an example, and for two other representative UNRES trajectories of 1E0L of the same duration at 330 K. The FEPs are shown in Figure 6 and are compared to

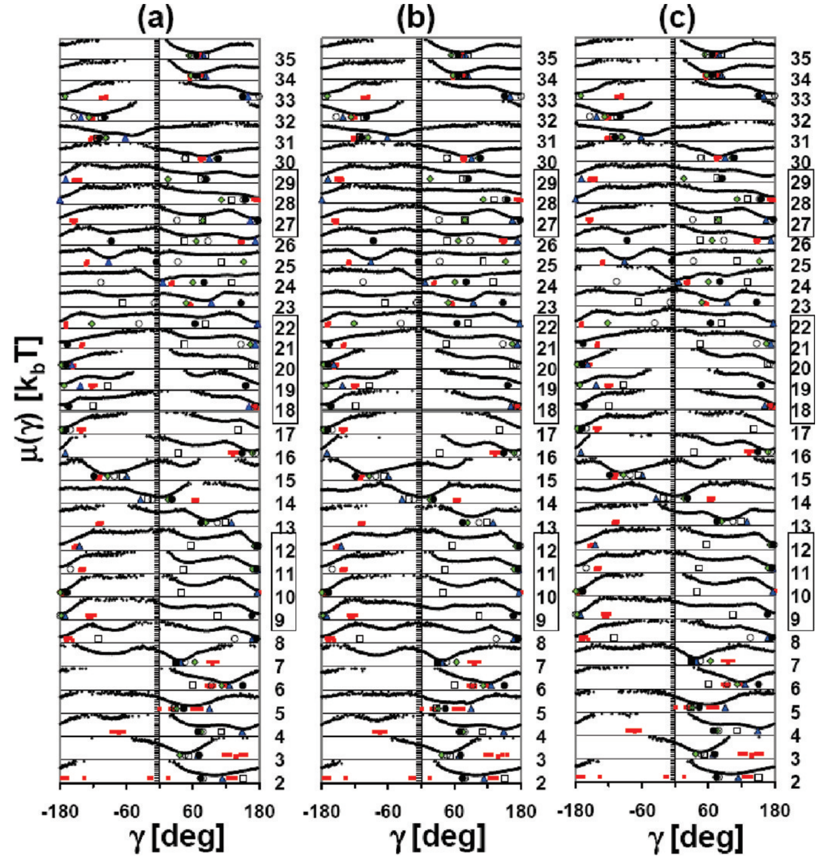


Figure 6. Effective FEP $\mu(\gamma_n)$ (in k_BT units) computed from UNRES MD trajectories of protein 1E0L ($T = 330$ K) along the primary sequence. The potential (full lines) for 1E0L γ_n ($n = 2-35$) for trajectories a−c is described in the text. The NMR-derived structural data (small red squares) are computed from the 10 models of PDB ID code 1E0L.[58] The dihedral angles of the native state 7 (blue triangle), substates of intermediate state 6 (black circle), 5 (green diamond), 4 (open circle), and non-native state 2 (large open square) (Figure 4c) were computed from their representative structures. The rectangles on the ordinates indicate the location of β strands.

the experimental data derived from 10 NMR models of 1E0L.[58] All three trajectories (not shown here) start from a fully unfolded conformation of the polypeptide. The three folded trajectories (for which the FEPs are shown in panels a−c of Figure 6) have different time-dependent profiles. In particular, before jumping to the native state, trajectory a remains in a non-native state for ~80 ns (UNRES time); trajectory b folds quickly but unfolds for ~100 ns (UNRES time), and then jumps back to the native state; trajectory c is similar to trajectory a, with only the rmsd in trajectory c decreasing continuously from the beginning of the trajectory until the polypeptide reaches the native state. As expected according to the ergodic hypothesis, the FEPs computed from trajectories a−c, which reach the native basin of the FEP, are very similar, as shown in Figure 6, despite the fact that exploration of the free-energy landscape in the course of time is very different. In other words, the FEL in these canonical simulations are close to the equilibrium FEL because the temperature of simulation (330 K) was close to the folding temperature (339 K), as stated above. There are some small differences between the FEP of the “folded” trajectories; in trajectory c, the FEP of γ_7 has a less well-defined second conformational substate (around −60°) than in trajectories a and b. In addition, the FEP of γ_{28} is flatter in trajectory b than in the two others.

The global minima of the FEP of each residue in Figure 6a−c are in rather good agreement for each of the ten models derived from the NMR data[58] at each dihedral angle γ. The values of each dihedral angle γ vary by several tens of degrees between the 10 experimental models as can be seen in Figure 6. The largest deviations (>70°) between the UNRES minima of the FEP and the experimental models in Figure 6a−c are for dihedral angles $\gamma_{3,}$ γ_4, $\gamma_{13,}$ $\gamma_{31,}$ and γ_{33}, which correspond to rotations around $C^{\alpha}_{n}-C^{\alpha}_{n+1}$ virtual bonds of residues n and $n+1$, which are part of the β strands (Figure 6).

The sequence of the FEPs in Figure 6a−c reflects the secondary structures of the polypeptide. Deep minima are observed around 180° for the angles γ_n in β strands ($n = 9-12$, $18-22$, $27-29$). In the first β-strand (9−12), most of the FEPs, however, have a second minimum between 60 and 90°. There are also small weak minima in the second and third β strands for most of the residues (see, e.g., $n = 18$ or $n = 27$). The FEPs corresponding to other dihedral angles, have double minima, for example, $n = 15$ and $n = 16$. The occurrence of multiple-minima within the secondary structures is worth noting. This was not observed in all-atom simulations of a 20-residue all-β BS2 polypeptide for which all FEPs were harmonic.[107] Such multiple minima correspond to a metastable (non-native) state of the protein.

Analysis of $\mu(\gamma)$ along the primary sequence reveals key residues involved in the transition between the basins of $\mu(q_1)$ of 1E0L (Section 4.2). Representative structures of the three principal basins of the FEP of $\mu(q_1)$ (centered around $q_1 \approx -1$, $q_1 \approx 2$, $q_1 \approx 5$) have been selected from the UNRES trajectories around the native 7, intermediate 6, 5, 4, and non-native 2, basins [indicated by arrows in Figure 4c]. The location of the dihedral angles γ_n of each of these representative structures around the native state 7, substates of intermediate basin 6, 5, 4, and non-native state 2, in $\mu(\gamma_n)$ are shown in panels a−c in Figure 6. As explained above, however, there are no major differences between the FEPs of the three folded trajectories a−c, and the

TABLE 1: Assignment of Minima in the FEP's of Folding Trajectories to Conformational Substates of 1E0L Found during the Folding Simulated with UNRES/MD and Subsequent PCA Analysis (Figure 4)[a]

γ	substate 6	substate 5	substate 4	non-native state 2
2	-	-	-	-
3	-	-	-	-
4	mss (80°)	mss (80°)	mss (80°)	off
5	-	-	-	-
6	-	-	-	-
7	-	-	-	-
8	-	-	-	mss (−80°)
9	-	-	-	w mss (90°)
10	-	-	-	mss (70°)
11	-	-	-	mss (60°)
12	-	-	-	mss (70°)
13	-	-	-	-
14	-	-	-	-
15	-	-	-	-
16	-	-	-	mss (70°)
17	-	-	-	mss (120°), off
18	-	-	-	mss (−120°)
19	-	-	-	mss (−85°)
20	-	-	-	-
21	-	-	-	mss (70°)
22	mss (80°)	mss (−100°)	off	mss (80°)
23	w mss (160°)	-	off	mss (−70°)
24	off	off	off	off
25	mss (0°)	w mss (100°)	w mss (100°)	w mss (100°)
26	mss (−85°)	mss (70°)	mss (70°)	mss (70°)
27	-	mss (70°)	mss (70°)	mss (70°)
28	-	off	-	off
29	mss (70°)	mss (70°), off	mss (70°)	mss (70°)
30	-	-	-	-
31	w mss (−95°)	w mss (−95°)	w mss (−95°)	w mss (−95°)
32	-	-	-	-
33	-	-	-	-
34	-	-	-	-
35	-	-	-	-

[a] For the non-native state 2, and substates 4, 5, and 6 of 1E0L (Figure 4c), each dihedral angle γ_n of each structure representative of these substates is compared to the corresponding minimum of the FEL $\mu(\gamma_n)$ (Figure 6) computed from trajectory (a). Each dash refers to the dihedral angle γ_n with the deepest minimum of $\mu(\gamma_n)$ for the substate considered. The expression "mss (60°)" means that the dihedral angle belongs to a metastable substate (mss) or secondary minimum located around 60°. The letter w indicates that mss is a weak secondary minimum of $\mu(\gamma_n)$. The entry "off" means that the dihedral angle was not found in any well-defined minimum of $\mu(\gamma_n)$ (flat regions of the FEL or barriers). Numbers in the first column indicate the location of the dihedral angle γ_n along the primary sequence and are in bold face for residues in β strands.

conclusions drawn from trajectory a apply to trajectories b and c. For these "folded" trajectories, we therefore limit the following discussion to the data shown in panel a in Figure 6.

The representative structure in non-native state 2 [$q_1 \approx 5$, Figure 4c] is very different from the structures in the intermediate basin ($q_1 \approx 2$) and from the native state ($q_1 \approx -1$). As shown in Figure 4a,b, the non-native state 2 is explored only during the first 80 ns (UNRES time) of the trajectory. No β strands are formed in state 2; in fact, the orientation of the virtual bonds within the β strands in non-native state 2 corresponds to the well-defined metastable states of $\mu(\gamma_n)$ seen within these secondary structures. In non-native state 2, only 14 dihedral angles γ_n (namely γ_2, γ_3, γ_5 to γ_7, γ_{13} to γ_{15}, γ_{20}, γ_{30}, γ_{32} to γ_{35}) out of 35 are located around the global minimum value of their FEP (Figure 6a, Table 1).

The native structure corresponds to state 7 and, as expected, each dihedral angle γ_n of the structure representing this state is located around the global minimum of $\mu(\gamma_n)$ (Figure 6a). Several transitions between the native state 7 and substates 6, 5 and 4, corresponding to an intermediate state (around $q_1 = 2$ in Figure 4c), were observed during several hundred nanoseconds (UNRES time) after the system has left state 2 which had been explored at the beginning of the trajectory (Figure 4b). For substates 6, 5 and 4, most of the dihedral angles γ_n of the respective representative structures are also located at the global minimum of the FEP $\mu(\gamma_n)$ of the native structure, and only a few dihedral angles γ_n have different locations which correspond, in most cases, to a metastable state (local minima) of $\mu(\gamma_n)$ (Table 1). These few dihedral angles that are not located at the global minimum of the FEP identify the segments which must be reoriented in order to move from the intermediate state ($q_1 \approx 2$) to the native state ($q_1 \approx -1$). The dihedral angles with locations different from those of the native state, which are common to all substates 6, 5 and 4, are γ_4, γ_{22}, $\gamma_{24,}$ γ_{25}, γ_{26}, γ_{29} and γ_{31} (Table 1). In addition, γ_{23} (substates 6 and 4), γ_{27} (substates 5 and 4),and γ_{28} (only substate 5) also have an orientation different from that of the native state.

From the analysis of the FEP of the folded trajectory of 1E0L (Figure 4c and Figure 6a), it appears that the dihedral angles of the structures in the intermediate state [$q_1 \approx 2$ in Figure 4c], which must be reoriented for the protein to reach the native state ($q_1 \approx -1$), correspond mainly to residues within loop 2 between β strands 2 and 3, but four residues (residues 3, 4, 5, and 6 contributing to γ_4) are at the N-terminus and four others (residues 30, 31, 32, and 33 contributing to γ_{31}) are at the C-terminus (see Table 1). These findings agree with the biphasic kinetics of folding of 1E0L found experimentally[103] and in off-lattice simulations.[108] The biphasic kinetics was explained by coexistence of two folding pathways: the most probable is a slow three-state (non-native, intermediate, and native) folding path and the less-probable is a fast two-state (non-native and native) folding path. Mutational analysis and simulations pointed to an intermediate state in the slow folding path in which loop 2 (residues 23−26) exists in a non-native conformation but which does not prevent the formation of most of the native contacts.[103,108] The intermediate state ($q_1 \approx 2$), revealed in Figure 4c by PCA agrees with this hypothesis, as shown by the analysis of the FEP projected on γ (Table 1). In addition, the non-native location of γ_4 and γ_{29} and, to a less extent, γ_{31} in substates 6, 5, and 4 (Table 1) could explain why the mutation of Trp30 and the truncation of the first five residues of 1E0L induced a loss of the slow-folding path.[103]

Comparison of FEP $\mu(\gamma)$ for each γ along the primary sequence with the FEP $\mu(q)$ computed along the collective PCA coordinate q provides a basis to discriminate between the roles of different residues in the major transitions between the basins of the FEP $\mu(q)$ visited by the protein in the folding process. Similarly, a comparison, between the diffusion of the main-chain on the FEL projected on the different main-chain segments [$\mu(\gamma)$], and on the FEL $\mu(q)$, should provide information about the contributions of the different residues to the conformational folding dynamics as will be explored elsewhere (unpublished results).

5. Conclusions

Use of the coarse-grained model UNRES has enabled us to study protein folding with Langevin MD and surmount the time-scale problem, that is, it has been possible to progress from the unfolded to the folded states of several proteins and to identify their native structures and the kinetics of their formation. Further, with a temperature-dependent version of UNRES, we have been able to include entropic effects and thereby determine thermodynamic changes between the unfolded and the folded states.

By extensive parallelization of the UNRES energy and force calculations, folding can now be achieved for larger proteins containing up to almost 1000 amino-acid residues with the force field parametrized by MD simulations.

Further extensions of UNRES/MD have been achieved by applications of generalized ensemble methods with UNRES. A detailed examination of several such methods has enabled us to focus on multiplexed-replica exchange MD (MREMD) as the best one to use with UNRES to determine structure and thermodynamics of protein folding, even though this approach cannot be used to determine folding kinetics.

With the aid of principal component analysis of UNRES protein folding trajectories, it has been possible to identify structures along the folding pathways and to demonstrate that only a few (low-indexed) principal components can capture the main structural features of a protein-folding trajectory. In addition, a comparison, between the structures that are representative of the FEP along the collective coordinate of protein folding (computed by PCA) and the FEP projected along the virtual-bond dihedral angles γ of the backbone, revealed the key residues involved in the transitions between the different basins of the folding FEP, in agreement with existing experimental data for 1E0L.[103]

These recent enhancements have increased the utility of the UNRES force field whose development was initiated a decade ago.

Acknowledgment. This work was supported by grants from the National Institutes of Health (GM-14312), the National Science Foundation (MCB05-41633), and the Polish Ministry of Science and Education (0490/B/H03/2008/35). P.S. thanks the Centre Nationale de Recherche Scientifique (CNRS) for support (délégation CNRS) during a sabbatic leave. This research was conducted by using the resources of (a) our 880-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 Supercomputer Center, (c) the John von Neumann Institute for Computing at the Central Institute for Applied Mathematics, Forschungszentrum Juelich, Germany, (d) the Beowulf cluster at the Department of Computer Science, Cornell University, (e) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, and (f) the Interdisciplinary Center of Mathematical and Computer Modeling (ICM) at the University of Warsaw.

References and Notes

(1) Leach, A. L. *Molecular modelling. Principles and Applications*. Pearson, Prentice Hall: New York, 2001; pp 303−558.

(2) Duan, V.; Kollman, P. A. *Science* **1998**, *282*, 740–744.

(3) Pande, V. S.; Rokhsar, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9062–9067.

(4) Daggett, V. *Chem. Rev.* **2006**, *106*, 1898–1916.

(5) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57–83.

(6) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309–1314.

(7) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1715–1731.

(8) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.

(9) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Ołdziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.

(10) Liwo, A.; Kaźmierkiewicz, R.; Czaplewski, C.; Groth, M.; Ołdziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259–276.

(11) Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Chem. Phys.* **2001**, *115*, 2323–2347.

(12) Liwo, A.; Arłukowicz, P.; Czaplewski, C.; Ołdziej, S.; Pillardy, J.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1937–1942.

(13) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035–8046.

(14) Ołdziej, S.; Liwo, A.; Czaplewski, C.; Pillardy, J.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16934–16949.

(15) Ołdziej, S.; Łągiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 16950–16959.

(16) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kozłowska, U.; Scheraga, H. A. *J. Phys. Chem. B* **2004**, *108*, 9421–9438.

(17) Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys: Condens. Matter* **2007**, *19*, 285203−1285203−15.

(18) Liwo, A.; Khalili, M.; Czaplewski, C.; Kalinowski, S.; Ołdziej, S.; Wachucik, K.; Scheraga, H. A. *J. Phys. Chem. B* **2007**, *111*, 260–285.

(19) Chinchio, M.; Czaplewski, C.; Liwo, A.; Ołdziej, S.; Scheraga, H. A. *J. Chem. Theory Comput.* **2007**, *3*, 1236–1248.

(20) Liwo, A.; Czaplewski, C.; Ołdziej, S.; Rojas, A. V.; Kaźmierkiewicz, R.; Makowski, M.; Murarka, R. K.; Scheraga, H. A. Simulation of protein structure and dynamics with the coarse-grained UNRES force field. In *Coarse-Graining of Condensed Phase and Biomolecular systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2008; pp 107−122.

(21) Shen, H.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2009**, *113*, 8738–8744.

(22) Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Comput. Chem.* [Online early access]. DOI: 10.1002/jcc.21399.

(23) Kozłowska, U.; Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *J. Comput. Chem.* [Online early access]. DOI: 10.1002/jcc.21402.

(24) Makowski, M.; Sobolewski, E.; Czapiewski, C.; Liwo, A.; Odziej, S.; No, J. H.; Scheraga, H. A. *J. Phys. Chem. B* **2007**, *111*, 2925–2931.

(25) Makowski, M.; Sobolewski, E.; Czapiewski, C.; Odziej, S.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2008**, *112*, 11385–11395.

(26) Noid, W.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(27) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.

(28) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316–3319.

(29) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 13785–13797.

(30) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys.* **1993**, *98*, 7420–7433.

(31) Kubo, R. J. *Phys. Soc. Jpn.* **1962**, *17*, 1100–1120.

(32) Lee, J.; Liwo, A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2025–2030.

(33) Ołdziej, S.; Czaplewski, C.; Liwo, A.; Chinchio, M.; Nanias, M.; Vila, J. A.; Khalili, M.; Arnautova, Y. A.; Jagielska, A.; Makowski, M.; Schafroth, H. D.; Kaźmierkiewicz, R.; Ripoll, D. R.; Pillardy, J.; Saunders, J. A.; Kang, Y. K.; Gibson, K. D.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7547–7552.

(34) Khalili, M.; Liwo, A.; Jagielska, A.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 13798–13810.

(35) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362–2367.

(36) Rakowski, F.; Grochowski, P.; Lesyng, B.; Liwo, A.; Scheraga, H. A. *J. Chem. Phys.* **2006**, *125*, 204107−1204107−10.

(37) Kubo, R. *Rep. Prog. Phys.* **1966**, *29*, 255–284.

(38) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.

(39) Guarnieri, F.; Still, W. C. *J. Comput. Chem.* **1994**, *15*, 1302–1310.

(40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(41) Kleinerman, D. S.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Chem. Phys.* **2008**, *128*, 245103.

(42) Nosé, S. *Mol. Phys.* **1984**, *52*, 255–268.

(43) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

(44) Nosé, S. *J. Phys. Soc. Jpn.* **2001**, *70*, 75–77.

(45) Murarka, R. K.; Liwo, A.; Scheraga, H. A. *J. Chem. Phys.* **2007**, *127*, 155103−1155103−16.

(46) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opinion Struct. Biol.* **2004**, *14*, 76–88.

(47) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665–9672.

(48) Karplus, M.; Weaver, D. L. *Biopolymers* **1979**, *18*, 1421–1437.

(49) Liwo, A.; Ołdziej, S.; Czaplewski, C.; Kleinerman, D. S.; Blood, P.; Scheraga, H. A. *J. Chem. Theory Comput.* [Online early access]. DOI: 10.1021/ct9004068.

(50) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(51) Khalili, M.; Liwo, A.; Scheraga, H. A. *J. Mol. Biol.* **2006**, *355*, 536–547.

(52) Bai, Y. W.; Karimi, A.; Dyson, H. J.; Wright, P. E. *Protein Sci.* **1997**, *6*, 1449–1457.

(53) Sato, S.; Religa, T. L.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6952–6956.

(54) Jagielska, A.; Scheraga, H. A. *J. Comput. Chem.* **2007**, *28*, 1068–1082.

(55) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *Phys. Rev. Lett.* **2009**, *102*, 238102−1238102−4.

(56) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *J. Mol. Biol.* **2009**, *385*, 312–329.

(57) Maisuradze, G. G.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2010**, *6*, 583−595.

(58) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. *Nat. Struct. Biol.* **2000**, *7*, 375–379.

(59) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.

(60) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249–253.

(61) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.

(62) Lee, J. *Phys. Rev. Lett.* **1993**, *71*, 211–214.

(63) Hao, M.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *98*, 4940–4948.

(64) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451–458.

(65) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776–1783.

(66) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(67) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.

(68) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(69) Geyer, C. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*; Keramidas, E. M., Ed.; Interface Foundation: Fairfax Station, VA, 1991; pp 156−163.

(70) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(71) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329*, 261–270.

(72) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332*, 131–138.

(73) Nanias, M.; Czaplewski, C.; Scheraga, H. A. *J. Chem. Theory Comput.* **2006**, *3*, 513–528.

(74) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *1−2*, 141–151.

(75) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *2*, 775–786.

(76) Czaplewski, C.; Kalinowski, S.; Liwo, A.; Scheraga, H. A. *J. Chem. Theory Comput.* **2009**, *5*, 627–640.

(77) Murtagh, F. *Multidimensional clustering algorithms*; Physica-Verlag: Vienna, Austria, 1985.

(78) Murtagh, F.; Heck, A. *MultiVariate data analysis*; Kluwer Academic: Dordrecht, Holland, 1987.

(79) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.

(80) Brooks, C. L. III.; Onuchic, J. N.; Wales, D. J. *Science* **2001**, *293*, 612–613.

(81) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, 2003.

(82) Gruebele, M. *Annu. Rev. Phys. Chem.* **1999**, *50*, 485–516.

(83) Myers, J. K.; Oas, T. G. *Annu. Rev. Biochem.* **2002**, *71*, 783–815.

(84) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193–197.

(85) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(86) Brooks, C. L. III. *Acc. Chem. Res.* **2002**, *35*, 447–454.

(87) Granakaran, S.; Nymeyer, H.; Portman, J. J.; Sanbonmatsu, K. Y.; Garcia, A. E. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.

(88) Boczko, E. M.; Brooks, C. L., III. *Science* **1995**, *269*, 393–396.

(89) Bursulaya, B. D.; Brooks, C. L., III. *J. Am. Chem. Soc.* **1999**, *121*, 9947–9951.

(90) Jolliffe, I. T. *Principal component analysis*; Springer: New York, 2002.

(91) Doruker, P.; Atilgan, A. R.; Bahar, I. *Proteins: Struct,. Funct., Genet.* **2000**, *40*, 520–524.

(92) Ozkan, S. B.; Dill, K. A.; Bahar, I. *Protein Sci.* **2002**, *11*, 1958–1970.

(93) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567–2572.

(94) Hess, B. *Phys. Rev. E* **2000**, *62*, 8438–8448.

(95) Maisuradze, G. G.; Leitner, D. M. *Proteins* **2007**, *67*, 569–578.

(96) Hess, B. *Phys. Rev. E* **2002**, *65*, 031910−1031910−10.

(97) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45–52.

(98) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111−1244111−10.

(99) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.

(100) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock., *J. Chem. Phys.* **2008**, *128*, 245102−1245102−11.

(101) Kitao, A.; Hayward, S.; Gō, N. *Proteins* **1998**, *33*, 496–517.

(102) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. *Phys. Rev. Lett.* **2007**, *98*, 028102−1028102−4.

(103) Nguyen, H.; Jäger, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3948–3953.

(104) Ansari, A.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 5000–5004.

(105) Frauenfelder, H. F.; Parak, F.; Young, R. D. *Annu. Rev. Biophys. Chem.* **1988**, *17*, 451–479.

(106) Nishikawa, K.; Momany, F. A.; Scheraga, H. A. *Macromolecules* **1974**, *7*, 797–806.

(107) Senet, P.; Maisuradze, G. G.; Foulie, C.; Delarue, P.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 19708–19713.

(108) Karanicolas, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3954–3959.

JP9117776